Fall 2012 Data Mining: Finding the Data and Models that Create Value 15.062 (ESD.754J)
(Welsch)

Information Sheet

Summary and Goals: Data that has relevance for managerial decisions is accumulating at an incredible rate due to a host of technological advances. Electronic data capture has become inexpensive and ubiquitous as a by-product of innovations such as the Internet, e-commerce, electronic banking, point-of-sale devices, bar-code readers, microarrays, genomic sequencing, and intelligent machines. Such data is often stored in data warehouses and data marts specifically intended for management decision support. Data mining is a rapidly growing field that is concerned with developing techniques to assist managers to make intelligent use of these repositories. A number of successful applications have been reported in areas such as credit rating, fraud detection, database marketing, customer relationship management, stock market investments, and bioinformatics. The field of data mining has evolved from the disciplines of statistics (multivariate analysis) and artificial intelligence (machine learning).

This course will examine methods that have emerged from both fields and proven to be of value in recognizing patterns and making predictions from an applications perspective. We will survey applications and provide an opportunity for hands-on experimentation with algorithms for data mining with easy-to-use software and cases.

Our objective is to develop an understanding of the strengths and limitations of popular data mining techniques and to be able to identify promising business applications of data mining. Students will be able to actively manage and participate in data mining projects that have been converted into cases. A useful takeaway from the course will be the ability to perform powerful data analyses in Excel as well as other data-mining systems.

Background: Material on statistics at the level of 15.060 (Data, Models, and Decisions) or 15.074 (Statistical Reasoning and Data Modeling) or 15.075 (Statistical Thinking and Data Analysis) or my permission. Perhaps the most important topic is regression and you might want to review your notes.

Instructor: Professor Roy Welsch, E62-564 (x3-6601), rwelsch@mit.edu. Often, I am available after class, but the best way to see me is to schedule some time by email or phone. Please also feel free to email me with your questions or comments. I will do my best to respond in a timely manner. We will use Stellar for communication with the whole class.

Establishing a Stellar Account: Check first to see if you are already registered on Stellar for 15.062. If not, a message will be generated for us to grant you permission.

Teaching Assistant: David Zhu (zhezhu@mit.edu), third year PhD student in Operations Research and Statistics (ORC). Office hours will be announced.

Course Assistant: Alison Prosek, E62-571 (x4-4378), aprosek@mit.edu will have extra copies of handouts not posted on the website and can often get a message to me.

Text and Data: Data Mining for Business Intelligence, 2nd Edition (2010) by Shmueli, Patel, and Bruce. The XLMiner add-in for Excel can be downloaded using the code inside the back cover. Data for cases and exercises is available at http://www.dataminingbook.com.

Lectures: These will be held in E51-395 from 4:05 to 5:25 on Mon. and Wed. I will use PowerPoint slides during lecture to provide an outline of what I want to cover. These will be available on Stellar (two slides per page) and will also be handed out (four slides per page) so that you can take notes during lecture. It helps if you skim the assigned textbook material before lecture in order to have some idea of what is coming even if you don't understand everything. Please ask questions as I go along.

Recitations: These will be held in room E51-395 on Tuesdays from 4:05-4:55. Generally the recitations will be conducted by the Teaching Assistant and cover some new material related to data analysis and statistical computing, e.g., XLMiner. There will also be time to discuss homework problems, examples, and clear up any confusion from my lectures.

Exams: None. Grades based on homework (cases), participation, and project.

Term Project: A term project will be required. This usually involves exploring data with the methods in the course (best if you are interested in the data, but see the resources listed below) or picking some material outside of the book that we do not plan to cover and demonstrating that you have gained a working knowledge of it. Computing algorithms can be appropriate. More theoretical issues may also be addressed. The report should be about ten pages with additional material (e.g., computer output and extra plots) included as appendices.

Homework: There will be homework about every ten days that will be graded and returned. (Sampling may be used, i.e. only a portion of the problems may be graded and the rest will be just counted. However, solutions will be provided to all of them.)

Grading: Our goal is to have everyone learn the material. If you are having problems, don't let them slide until the end. Talk to us. The homework will count 40%, class participation 20%, and the project 40%. Once we have handed out the solution sheet for a homework set, **late homework will not be accepted.**

Work Load: This is a 4-0-8 course for half a semester for a total of 6 units. We will have three main hours of lecture and one additional hour of recitation or demonstration each week. Homeworks should take the median student about 8 hours each week. If we have misjudged this load (most often because computing can sometimes take more time than we think), please let us know.

Feedback: Let me (or the TAs) know (anonymously, if you wish) what is going right and what is going wrong with lectures, homework, content, etc. I will occasionally invite a random sample of you to talk to me about the course and/or to fill out evaluation forms during the course.

Academic Honesty: It is best to attempt the homework on your own and then ask us questions. In a pinch, talk to your classmates for clarification. What goes on your homework paper **should be your own work. The project should, of course, be entirely your own work.** Please see the

statement about MIT Sloan Academic Standards posted on the 15.062 Stellar site for further details.

Computing: We will be using a data mining package called XLMiner that is an Excel add-in and comes (via download) with the textbook for the course. XLMiner does not work on Macs unless you are using Windows. We will try (if needed) to have some copies of XLMiner on computing lab machines at Sloan in case you have a Mac and are not running Windows. JMP or Matlab (see below) are good substitutes.

There are many other data mining and data analysis packages available that you can use. For example, JMP from SAS and Matlab (with the statistics toolbox plus other data mining toolboxes such as neural networks and bioinformatics). You can obtain these free from http://ist.mit.edu/software-hardware. There are other Excel add-ins on the market as well.

Data for Homeworks and Projects

Here are some places to get datasets for projects and other uses.

1. KDNuggets http://www.kdnuggets.com/datasets/index.html. For a great deal of additional information try the core site at http://www.kdnuggets.com.

2. UCI http://www.ics.uci.edu/~mlearn/MLRepository.html with summary descriptions at http://mlearn.ics.uci.edu/MLSummary.html.

3. DASL http://lib.stat.cmu.edu/DASL/Datafiles/. Other data sets and software may be found at the core link http://lib.stat.cmu.edu/.

4. JSE http://www.amstat.org/publications/jse/jse_data_archive.htm.

10/18/2012v1

Fall 2012 Data Mining: Finding the Data and Models that Create Value 15.062(ESD.754J)
(Welsch)

Tentative Schedule

All readings listed here are in the book by Shmueli, Patel, and Bruce, Data Mining for Business Intelligence, 2nd Edition (2010) (denoted as DMBI) and class notes.

Date (L#)	Topics	Reading
Oct.		
29 M (1)	What is Data Mining?	1, 2
30 T	Rec.: Getting Started with XLMiner (and other software)	
31 W (2)	Data Visualization	3
Nov.		
5 M (3)	Evaluating Classification and Predictive Performance	5
6 T	Rec.: Visualization and Performance Computing	
7 W (4)*	Near Neighbor and Naive Bayes Methods	7, 8
12 M	Holiday	
13 T	Rec.: Near Neighbor and Naive Bayes Computing	
14 W (5)	Classification and Regression Trees	9
19 M (6)	Regression Review and Selection of Variables	6
20 T (7)	Logistic Regression	10
Note interchange of Recitation and Lecture. Class will go until 5:30 on the 20th (5-5:30 optional) and recitation until 5 on the 21st.		
21 W	Rec.: Regression and Classification Trees Computing	
26 M (8)*	Discriminant Analysis	12
27 T	Rec.: Logistic Reg. and Discriminant Analysis Computing	
28 W (9)	Neural Networks	11
Dec.		
3 M (10)	Cluster Analysis	14
4 T	Rec.: NN and Cluster Analysis Computing	

5 W (11)	Affinity Analysis and Association Rules	13
7 F*	Homework Due	
10 M (12)	Dimension Reduction; Bagging and Boosting;	4, Notes
11T	Rec.: Assoc. Rules; Dimension Reduction; Bag and Boost	
12 W (13)	Time Series Forecasting **Final Project Due**	15, 16, 17

There is no final examination in this course. Grades are based on homework, projects, and case studies.

* Denotes tentative homework or case due dates.

10/18/2012 V1

Fall 2011 (Welsch) Data Mining: Applications and Algorithms 15.062

Reserve Books

Here is a list of books that I have placed on reserve:

Berry, M., Linoff, G., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd ed., Wiley, 2011 (ISBN 978-0-470-65093-6).

Berry, M., Linoff, G., Mastering Data Mining: The Art and Science of Customer Relationship Management, Wiley, 1999 (ISBN 978-0-471-33123-0). (on MIT Libraries site at Books24x7.com)

Dunham, M. Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003 (ISBN 9780130888921).

Green, P., Carmone, F., and Wachspress, D. (1977). "On the Analysis of Qualitative Data in Marketing Research," *Journal of Marketing Research*, **14**, 1, pp. 52– 59. http://www.jstor.org/stable/3151054

Han, J., Kamber, M., and Pei, J., Data Mining: Concepts and Techniques, 3rd ed., Elsevier, 2011 (ISBN 978-0-12-381479-1).

Hand, D., Mannila, H., Smyth, P., Principles of Data Mining, MIT Press, 2001, (ISBN 978-0-262-08290-7). (on MIT Libraries site at Books24x7.com)

Hastie, T., Tibshirani, R., and Friedman, J., The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2nd ed., 2009 (ISBN 978-0-387-84857-0).

Hosmer, D.W. and Lemeshow, S., Applied Logistic Regression, 2nd ed., Wiley, 2000 (ISBN 978-0-471-35632-5).

Johnson, R.A. & Wichern, D.W., Applied Multivariate Statistical Analysis, Prentice-Hall, 6th Ed., 2008 (ISBN 9780131877153).

Kutner, M., Nachtsheim, C., Neter, J., Li, W., Applied Linear Statistical Models with Student CD, McGraw-Hill/Irwin, 5th Ed., 2005 (ISBN 9780073108742). (edition w/o CD ISBN 9780072386882)

Labe, R.P. (1994), "Database Marketing Increases Prospecting Effectiveness at Merrill Lynch," *Interfaces*, **24:5**, pp. 1–12. http://www.jstor.org/stable/25061926

Markov, Z. and Larose, D., Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage, Wiley, 2007 (ISBN 978-0-471-66655-4).

Montgomery, D., Peck, E., and Vining, G.G., Introduction to Linear Regression Analysis, 5th ed., Wiley, 2012 (ISBN 978-0-470-54281-1).

Pyle, D., Business Modeling and Data Mining, Elsevier, 2003 (ISBN 978-1-55860-653-1). (on MIT Libraries site at Books24x7.com)

Roiger, R., Geatz, M., Data Mining – A Tutorial-Based Primer, Addison-Wesley, 2003 (ISBN 9780201741285).

Shmueli, G., Patel, N., and Bruce, P., Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Excel with XL Miner, 2nd ed., Wiley, 2010 (ISBN 978-0-470-52682-8).

Tan, P., Steinbach, M., Kumar, V., Introduction to Data Mining, Addison Wesley, 2006 (ISBN 9780321321367).

Trippi, R. and Turban, E. (ed.), Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real-World Performance, Irwin Professional Publishing, 2nd ed., 1996 (ISBN 9781557389190).

15.062 10/30

First Lecture

Schedule will be juggled since Hurricane

Purchase book

Uses code to download XL Miner
└ becomes piece of Excel

Can use Matlab instead Windows Only

Jump

Prof uses

Lecture notes online

Presumption is some bg in stats - Regression
- t-tests

(weird to start class half way through!)

Will generalize some things learned before

Try to get away from linear

Talk about outlyers

(2)

lst hw due Nov 14

Project Due Dec 12

No Final exam

Lots of data is collected + warehased

Want to be able to put it to use

Poorer mans statistics

Asy very large → only then shows up

But often only a finite sample

Can compute t-dist for small samples

If have lots of data → just use it for prob

Don't need to do normalize data forms!

(3)

targeting in marketing
Personalization
Understanding customer

Topics

reduction, exploration, visualization
Classifications, prediction

Supervised → for part of the data we actually know truth (has cancer)
but for other parts of data we don't
(does this other person have cancer?)

Unsupervised → more explatory
to cluster data
Netflix → collaborative filtering
(good way to think about)

(4)

Classification

Given a collection of records → training set

each record contains a set of attributes

One of the attributes is a class

find a model for class attributes as a function of other attributes

Goal: previously unseen records should be assigned as accurately as possible

Can try it out on a test set

Much of states does not tell you about new data

R^2 - penalty for overfitting

(5)

ie: target on customers if they buy a lawn mower or not

Logistic regression (missed)

try to predict response & product
(who buy)

CART - Classification + regression trees

not just who buys

but who will be a good customer

Clustering - points in cluster more similar to other points in cluster vs ~~other~~ pts in ~~not~~ other clusters

How ~~can~~ measure distance?

ie Document clustering - somewhat how Google works

(6)

Deviation / Anomaly Detection

- CC Fraud Detection
- Network Intrusion Detection

Association Rule Discovery

try to produce dependency rule

Detecting insider trading
but must do in nanoseconds
algo trading
how to find a flash crash?

loyalty cards at grocery store
gives you coupons to come back
what it can entice ya to buy
tells how to arrange store

(7)

Recomendation engines

Netlix → large sparse of matrix

~~millions~~ of titles

thousands

but each person only rented 10-100

decrease dimesion of matrix

Spam filters

Gotten pretty good

So people rarely look in their spam filte

Medical insurance history

how make them anonmous enough

would be huge

So many happenstance expeiments

doctors just naturally treat patients differently

(8)

Drug interactions

like to know fast

Shipping tracking

w/ RFID tags

Compare freq of word counts for spam + real email

freq that the word appears

Must deal w/ lots of diff types of data

Continuous, discrete

Integer

bins - but lose resolution

Ordered

Generalization

Overfitting

Regularization

(9)

Adjusted R^2 - builds in a penalty

Diff kinds of big data

- could still be sparse
  (this is silly)
- Or lots of ind w/ few features
  or few ind w/ lots of features
  (n) (p)

Sampling

tons in elections
margin of error
where target advertising + visits
but are looking for small differences

Loss function

Cost of missing customer
vs mailing too many flyers
often not symmetric

Regression w/ over fit

When # parameters = order → perfect fit!

but won't generalize well

Data Partition

Training data

↓

Validation data

↓

Test data ← client has new data

↓

New data

15.062 10/3

Lecture 2

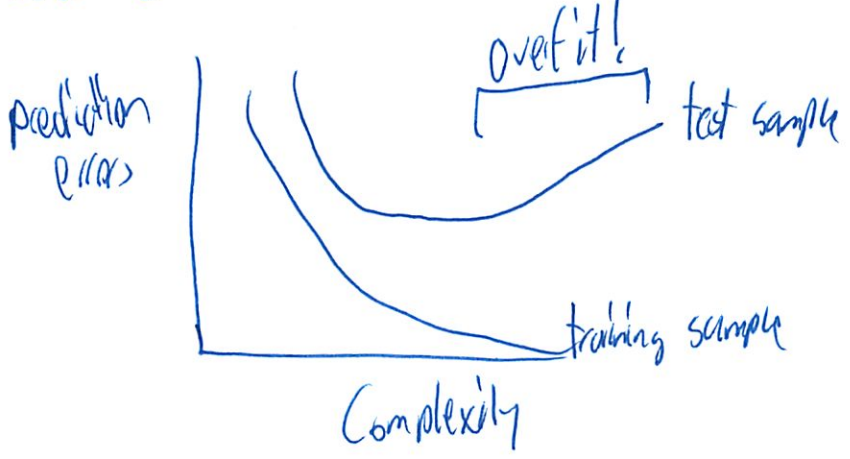
Seating - none left (finish Lecture 1 + Do Lecture 2 - Data Visualization)

Faculty get points for each student in class
↳ not listen

Last time: motivational issues

Paradigm: not using training data to test model
↳ in hands of client
So keep some data at to check what you are doing

Model Complexity



↑ How do we find the sweet spot?

Area

(2)

Need right # of explanitory variables
Don't use up all your degrees of freedom

SEMMA Methodology

Sample
Explore
Modify
Model
Assess

(I think I actually know more about this class than I think)

Every data set has some missing observations

Ensemble methods = combination of methods

Can see some methods good on some data sets

Decision trees hard to explain
Neural Nets hard to explain

(3)

Tuning
robustness to outliers
↳ some data has no value

table where he created these

(I learned about most of these in AI
should just rename / rebrand that class)
or just need a bit pre explanation

Bg on data

Collection of objects + attributes
Use some attributes to predict others
Find w/ random audits

Continuous vs discrete
↳ might need to bin

Ordered vs unordered
↳ can just use ↳ need dummy variables (one fewer than # of cat)
but beware non-equal increments

(Stuff never explained this well before - efficient studying :))

(9)

does default in loan file suggest you will default
└ missing paper

So convert to dummy variables

---

many rows many cols → big data
other files other way

* can't control data cleanly like an experiment
  Often missing values / other factors

Can we ↓ dimensionality of space
lots of diff ways to represent the data

data not perfect
always be skeptical of what you see

outliers
- own cluster
- set aside
  └ can still contain important info
  but perhaps try w/ + w/o models

- info could not be collected
- Or info might not be applicable

how to handle missing data?
diff models do it differently

Sampling

Can hurt if looking for a few key values
data might be huge → but looking for few options might miss it - even w/ random sample

Could we do something upfront?

Types

Simple random sampling
Or don't replace item
LS does affect underlying prob/sample

Stratified
split into groups
then sample
ie draw 10 loan files from each state

(6)

it needs to be somehow related to the outcome
or what you predict

how big a sample do you want?

Curse of dimensionality

high dimensions → means likely that points are sparse
harder to find nearby pts
this is one of the major challenges of data mining!

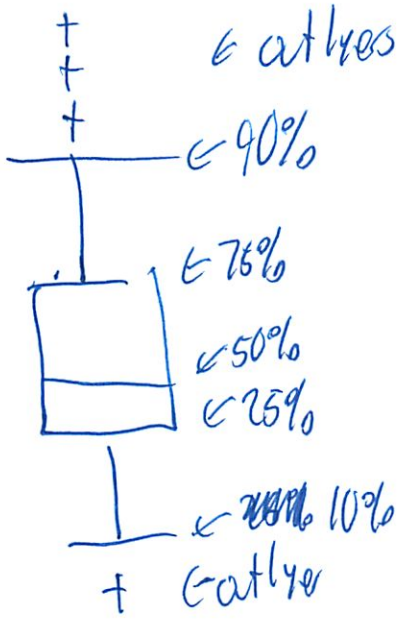
- principal component analysis
- Singular value decomposition

Data Visualization

How do you visualize this?
Want to explore data up front
John Tukey: Exploratory Data Analysis
Do this before running st. techniques
* look for relationships

6

Box Plot



Makes it easy to see things fast

Scatter Plot Iris Attributes

try to show multiple dimensions

Tufte's Graphical Excellence

Clarity, precision, efficiency
no chart junk → fancy styling
encourages eye to compare

Draft picture
↳ shows Iris was not stirred
but see from picture → quickly

(8)

Try diff ways

: over time

bar vs line vs scatter

What is the best way to process?

Cell Imaging

bigger issue

6v extracts features

building matrix of features vs picture of cell

could compare a bunch of box plots

but let's extract a key idea

Star Plots distance is how big feature is

Chernoff Faces

high dimensional data collapsed in a face

but what feature in data do we assign to what face

Bi Clustering

feature + dose

tricky

(9) Wind rose

Certain cells move differently depending on what is going on

Boston housing data
- looked at a lot of variables
- laid at horizontally
- Spotfire

Networks
- ie ebay auctions

Bootstrap Resampling Technique
- Randomly sample your random sample
- we don't know if original data was random (missed)
- Only thing we have is the data
  - Here imposing pdf prob on it
- resample the replacement

(10)

Shows something about the uncertainty

Can do this 1000 x

Can make histogram of means

Chop of 2.5% on each side

Shows sample variability

Why not in textbooks? it costs a lot of computing power

Must break habit of thinking data is randomly distributed

Actual mean + scatter ~~plot~~ mean

Many people think scatterplot is truth

but when randomly sample → it looks diff!

Could do a bunch of boot straps

Draw a bunch of lines a top of each other

That shows you the uncertainty!

(11)

"Could you find a distance b/w scatter plots?"
"Earth Mover Distance"

how much work do ya need to do to move piles of dirt

Distance b/w plots is assignment problem

* is an optimization problem

* Don't accept scatter plot at face value

15.062

Parts 3+4

11/5

HW Wed is due a week later

HW 2 out on Wed

Recitation time: Computing

Wed before Thanksgiving will be class again

Notes on Stellar

Last time: Visualize data

Today: Classification

Stocastic: Sporadic, non-deterministic but not intermittent

Classifier = discriminator

Training data → know actual class

Actual data → off diagonal you lose

Error misclassification rate

$Err = (n_{0,1} + n_{1,0})/n$ symetric cost

(2)

Costs may be asymmetric: extra piece of junk mail vs extra piece of junk mail

Accuracy = 1 − Err

But if few legit yes points

then could classify all as no

then looks like high accuracy

Sensitivity: percent of C_0 classified as C_0

Specificity: " " C_1 " " C_1

Want high #s on both!

but don't want to over treat

Costs of mistakes

Courts inconsistent on in Courts

Lift charts/ROC

Q3

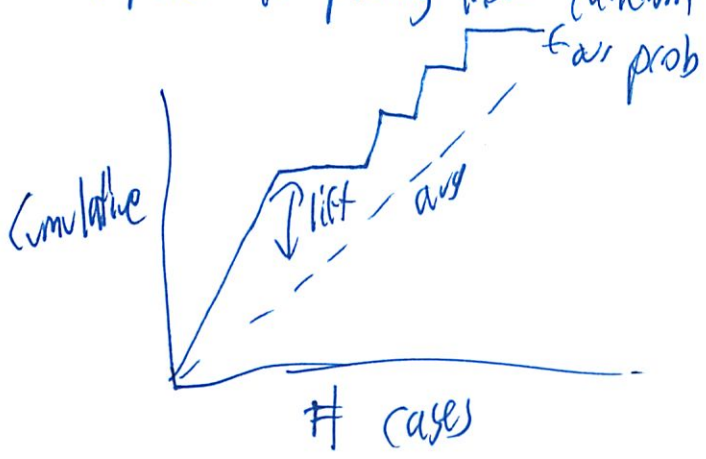
Often times we don't know an exact cost
instead probability

Classifier that gives ya probabilities
the prob that ya will buy

Aka skinning the creme - picking top customers

default usually .5
but can adjust misclassify harms around costs of
making a mistake

Can get prob w/ logistic regression
Convert to lift chart

Compare to picking from random


(4)

hope lift outweighs data analytics base cost

but what are real costs of sucess + failure?

asymetric costs

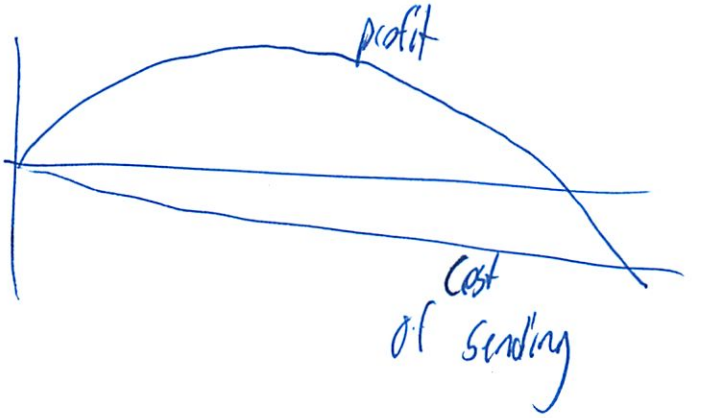
$$\frac{q_0 n_{0,1} + q_1 n_{1,0}}{n}$$

q = Cost of misclassifing

n = number

Can generalize to more than 2 costs

So diff w/ costs of sending mail



(I wish I was much better at this)

(5)

Predictive Performance

he doesn't like R^2, adj R^2
based on training data

not new data

diff is the mistake you made
lots of ways of taking the error
(see slides)
RMSE - cost of mistake is quadratic

but costs to errors - sometimes can actually know
often no reason its symmetric

divisions often use diff methods of forecasting

---

but do we not want to overweight
"black swan" errors?
don't want to reject model if one of these
in test data...

(6)

Oversampling

Maybe only a small fraction of actual data useful!
One cost is also usually substantially higher

Can build training set that is actually 50-50
Still validate it like normal

Bayes Rule

$$\frac{P(S|M)\,P(M)}{P(S)}$$

(missed some - some - titled)

Some people in class
1/w women + 2/3 male is prediction

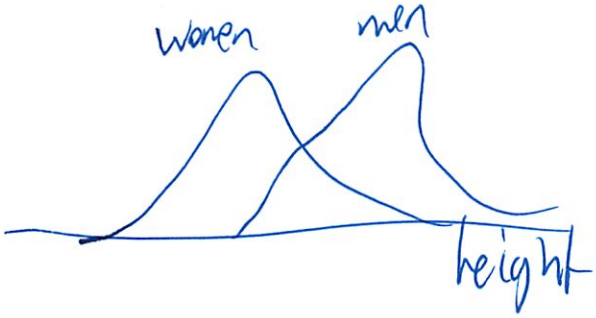
Prob make a mistake on
Can look at a-priori prob

---

will put in classification w/ highest prob

(1)

Bayes' Rule for Min Error



Some overlap
but distinct height

Person covered up
- so know ~~same~~ height - not gender!

$f_0(x) = f(x \mid C_0)$ = density fn given you are a female

- we can gather this pretty easily

Now how are going to classify people
Person walks in height x_0

(8)

$$p(C_1 \mid x = x_0) = \frac{P(x = x_0 \mid C_1) P(C_1)}{P(x = x_0 \mid C_0) P(C_0) + P(x = x_0 \mid C_1) P(C_1)}$$

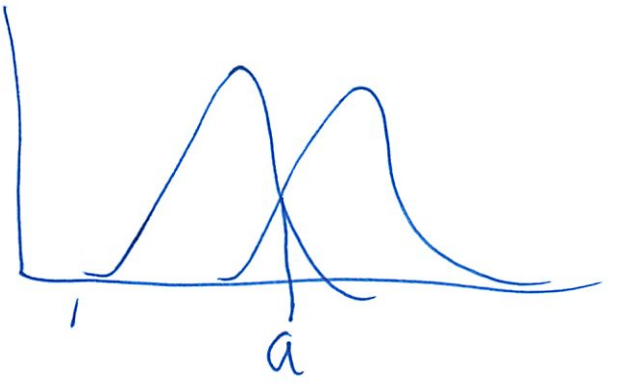
If asume $p(C_0) = P(C_1)$

↳ That males - females 50 - 50

Then can simplify

$p(C_1 \mid x = x_0) > p(C_0 \mid x = x_0)$ if

$f_1(x_0) > f_0(x_0)$



Only good if priors = 1/2 ! (50-50 gender split)

So classify as

(9)

About the best we can do if know density fn
often have a sample

Can compare the outcome of what we have w/ Bayes rule

Assumes we have a prob function/model

Examples

Auditing

$$P(C_0 \mid \text{legal charges}) = \frac{P(\text{legal charges} \mid \text{fraudulent})\, P(\text{fraud})}{P(\text{legal charges})}$$

C_0: fraudulent

So priors → truthful ↳ just client
but when know if legal challenges have been filed or not
we get more info

(10)

Two variables

if there is a covariate b/w them

need more data

Wiew Bayes Classifier

assumed independence of 2 predictive variables

Example mammals vs non-mammals

$$P(A|M)\,P(M) \gtrless P(A|N)\,P(N)$$

Note→ used independence

(Mi is that legal? or is that the point of this?)

Q: what do we do w/ sometimes?

- Coin toss
- missing data
- more later

(11)

banking → will customer get a CC?

Not a lot of discriminating power here
at least at face value

So very high error
when store training data

Nieve Bayes Classifier

In many situations it works pretty well
Often worth a try

But if something is 0 - ruins it
Can fix w/ LaPlace or m-estimate

free variables to pick
Some tricks in lit
Could choose for lowest classification error
Could be called tuning param

Advantages + Disadvantages

(see slides)

might be good enough to just do a ranking

Nearest Neighbors

more complicated than everything is ind

find 5 near neighbors

Majority rules

draw some sort of radius

- can be k-nearest
- or 1 cm radius

Prof: this is one people can understand

if look / don't look like others folks

15.062 11/7

4 Nearest Neighbor + 5 Classification + Regression Trees

Regression → one model for the entire space!

11/28 1pg Project proposal due at end

Something you are interested in!

does not have to be data

Could be algorithms

Last Time: Naive Bayes

nice if could write prob dist w/ good precision

(need to be better at what to use when)

Bootstrap

If don't have multivariate dist

just assume its ind

that is not that bad!

(2)

Response variable
or Continuous

Some fn of input attributes
We don't know what that fn is

Regression → We assume it is a line

We need to think more than linear - but must still restrict

Not <u>parametric</u> → exponential → λ
normal → μ, λ
etc

But this is not parametric
Don't need to find the parameter
he prefers "<u>distribution free</u>"

(3)

New observation → look for nearby values in training data

Look for values that are near what we have

Can think of Euclidean distance

Were have been normalized

Comparable footing / scale free

Some alg that takes account of near neighbors

Simplest : majority rules

How big of a circle?

- fixed radius
- k-Nearest Neighbors

Can show misclassification prob is no worse than twice that ———

If k=n you are overdoing it!

Regression
using up all our degrees of freedom

k-NN: if misclassify error > 50% you've gone way too far!

look at % error validation
find min
w/ least complexity

Shortcomings

- Can be computationally intensive
- # training obs need grows exponentially
- apply dimensionality reduction methods
  - stats
  - optimization
  - computation

(5)

In high dimensions → nothing is nearby

But are many of these variables noise / just we don't need!

Relation to Bayes Classifier

Approx for densities from NN

Count nearby ones in each class

Assuming priors =

Majority rules

So approx. to Bayes classifier

Bayes ≤ NN

error

Regression

$$Y = B_0 + B_1 X_1 + \dots + B_p X_p + \underbrace{\epsilon}_{\text{everything else}} \quad \epsilon = 0$$

maximum likelihood ~~max~~

least squares of errors

(6)

Plug into regression to get the coeffs

But ya violate all sorts of assumptions if $Y = 0, 1$

Stuff is <u>not</u> normally distributed

Could get a <u>logistic regression</u> from <u>Bernoulli regression</u>

Classify to $\hat{f}(x)$ or $1 - \hat{f}(x)$

but $0 \leq \hat{f}(x) \leq 1$ for linear regression

So instead logistic regression

But actually least Sqr Regression ~~if~~ isn't That bad!

<u>Examples</u>

Classify stuff as 0 or 1

Line is where it breaks from majority

Orange to Blue

(7)

We've gotten away from line

Now complicated curve

1-NN will overfit

Bayes Optimal

* ↳ if knew underlying prob distibution

Chart w/ error

* Are trying to find min error
So we have minimal error on test data
How well could we do?
Training → Validation → Test

Distance

Euclidean $|x - \bar{x}|$ ← mean of group

$$\sqrt{(x_1 - \bar{x}_1)^2 + \cdots + (x_p - \bar{x}_p)^2}$$

(8)

With some sort of scaling
how can we compare fairly?

Euclidean does not account for variability or correlation

So instead: Mahalanobis-distance

aka statistical distance

5: Classification + Regression Tree

fitting to a plane?
multiple variables
response variables (missed)

Only rectangles

~~Broad~~ very local
but can overfit

Could do better w/ continuous splice → connections smooth

(9)

Regression Tree

Greedy alg

Regression

Split into 2 halves

Pick a split point

if split there → what happens to sum of split of errors

This alg goes through all possible explanatory variable and decision points

Then could split those regions again

Only go forward

Going until each is in rectangle → overfitting

So need some way to prune

Could also do nested rectangles

(10)

Since greedy - we never look back

Won't give optimal → but hopefully a good sol

inner minimization

We could split again on same variables

Classify vehicle by dropping through the tree

- Goal: classify vehicles into 4 ~~overall~~ miles per gallon categories

Easy to describe

fairly robust to outliers

handles missing data well

splits somewhere else

---

robustness what if we perturb a little

data lot

- remove some data a lot

Sensitivity what if we perturb a lot of data a little

- add in a little noise to all data

11

Prune w/ validation data

Start wacking at tree
to minimize misclassification error
its like ceding # of parameters

Example: Tax Cheat Decision tree

How do we want to split?
Continuous
Recursive Partitioning

Criteria: how pure is the rectangle
Make splits so get as much purity as we can

* How do we measure purity?
  - Gini Index
  - Entropy
  - Misclassification error

  Gini: its like sum of square of errors in regression

Most alg will come w/ default - sometimes can switch

Gini

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

What happens when things =ly distributed among all classes?

Maximum: $1 - \frac{1}{nc}$ when records =ly distributed

Minimum: 0 when all belong to one class

We want small Gini

p is relative probability

Fall 2012 Data Mining: Finding the Data and Models that Create Value 15.062 (ESD.754J) (Welsch)

Homework #1

Due: Wednesday, November 7, 2012

Reading:

DMBI Chapters 1-3 and 5.

Problems (individual work unless otherwise noted):

1. 2.11

2. 3.4 (Teams of up to two allowed—turn in one paper with both names). If your computer will not handle this much data, take a sample. If there are problems with the software, then do 3.2 (a,c) and 3.3.

3. 5.4

10/25/2012v1

Homework 1

1/12

What are HWs like in this class?
Readings seem to be similar as class...

2.11 Toyota Corolla ix ls
"i where i
Got it - book site

a) Use matrix plot
Partition data
Which ones?
explore pairs...
Should flip through book... don't rush!

Partition
how well will it work on new data
partition data so only use training
half or verification half

(2)

Sometimes 3 sets

k-NN requires it (where?)

Linear Regression

Boston Housing data

Categorical values → yes or no if high

look for errors

might want to reduce

XMLine partition into training + val

Linear Regression it

Minimize total sum of square errors

RMS ↗

try on validation data

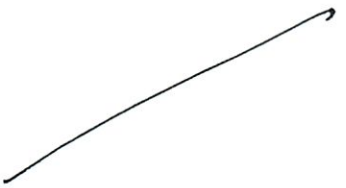
Oh I was using the wrong category!

read instructions wrong

"Scatter plot Matrix"

(3)

What is correlated?



line would be fully correlated

fully neg correlated

fully uncorrelated

Age + km

Price + km

Price + Age

(36)

Variable	Description
Id	Record_ID
Model	Model Description
Price	Offer Price in EUROs
Age_08_04	Age in months as in August 2004
Mfg_Mont	Manufacturing month (1-12)
Mfg_Year	Manufacturing Year
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Color	Color (Blue, Red, Grey, Silver, Black, etc.)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Cylinders	Number of cylinders
Gears	Number of gear positions
Quarterly_	Quarterly road tax in EUROs
Weight	Weight in Kilograms
Mfr_Guara	Within Manufacturer's Guarantee period (Yes=1, No=0)
BOVAG_Gu	BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)
Guarantee_	Guarantee period in months
ABS	Anti-Lock Brake System (Yes=1, No=0)
Airbag_1	Driver_Airbag (Yes=1, No=0)
Airbag_2	Passenger Airbag (Yes=1, No=0)
Airco	Airconditioning (Yes=1, No=0)
Automatic_	Automatic Airconditioning (Yes=1, No=0)
Boardcomp	Boardcomputer (Yes=1, No=0)
CD_Player	CD Player (Yes=1, No=0)
Central_Lo	Central Lock (Yes=1, No=0)
Powered_\	Powered Windows (Yes=1, No=0)
Power_Ste	Power Steering (Yes=1, No=0)
Radio	Radio (Yes=1, No=0)
Mistlamps	Mistlamps (Yes=1, No=0)
Sport_Mod	Sport Model (Yes=1, No=0)
Backseat_[	Backseat Divider (Yes=1, No=0)
Metallic_Ri	Metallic Rim (Yes=1, No=0)
Radio_cass	Radio Cassette (Yes=1, No=0)
Parking_As	Parking assistance system (Yes=1, No=0)
Tow_Bar	Tow Bar (Yes=1, No=0)

Why did color report wrong?

(5)

b) Prepare to use data

ia) Explain how convert Fuel Type + Metallic to binary

c) partition

3-4 | Laptop Sales at London Chain
52 MB txt file

Data visualization

help us find duplicates
and select which variables to use
and which are redundant
or find bin size

explore data

Supervised = training data provided

(6)

Line graphs → good for time series

Box plots + histograms distribution
↑ more than 1 #

Heatmaps color to denote stronger values
or missing data

Categorical → bes w/ hue, shape
or multiple panels
(can't do scatter plot)

+ rellising splitting ~~opese~~ observations
according to categorical value
and create sep plot for each cat

Scatter plot matrix what we saw earlier

Preprocess
rescaling
aggregation + hierarchies
↑ monthly

7

Zooming + panning showing only some parts

trend lines + lables

Ways to deal w/ large data
- sampling
- reducing maker size
- transparent makers
- panels
- aggregate
- jitter - move up

interactive visualization
- Spotfire
- Tableu

Network data

Tree maps
- if hirarchical
- Color
- size

Maps

(8)

a) try to project revenues

figuring out sw!

---

how does map work

Where are post codes?

Mapping not working!

Oh only the last 3!

Stores selling the most

Coordinate high lighting

$$\sqrt{(x-x)^2 (y-y)^2}$$

dimension ind variable

measure dep variable

(I'm not thinking objectively about this!)

(9)

Think my problem was not using cont.

I like Tableu!

No - how to aggregate scatter plot

"Number of Records"

Ahh

So thats how can see what is common!

Still did not get the geo right

but that was since their lat lng scales

Plus no cust and store at same time!

Or a travel "line"

5-4 Classifiers Insurance data

(10)

1% Fraud
Oversample

Correct 310 Miss 90
270 130

Are	\ Claim	F	NF
F		310	90
NF		130	270

Tbl what I naturally did was a qv!

(I'm starting to guess how this works!)

b) adjusted misclassification rate

Oversampling → include a bunch of the fraudulent data

(11)

Misclassification $\frac{95 + 1130}{800} = 27.5\%$

(a)

Can take away 1s or add 0s

Add 0s so 1s are only 2% of sample
There 1%

$400 + .99x = x$

Solve for x

$40,000$

So 39600 0s should be

60

$$\begin{array}{r} 39600 \\ 400 \\ \hline 40000 \end{array}$$

(12)

Bump up by ratio $\frac{400}{3960} = \frac{1}{99}$

$310 \cdot 99 = 3069 0$

$90 \cdot 99 = 8910$

So adjusted $\frac{8910 + 130}{40000}$ 22.6%

What % of new records had last

$440 \cdot 99$

3069 0 8910

~~4~~130 270

I think I did that wrong

(43)

130.99	270.99
12870	26730
F	NF
F 310	90
NF 12870	26730

$$\frac{12960}{40000} \quad 32.4\%$$

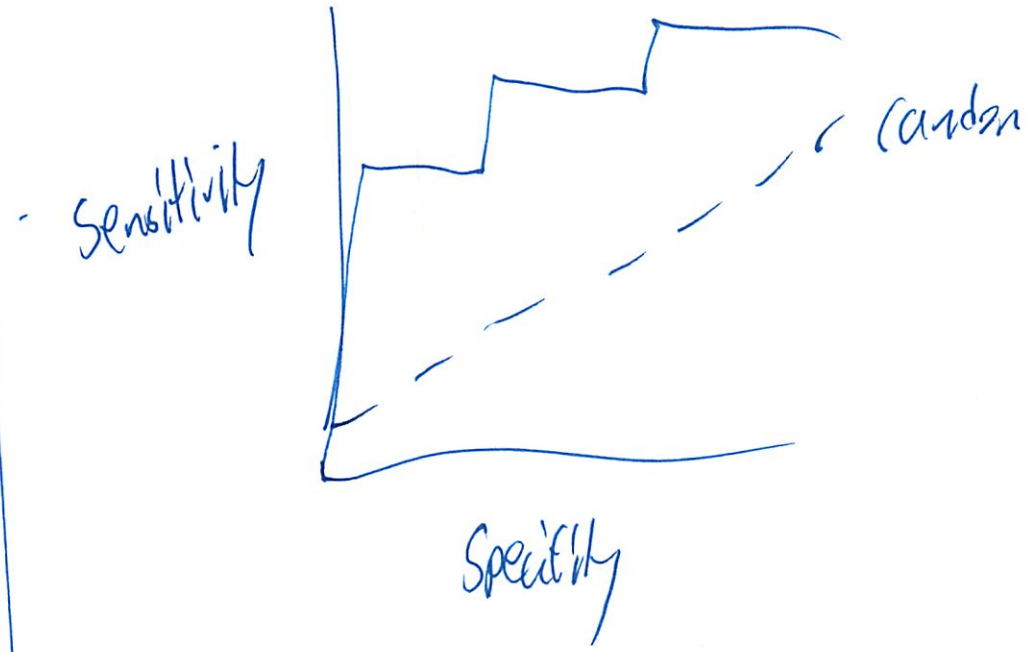
c) $$\frac{90 + 26730}{40000} = 67.05$$

Not going slow + understanding

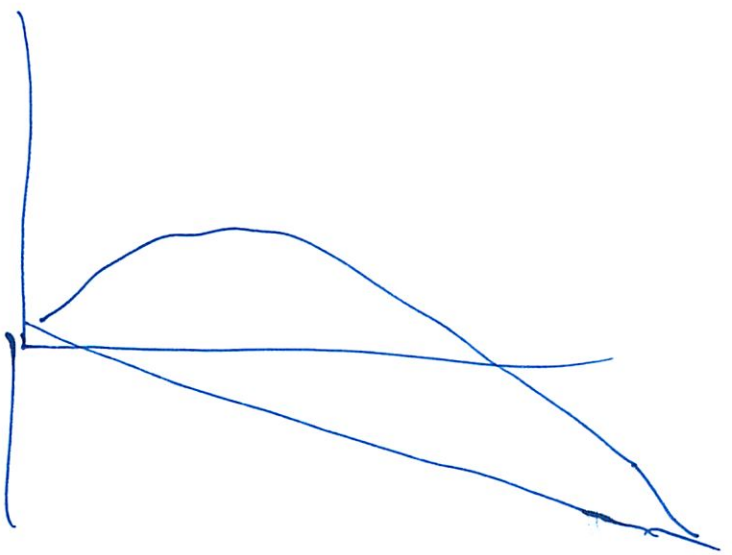
(14)

Lift chart used in direct marketing

What provides the best lift



Can include cost



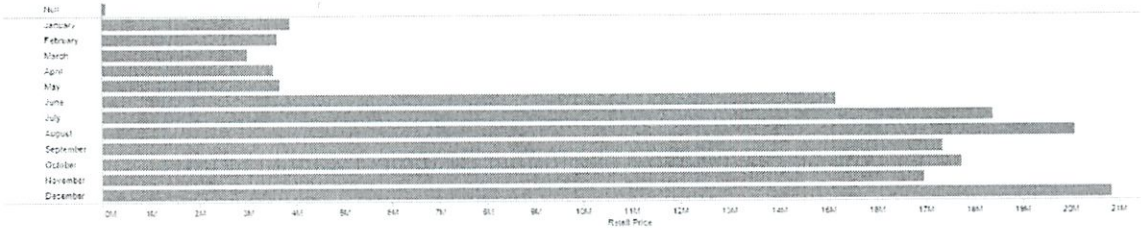
Can use as a cutoff

11/12

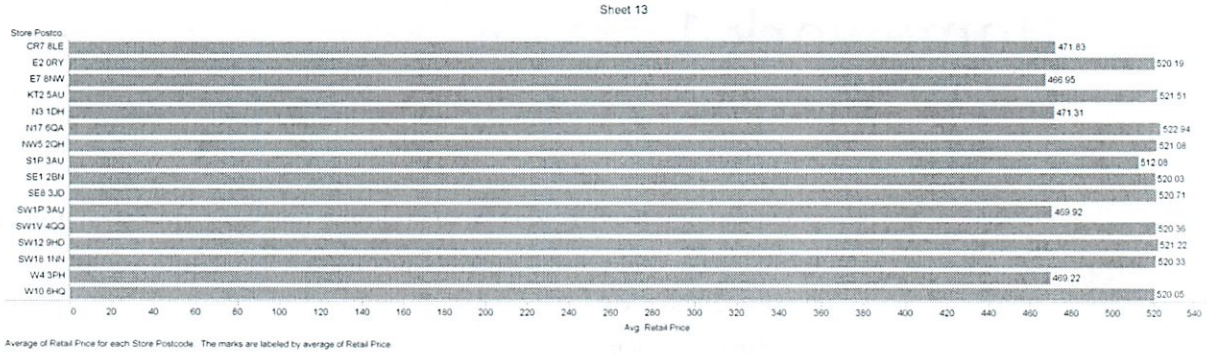
Homework 1

Michael Plasmeier

1. 2-11
    a. From a visual perspective, we can see that the following are correlated. It is difficult to see if the binary values are correlated just from looking at the chart.
        i. Price and KM
        ii. Price and Age
        iii. Age and KM
        iv. Price and HP we can kinda tell
    b. Both Metallic Color and Metallic Rims are already a binary value, meaning they are 1 if present and 0 otherwise. We could convert Fuel Type into two or three binary values. For a regression, we only want two binary values (Petrol Yes/No and Diesel Yes/No). For some other methods we use all three (Petrol Yes/No, Diesel Yes/No, CNG Yes/No).
    c. XLMiner has a function Transform>Transform Categorical Data>Create Dummy Variables which will do this for us.
    d. We would want to make sure we remove the original Categorical Data column. We also want to be careful about when we should or should not use the last dummy variable (as above sometimes we use one less) when we should not.
    e. Partitioning allows us to make sure we are not overfitting. We first run our algorithm on the training data which gives us a model. We then verify our model on the validation data. Finally, to be sure that we have a robust model, we add in new test data. This prevents us from fitting too closely to the verification output, making sure our model is the best generally.
2. 3-4
    a. The average price of a laptop is $508. However, many laptops are sold between $450 and $550.
    b. Laptops are selling mostly in months at the end of the year



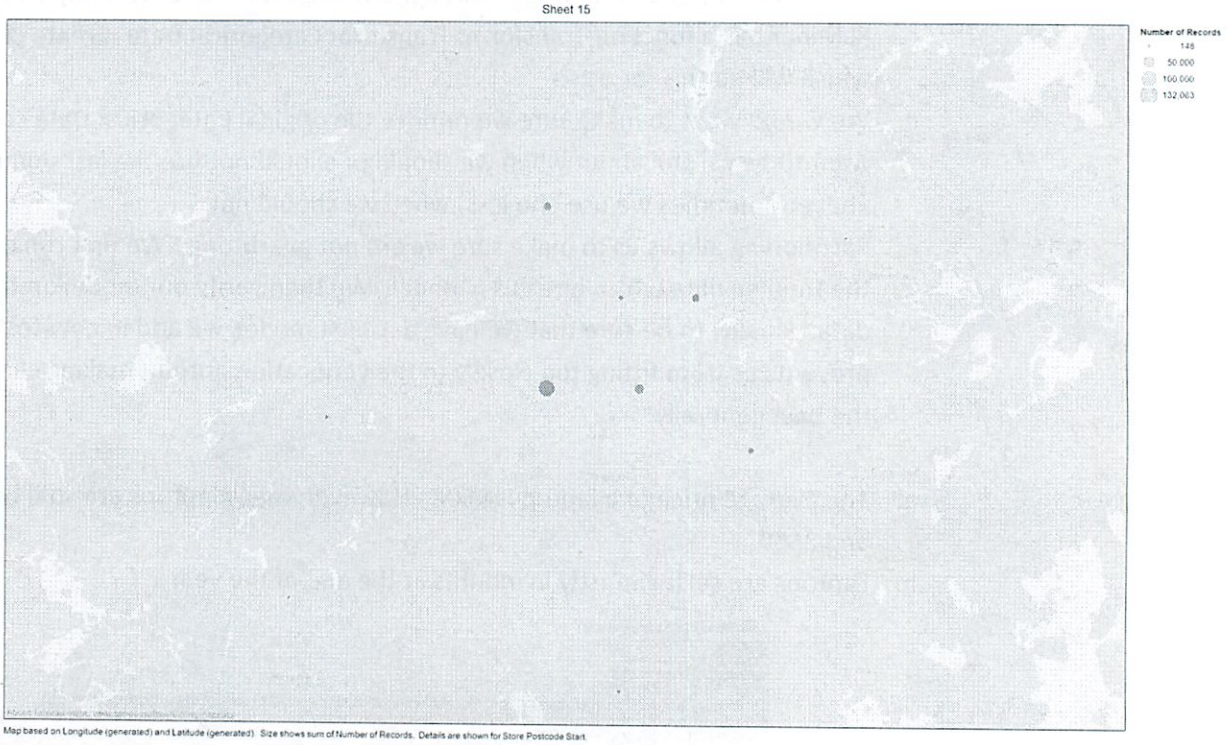
    c. Some postcodes sold machines at cheaper average sale price per unit



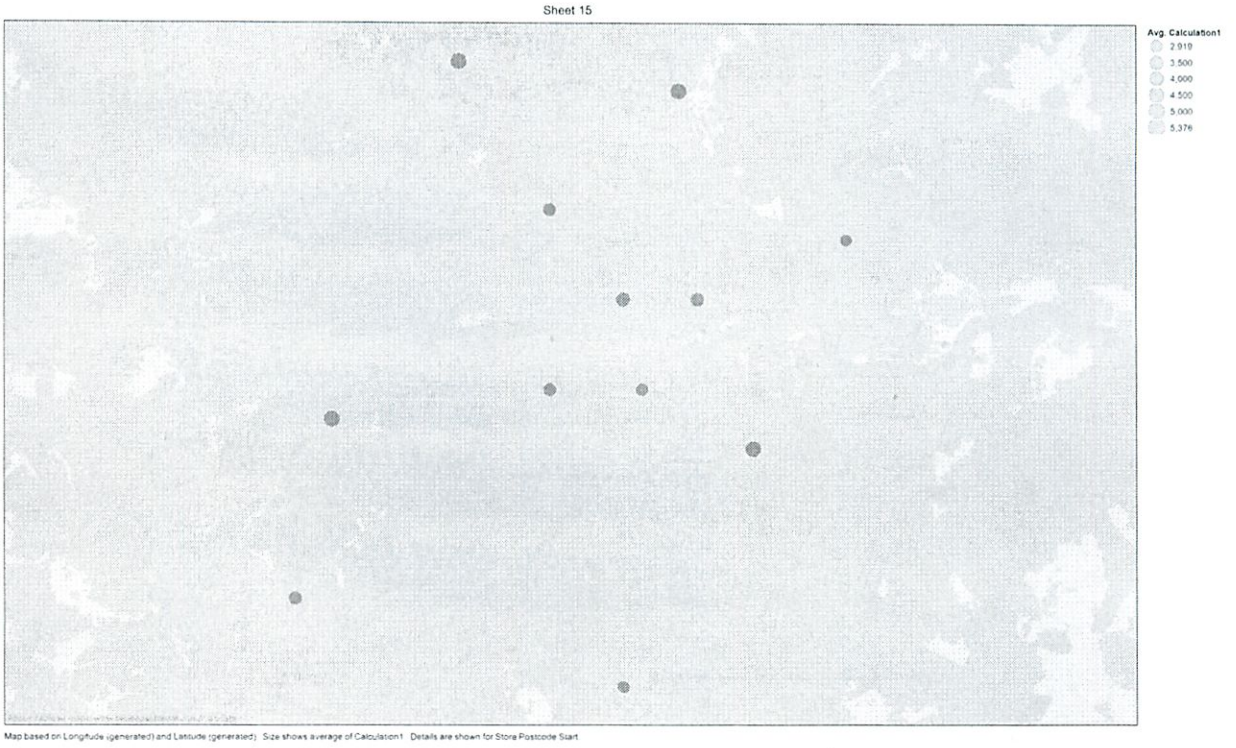
d. The price increases as the processor speed increases. The price increases as HDD space increases.

e. In London.

f. The store at SW1P 3AU sells the most product; generally the stores downtown sell more than the stores in the suburbs.



g. The bigger the circle, the further customers wanted to travel

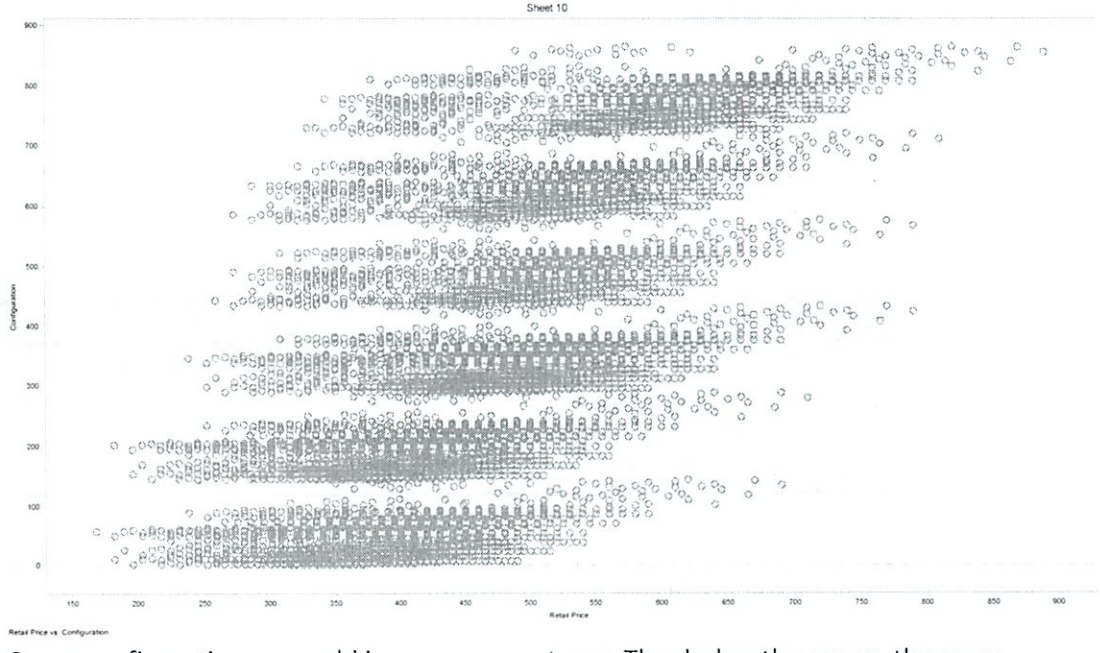


h. The average guest traveled 3,800 units. (I could not tell how your lat/long were formatted)

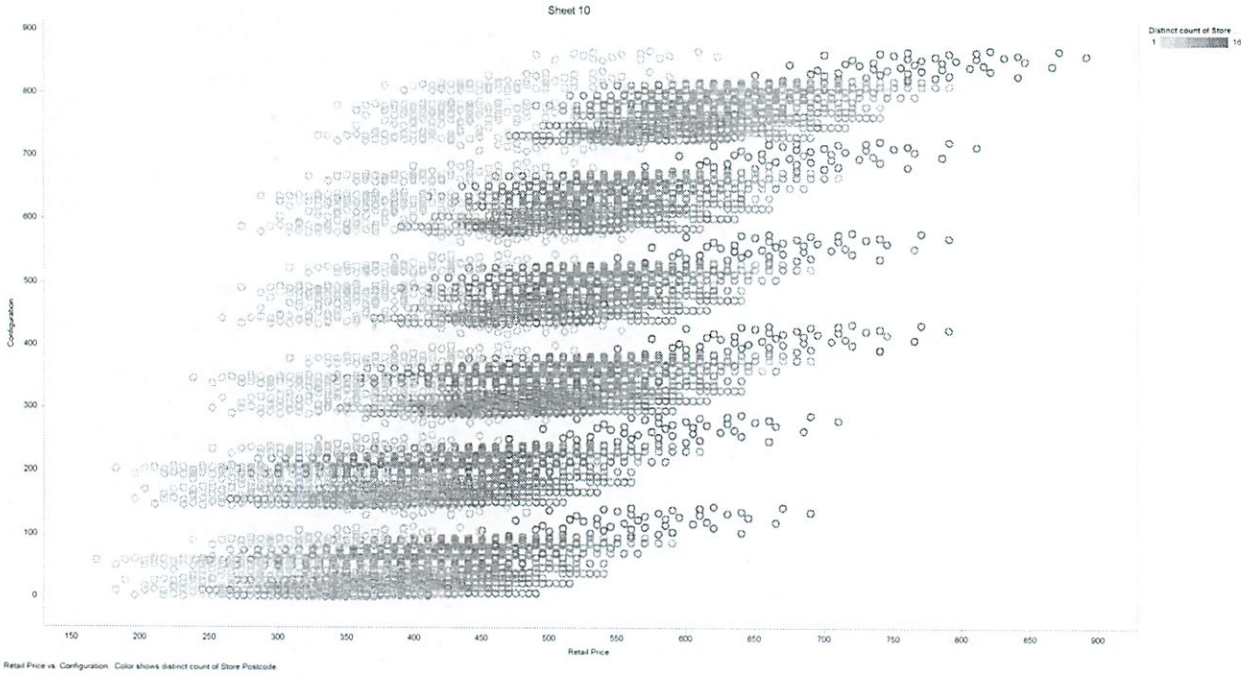
i. There is no data in the data set about revenue (only retail price)

j. There is no data in the data set about revenue (only retail price)

k. There are many configurations, but are 6 main groups. These increase in price as configuration number increases



l. Some configurations are sold in many more stores. The darker the green, the more stores the config is sold in.



3. 5-4

a. Table

		Predicted		
		Fraud	**Not Fraud**	**Total**
Actual	**Fraud**	310	90	400
	Not Fraud	130	270	400
	Total	440	360	800

b. The adjusted misclassification rate is 32.4%

i. Since we have 99% are actually non fraudulent, we need to add new non fraudulent transactions in the same proportion as above. So we add so we have 39,600 non frauds and 400 frauds for 40,000 total records. We multiply 130 and 270 each by 99 to get

		Predicted		
		Fraud	**Not Fraud**	**Total**
Actual	**Fraud**	310	90	400
	Not Fraud	12870	26730	39,600
	Total	13180	26820	40,000

c. 67.05% will be classified as non-fraudulent under this classification

12/15 11/12

Homework 1

Michael Plasmeier

3/5 1. 2-11

a. From a visual perspective, we can see that the following are correlated. It is difficult to see if the binary values are correlated just from looking at the chart.
   i. Price and KM
   ii. Price and Age
   iii. Age and KM
   iv. Price and HP we can kinda tell

↑ where!?

b. Both Metallic Color and Metallic Rims are already a binary value, meaning they are 1 if present and 0 otherwise. We could convert Fuel Type into two or three binary values. For a regression, we only want two binary values (Petrol Yes/No and Diesel Yes/No). For some other methods we use all three (Petrol Yes/No, Diesel Yes/No, CNG Yes/No).

c. XLMiner has a function Transform>Transform Categorical Data>Create Dummy Variables which will do this for us.

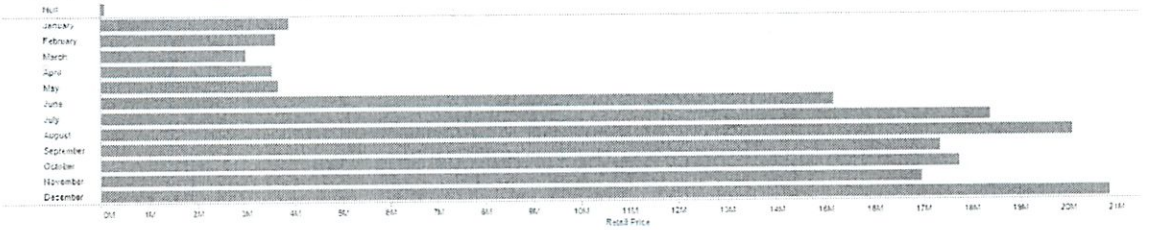
d. We would want to make sure we remove the original Categorical Data column. We also want to be careful about when we should or should not use the last dummy variable (as above sometimes we use one less) when we should not. ?

e. Partitioning allows us to make sure we are not overfitting. We first run our algorithm on the training data which gives us a model. We then verify our model on the validation data. Finally, to be sure that we have a robust model, we add in new test data. This prevents us from fitting too closely to the verification output, making sure our model is the best generally.

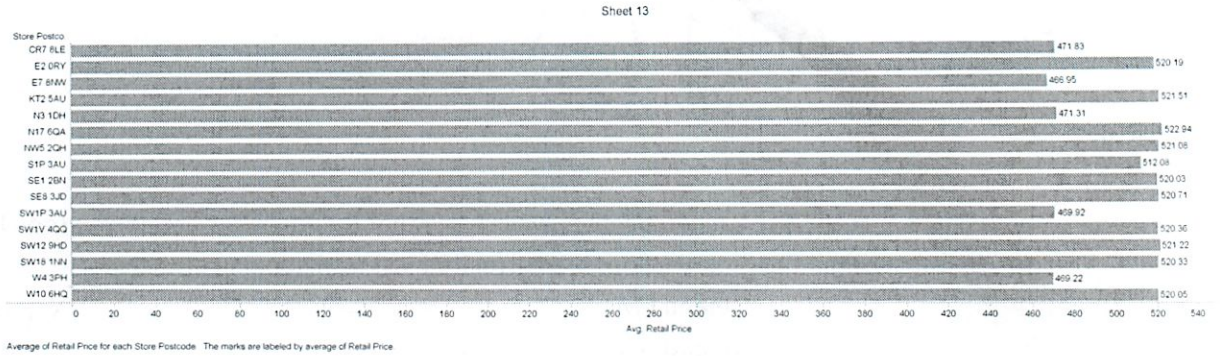
5/5 2. 3-4

a. The average price of a laptop is $508. However, many laptops are sold between $450 and $550.

b. Laptops are selling mostly in months at the end of the year



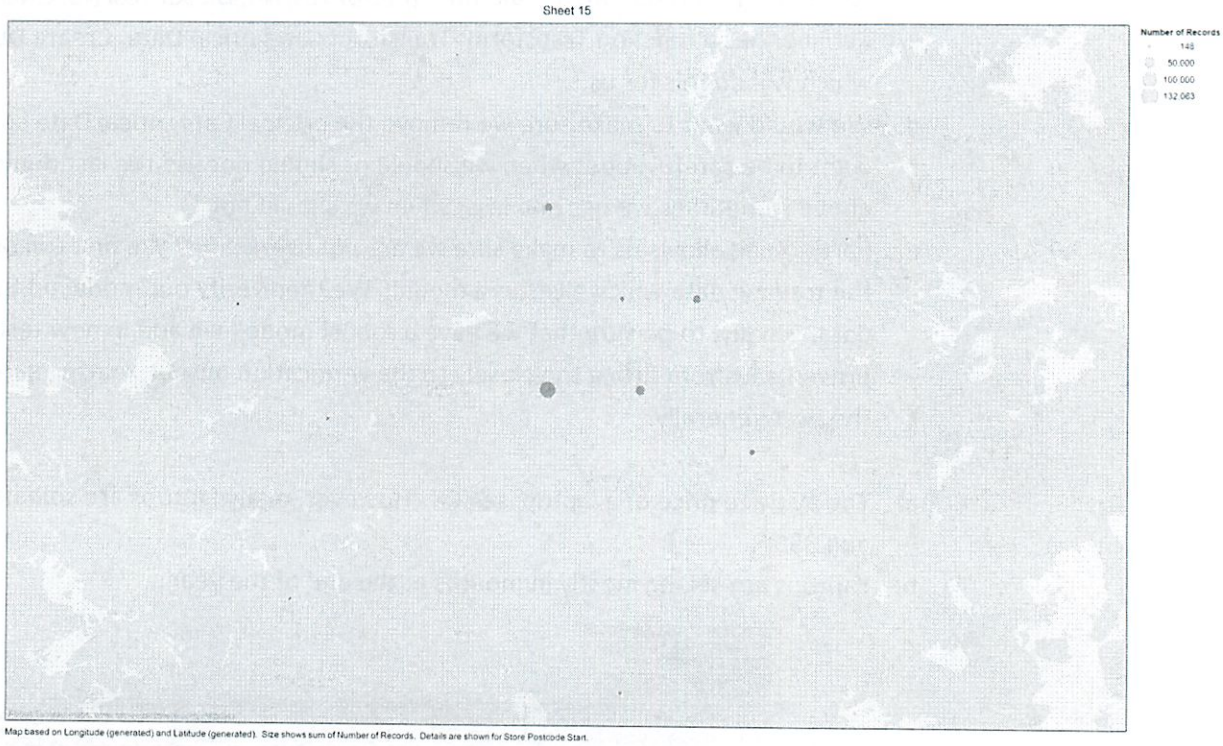
c. Some postcodes sold machines at cheaper average sale price per unit



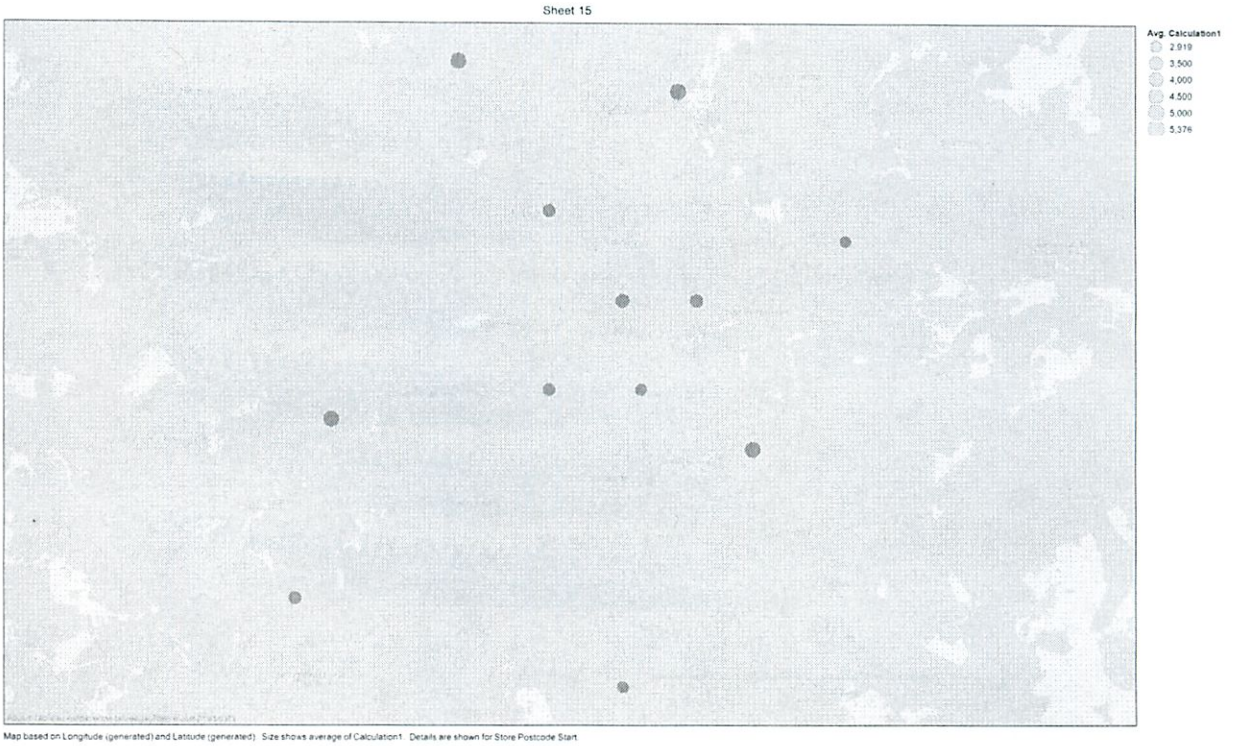
d. The price increases as the processor speed increases. The price increases as HDD space increases.

e. In London.

f. The store at SW1P 3AU sells the most product; generally the stores downtown sell more than the stores in the suburbs.



g. The bigger the circle, the further customers wanted to travel

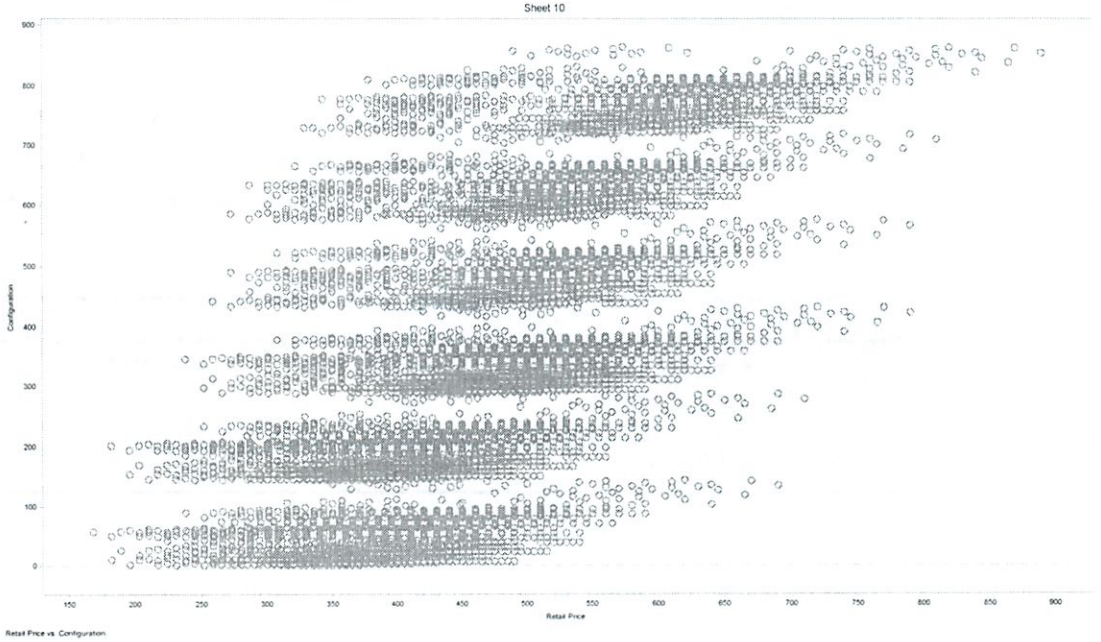


h. The average guest traveled 3,800 units. (I could not tell how your lat/long were formatted)

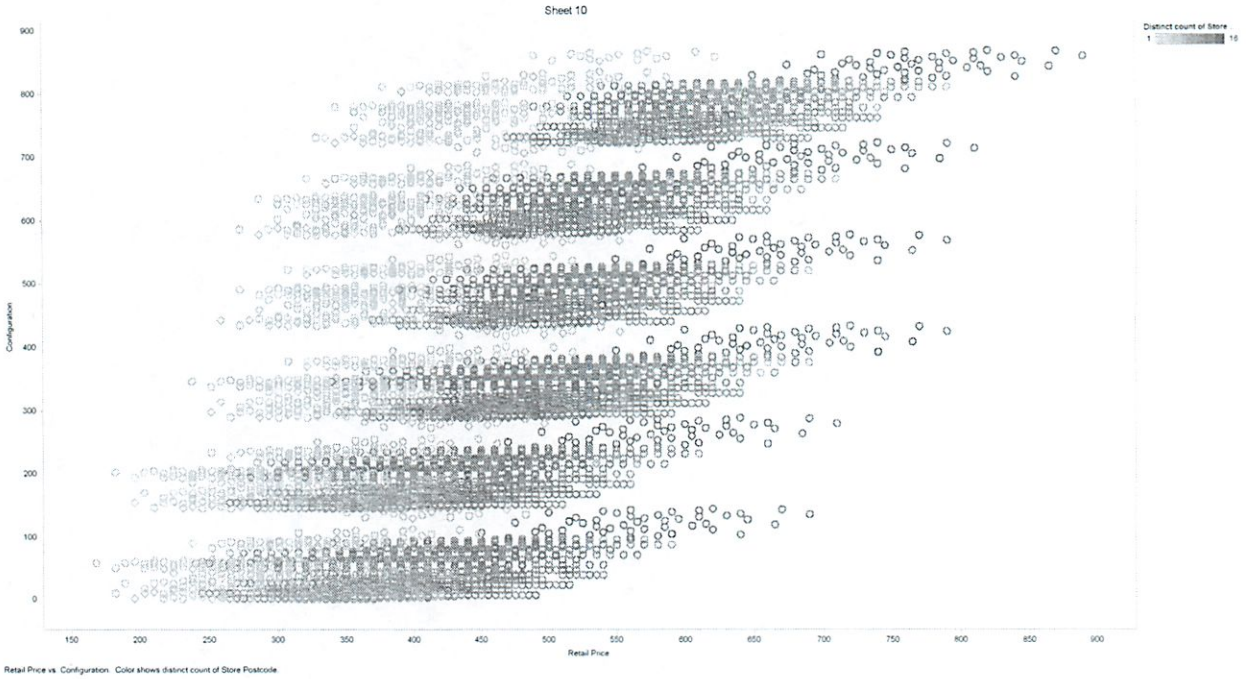
i. There is no data in the data set about revenue (only retail price)

j. There is no data in the data set about revenue (only retail price)

k. There are many configurations, but are 6 main groups. These increase in price as configuration number increases



l. Some configurations are sold in many more stores. The darker the green, the more stores the config is sold in.



4/5 3. 5-4

a. Table

		Predicted		
		Fraud	**Not Fraud**	**Total**
Actual	**Fraud**	310	90	400
	Not Fraud	130	270	400
	Total	440	360	800

b. The adjusted misclassification rate is 32.4%

i. Since we have 99% are actually non fraudulent, we need to add new non fraudulent transactions in the same proportion as above. So we add so we have 39,600 non frauds and 400 frauds for 40,000 total records. We multiply 130 and 270 each by 99 to get

		Predicted		
		Fraud	**Not Fraud**	**Total**
Actual	**Fraud**	310	90	400
	Not Fraud	12870	26730	39,600
	Total	13180	26820	40,000

c. 67.05% will be classified as non-fraudulent under this classification

↑ did not answer the question!

Fall 2012 Data Mining: Finding the Data and Models that Create Value 15.062 (ESD.754J) (Welsch)

Homework #2

Due: Monday, November 26, 2012

Reading:

DMBI Chapters 6-10 (not at much as it looks and regression review is included).

Problems (individual work unless otherwise noted)**:**

1. 7.1

2. 10.1

3. **Case** (up to two may work on together and submit one write-up):

German Credit case at the end of the book (18.2). Use the following methods on these data: k-NN, naive Bayes, classification trees, and logistic regression. By using one data set, I am hoping to keep the data manipulation time to a minimum. However, I would like for you to compare and contrast the results you obtained using the different methods. To do this please modify part 2 of this case and divide the data into training, validation, and test data sets as follows: Train with 600, validate with 200, and test with 200. Please also let us know what you think your best model is. We may pick a random test set to compare the final models suggested by each of you. Save your files since we will use neural nets and discriminant analysis on this data set in the next homework assignment.

11/7/2012v1

Homework 2

11/12

7-1 | Personal Loan Acceptance

Wants to turn depositors into loaners

9% conversion rate

k = 1

Use data utils → score from stored model

Predicts 0

b) Pick k

15

c) Classification matrix

Copied

d) Prediction still 0

(7)

e) Repartition

Compare test set

10-1 Financial Condition of Banks

Banks, x/s

Financial condition of bank

Run a logistic regression

Sucess = weak

(So not Total Cap/Assets

Logistic Regression

like linear regression

Uses prediction variables

Variable selection algorithms

Y = categorical

(3)

Use to classify a new observation

or profile factors that differ b/w the two

Categorical or Continuous

Prob of belonging to each class

$P(Y=1) \geq 0.5$ belong to class 1

$p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots)}}$$

↑ logistic response fn

$$\text{odds} = \frac{p}{1-p}$$

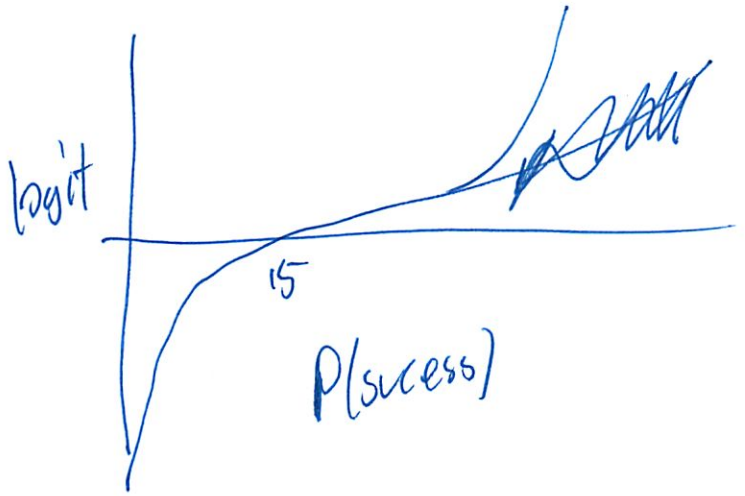
· if $p(\text{winning}) = .5$, odds $= 1$

$$p = \frac{\text{odds}}{1 + \text{odds}}$$

$$\text{odds} = e^{\beta_0 + \beta_1 x_1 + \cdots}$$

(9)

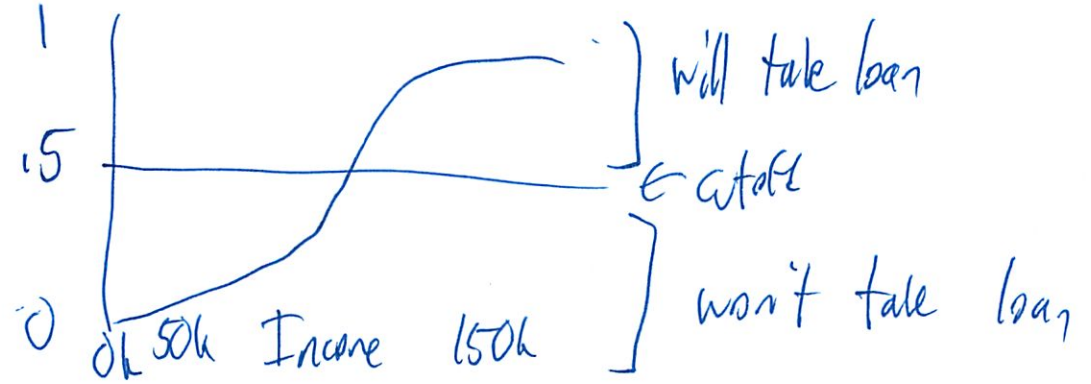
$$\underbrace{\log(\text{odds})}_{\text{logit}} = \beta_0 + \beta_1 x_1 + \cdots$$



Single Predictor

$$P(\text{Personal Loan} = \text{Yes} \mid \text{Income} = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

estimate coeffs from training
then try in verification



(5)

Note → Y, B non linear

So no least squares
instead maximum likelihood

So for each coeff, st err, p-value, odds
↳ what is this?

0 pvalue means likely to take
↳ why?

$\frac{.5280}{1-.5280}$

odds $\frac{p}{1-p}$

So .0416

= .0434

no that is not consistent w/ odds ... hmm

$P = \frac{1}{1+e^{-(B_0+B_1x)}}$

I don't get it!

c) Threshold
often .5
or maximize overall accuracy
w/ one way data table
- it tries all of them
or give costs

(6)

But how do we base this on the odds?

Used to ↑ precision (correct is return correct)

↓ recall (correct → incorrect)

d) Skip

e) Got

---

3. Case

Compare w/ a bunch

Prior app for credit

Rated good or bad

Try new ones

k-NN

Codelist

Var. #	Variable Name	Description	Variable Type	Code Description
1.	OBS#	Observation No.	Categorical	
2.	CHK_ACCT	Checking account status	Categorical	0 : < 0 DM 1: 0 < ...< 200 DM 2 : => 200 DM 3: no checking account
3.	DURATION	Duration of credit in months	Numerical	
4.	HISTORY	Credit history	Categorical	0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly till now 3: delay in paying off in the past 4: critical account
5.	NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
6.	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
7.	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
8.	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes
9.	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
10.	RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes
11.	AMOUNT	Credit amount	Numerical	
12.	SAV_ACCT	Average balance in savings account	Categorical	0 : < 100 DM 1 : 100<= ... < 500 DM 2 : 500<= ... < 1000 DM 3 : =>1000 DM 4 : unknown/ no savings account
13.	EMPLOYMENT	Present employment since	Categorical	0 : unemployed 1: < 1 year 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years 4 : >= 7 years
14.	INSTALL_RATE	Installment rate as % of disposable income	Numerical	
15.	MALE_DIV	Applicant is male and divorced	Binary	0: No, 1: Yes
16.	MALE_SINGLE	Applicant is male and single	Binary	0: No, 1: Yes
17.	MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
18.	CO-APPLICANT	Application has a co-applicant	Binary	0: No, 1: Yes
19.	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
20.	PRESENT_RESIDENT	Present resident since - years	Categorical	0: <= 1 year 1<...<=2 years 2<...<=3 years 3:>4years
21.	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes
22.	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
23.	AGE	Age in years	Numerical	
24.	OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes
25.	RENT	Applicant rents	Binary	0: No, 1: Yes
26.	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes
27.	NUM_CREDITS	Number of existing credits at this bank	Numerical	
28.	JOB	Nature of job	Categorical	0 : unemployed/ unskilled - non-resident 1 : unskilled - resident 2 : skilled employee / official 3 : management/ self-employed/highly qualified employee/ officer
29.	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30.	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
31.	FOREIGN	Foreign worker	Binary	0: No, 1: Yes
32	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

(8)

2c) I think I get it now

Luse plias

lO weak

lO strong

---

reacts extremly to subtile changes ←

or reacts subtily to extream cases

(9)

3. Case

1 = good credit

0 = bad credit

I do a crappy job

; Too many

I didn't use cost either...

; then include that

; then sort by predicted prob of sucess

---

how do cost;

cutoff?

ratio of costs

$$\frac{n_{01}}{n_{00}+n_{01}} * \frac{n_{00}+n_{01}}{n} q_0 + \frac{n_{10}}{n_{10}+n_{11}} \frac{n_{10}+n_{11}}{n} q_1$$

q_0 = cost of says 0 belongs to 1

q_1 1 0

(10)

saying
So q_0 = bad belongs good = ~~##~~500
q_1 good bad = ~~##~~100) costs

Want to minimize

$$\frac{q_1}{q_0} = \frac{100}{500}$$

Use priors $\frac{p(C_0)}{p(C_1)} = \frac{300}{700}$

"So what" Use .8 as cutoff

Sort based on prob of sucess

"So .6 prob of sucess cutoff
Since a priori
(this is a cool method - next wealthiest cust!)
What I had set before did not matter
I just looked at prob

11/12

Homework 2

Michael Plasmeier

1. 7-1
    a. It predicts that customer will not open an account.
    b. The best k is 15.
    c. See table:

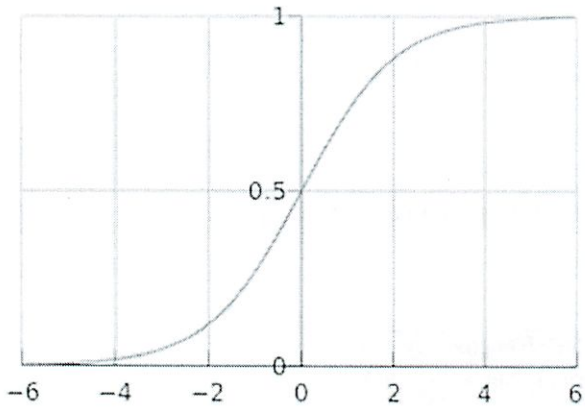
Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	44	127
0	53	1776

    d. It still predicts the customer will not open a new account
    e. The overall percentage of errors is lower than the validation data, but higher than the training data. We would expect the training data to fit well. However, we might expect the test error to be higher than the validation error, since we have been adjusting the model with the validation data. However, this is not true, meaning our data likely fits the model well.
2. 10-1
    a. The logit is -14.18 + 79.96 (TotExp/Assets) + 9.17 (TotLns&Lses/Assets)
        i. The odds of being weak are e^(-14.18)*(*[1])Exp*(9635.41)Lns
        ii. The probability of being weak is odds/(1+odds)

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-14.1875391	6.12205267	0.02047934	*
TotExp/Assets	79.96391296	39.26251602	0.04168537	*
TotLns&Lses/Assets	9.17319965	6.86388016	0.1814038	9635.410156

    b. The bank is found to be weak. The probability of being weak is .5280, the logit is .1124, the odds are 1.1186
    c. The threshold is the same since we have 10 strong banks and 10 weak banks so banks have a .5 *a priori* chance of being weak.
    d. A high Loans and Leases to Assets ratios correlates well with being weak. We can confirm this since weak banks have a higher average Loans and Leases to asset ratio than strong banks. The coefficient represents the change in the logit for each unit change in the predictor. Since the logit is logarithmic, it reacts strongly to small changes.

[1] I don't understand why this value was undefined



e. Increasing our classification cutoff value increases precision but decreases recall (correct items that are incorrect). So to prevent strong banks from being marked as weak, we should raise our cutoff.

3. Case

We can run our case with different methods. However, they all do somewhat poorly in predicting the final error. This is running with all of the variables, and without oversampling, and without adjusting for the cost of misclassifying.

a. Having a used car, being foreign, being divorced, being unemployed, and having a large amount of money in your savings meant you were likely to default. The having a lot of money in your savings was surprising.

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	6.30225992	1.46662378	0.0000173	*
CHK_ACCT_0	-2.29585409	0.33491024	0	0.10067537
CHK_ACCT_1	-1.63414514	0.31551567	0.00000022	0.1951191
CHK_ACCT_2	-1.54444897	0.49715993	0.00189289	0.21342945
DURATION	-0.02587759	0.01276775	0.04268355	0.97445434
HISTORY_0	-1.33374071	0.61145669	0.02916484	0.26348978
HISTORY_1	-1.51970685	0.57911003	0.008685	0.21877603
HISTORY_2	-0.68554056	0.33868772	0.04295903	0.50381786
HISTORY_3	-0.18410464	0.45589614	0.68633759	0.83184874
NEW_CAR	-0.88353533	0.5033834	0.0792262	0.41331911
USED_CAR	0.95051974	0.63732195	0.13584919	2.58705378
FURNITURE	0.18004292	0.52236569	0.7303437	1.19726872
RADIO/TV	0.07831579	0.49521932	0.87434363	1.08146417
EDUCATION	-0.9208132	0.64983028	0.1564813	0.39819512
RETRAINING	-0.57101339	0.60442251	0.34479898	0.56495261
AMOUNT	-0.00017915	0.00006107	0.00335154	0.99982089
SAV_ACCT_0	-0.49036461	0.34389061	0.15388799	0.61240304
SAV_ACCT_1	0.18633856	0.46812138	0.69058889	1.20483005

SAV_ACCT_2	-0.53209227	0.57529843	0.35501906	0.58737475
SAV_ACCT_3	1.15361702	0.80781066	0.15326928	3.16963673
EMPLOYMENT_0	-0.44411018	0.55798197	0.42607731	0.64139473
EMPLOYMENT_1	-0.49783039	0.37889427	0.18887877	0.60784805
EMPLOYMENT_2	-0.23545948	0.32984331	0.4753184	0.79020768
EMPLOYMENT_3	0.66987807	0.4155544	0.10695966	1.95399904
INSTALL_RATE	-0.39909413	0.12076842	0.00095104	0.67092752
MALE_DIV	0.84928334	0.58631617	0.14747518	2.33797073
MALE_SINGLE	0.46977305	0.26900393	0.08075141	1.59963119
MALE_MAR_or_WID	0.52624226	0.4378055	0.22936352	1.69256008
CO-APPLICANT	0.63995582	0.62988758	0.30963692	1.89639711
GUARANTOR	0.69741499	0.50932699	0.17090948	2.00855398
PRESENT_RESIDENT_1	0.42877972	0.41927075	0.30645928	1.53538275
PRESENT_RESIDENT_2	-0.76593876	0.30299652	0.01147536	0.4648973
PRESENT_RESIDENT_3	-0.45220718	0.3728238	0.22515862	0.63622236
REAL_ESTATE	0.10405345	0.28950137	0.71927869	1.10965979
PROP_UNKN_NONE	-0.51981235	0.53935426	0.33516231	0.59463215
AGE	0.00587227	0.01183509	0.61977094	1.00588953
OTHER_INSTALL	-0.6189636	0.28184229	0.02808231	0.53850228
RENT	-0.83508736	0.63786721	0.19047107	0.43383658
OWN_RES	-0.26296562	0.61249375	0.66767871	0.76876831
NUM_CREDITS	-0.38896427	0.24698998	0.11529846	0.67775846
JOB_0	1.4665575	1.04517829	0.16056766	4.3342886
JOB_1	-0.31274748	0.47036251	0.50610983	0.73143458
JOB_2	-0.2788668	0.39473286	0.4798961	0.75664067
NUM_DEPENDENTS	0.19530804	0.33370745	0.55836803	1.21568537
TELEPHONE	0.15822937	0.26449308	0.54968226	1.17143488
FOREIGN	3.14732909	1.41494763	0.02612536	23.27381897

Without cost control

b. k-NN with k = 10

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
1	138	8	5.80
0	62	57	91.94
Overall	200	65	32.50

c. Naive Bayes

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
1	138	27	19.57
0	62	31	50.00
Overall	200	58	29.00

d. Classification Tree without Pruning (Pruning removed the whole tree)

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
1	138	17	12.32

0	62	36	58.06
Overall	200	53	26.50

e. Logistic Regression

Error Report			
Class	# Cases	# Errors	% Error
1	138	22	15.94
0	62	36	58.06
Overall	200	58	29.00

f. All of the methods did pretty poorly, but Classification Trees without Pruning seems to be the best. It looks pretty hard to pick out the factors that lead to credit approval.

Now use .8 as the cutoff to indicate the higher cost of marking a bad credit risk customer as good:

g. 10-Nearest Neighbor

Error Report			
Class	# Cases	# Errors	% Error
1	138	80	57.97
0	62	17	27.42
Overall	200	97	48.50

But we must now calculate the cost for this.

80*100+17*500=$16,500

Before:

8*100+57*500=$29,300

h. Naive Bayes

Error Report			
Class	# Cases	# Errors	% Error
1	138	54	39.13
0	62	18	29.03
Overall	200	72	36.00

54*100+18*500=$14,400

i. Best Pruned Tree

Fully prunes

j. Non pruned

Error Report			
Class	# Cases	# Errors	% Error
1	138	66	47.83
0	62	13	20.97

Overall	200	79	39.50

66*100+13*500=13,300

k. Logistic

Would not run: *Number of rows is less than number of columns. Regression computation failed!*

l. So the least costly for the bank is still the non-pruned classification tree.
m. Doing this for non-cost data since Logistic Regression worked there.
n. You go 410 rows in to a 1000 row data set. You make a profit of $21904.
o. This means you should allow .6 as the probability of success cutoff.

11/12
11:40

Study

Sensitivity the classify correct as correct

$$\frac{n_{11}}{n_{10} + n_{11}}$$

n_{10} called correct, n_{11} actually correct

Specificity rule out 0 correctly

$$\frac{n_{00}}{n_{00} + n_{01}}$$

Lift charts (why is largest one not 1st?)

Rich checks them in serial I think
or most probable ones 1st?

decile chart 10% at a time
the 1st 10% are most probable

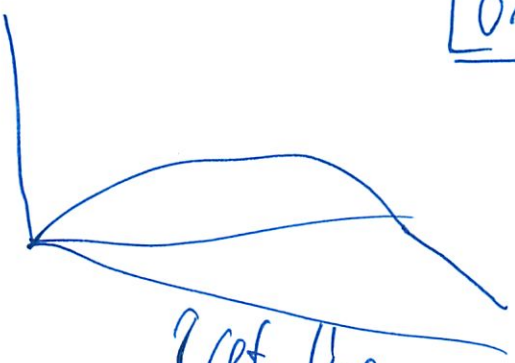
(2)

implicit is that one class or another is better/worse...

Can do w/ cost
↳ avg misclassification cost
minimize this
both sides

oversampling
including more data
in ratio of costs ideally
or sample each =ly

lift curve



ref line

(# records, total net benefit)

Logistic Regression

Ah if $p() > .5$ then predict true

$$p = \frac{e^{logit}}{1 + e^{logit}}$$

$$logit = \beta_0 + \beta_1 x_1 + \dots$$

(3)

Cutoffs selected during validation

w/ maximum likhood

estimate that maximizes chance of obtaining the data we do have

vp

Usually categorically dep variable and several continuous ind variables

converts dep variables to prob scores

How is linear regression actually found?

line of best fit

Some matrix thing

3 matrix multiplications and an inversion

beyond scope of this class

Std error

High when collinearity

Could not find exact meaning

just lower = better

(4) Then add stat significant

~~Then~~ confidence interval right,

$p < .001$ the astorix

~~$p > |z|$~~

which has nothing to do w/ p-value... I think

~~*~~ The std error is that $p < .001$ thing

astrix

~~under confidence level~~

and confidence level

Std error take std dev of #s

$\sqrt{\#\# \text{ in set}}$

Confidence interval ya chose

95% is $\pm$ 1.96 * Std error

(5)

So 100 widgets

mean = 10

std dev = 2

$$\text{Std error} = \frac{2}{\sqrt{100}} = .2$$

95% CI : $\pm 1.96 \cdot .2 = \pm .39$

So std error comes from just the values and how close they are?

Not about predictor?

So the values of 95% CI give the values at that

I guess $16 \pm .39$ as above ...

(6)

So margin of error,

Std error $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p-p^2}{n}}$ * when chosen randomly

? std dev of percentage

So here $p = .47$

$n = 1013$

So std error = 1.6%

We don't know the "true" percentage

inside ~~some~~ this w/ some level of confidence

So percentage ± margin of error is CI

1 std error = 68% CI

1.96 std errors = 95% CI

(7)

So result w/ 95% confidence is

$.47 \pm 1.6 \cdot 1.96$

$.47 \pm 3.13$

So can be bw 43.87 and 50.13
w/ 95% confident

On tv they say Kerry 47%
fine print ±3.1% error

One can say w/ 95% ±4 percentage pt margin of
Gallup cofidence ↗ max error

Std error = $\frac{\text{std dev}}{\sqrt{\#\text{ individuals}}}$

but assumes normal dist
↳ which is usually good right...

8

Central limit large # of random variables ind drawn from same dist is approx normal

(don't go into this now
- should revisit at some point...)

Data Roles; Dimension vs. Measure

Dimensions

Dimensions typically produce headers when added to the rows or columns shelves in the view. By default, Tableau treats any field containing qualitative, categorical information as a dimension. This includes, for instance, any field with text or dates values. However, in relational data sources, the actual definition of a dimension is slightly more complex. A dimension is a field that can be considered an independent variable.

This means that a measure can be aggregated for each value of the dimension. For instance, you might calculate the Sum of "Sales" for every "State". In this case the State field is acting as a dimension because you want to aggregate sales for each state. The values of Sales are dependent on the State, so State is an independent field and Sales is a dependent field.

Such aggregation could also be computed for numeric fields that are treated as dimensions. For instance, you might want to calculate the SUM of Sales for each "Discount Rate" offered to customers. In this case the Discount Rate field acts as an independent field and the Sales field is dependent even though both fields are numeric. You can use a numeric field as the independent field by first converting the Discount Rate measure to a dimension.

Measures

Measures typically produce axes when added to the rows or columns shelves. By default, Tableau treats any field containing numeric (quantitative) information as a measure. However, in relational data sources, the actual definition of a measure is slightly more complex. A measure is a field that is a dependent variable; that is, its value is a function of one or more dimensions.

This means that a measure is a function of other dimensions placed on the worksheet. For instance, you might calculate the Sum of "Sales" for every "State". In this case, the Sales field is acting as a measure because you want to aggregate the field for each state. But measures could also result in a non-numeric result. For instance, you might create a calculated measure called "Sales Rating" that results in the word "Good" if sales are good and "Bad" otherwise. In this case the "Sales Rating" field acts as a measure even though it produces a non-numeric result. It is considered a measure because it is a function of the dimensions in the view.

Parent topic: Data Roles

(10)

Nieve Bayes

1. Find records like it
2. See what they belong to
3. Assign it

* What is the estimated prob of being in a class of interest

Sliding cutoff to classifing it as i

Establish cutoff

Find P(record belongs to class)

If above cutoff → assign it

esp good for categorical

(11)

Example: truthful + fraudulent customers

Exact

$P(\text{frad} \mid \text{prior legal})$

but would assign all to non-fraud

So use cutoff prob method

$$P(C_i \mid X_1, \ldots) = \frac{P(X \ldots \mid C_i) P(C_i)}{P(X \ldots \mid C_1) P(C_1) + P(\text{not } C_1) P(C_n)}$$

~~but need covalent~~ covalent problem

but need an exact match

that might not happen

esp as add variables!

(17)

Niave

1. For Class 1 find ind prob each predictor in the record ~~of~~ $(x_1 \dots x_p)$ occurs in class 1
2. Multiply these by each other * proportion in class 1
3. Repeat 1 + 2
4. Estimate $\frac{p(\text{class } i)}{\sum \text{values for all classes}}$
5. Assign record to class w/ highest p value

From WP: So get mean and variance of each?

Whatever - care what output

but what does our ps tell us

Classification for if fraudulent

ie $p(\text{fraud} \mid \text{prior legal} = y, \text{size} = \text{large}) = .87$ over .5 ✓ fraud

$p(\text{"} \mid \text{"} = n, \text{size} = \text{large}) = .31$ ✗ not fraud

(13)

So its the prob that is true
and is a fraud

And since we don't know that exact
we multiply together

$P($ Prior legal = y among all fraud cos $)$ $P($ size = large among all fraud cos $)$ $P($ is fraud among all $)$

$P($ y, large is fraud $)$ + $P($ y, large is truthful $)$

So each ind

Train on training data

Then test...

Calc prob message is spam by P(each word)
Anything above cutoff is spam

(14)

Classification Tree Pruning

CART
Uses validation tree

So actually removing params should not help
- it would ignore them anyway
- most methods at least ...

tries to lop off subtree to see how misclassification happens in validation data set

penalty function α

Want min error on validation data

15.062

11/14

5 Regression + Classification Trees
6 Regression + Variable Selection

(7 min late)

Entropy - like from com theory

Error

Diff metrics for classifying error

Want to be on edges
└ high priority

Misclassification has corner
algs don't like

So Gini used

Gini + Entropy are comly used to grow trees
Favs less on misclassification
More likly to create pure nodes

(2)

Bernoli variance
for each j
Summing over all j gives us Bernoli

Sum up binomial variance
Try to ↓ variance

Example: Own/not own a lawnmower

So 1st split
Seeking purity
- Ls few missclassifications

But all the way → all pure
Ls overfitting!

Difficulties

Over + Under fit
Missing values - can't process tree
Cost of classification

(3) (He didn't ans if exponential)
Note is Greedy
And lots of optimizations

But we want fewer end nodes

So prune
if validation set improves
then make it an end node

CART pruning
$L(\alpha)$ = # leaves on tree
Tradeoff α mistakes vs. complexity
So chose α based on validation data
↳ mechanism for trading off complexity

(9)

Can look at prune log

min error prune

but best prune looks at tradeoff of error vs #
of nodes

Not cast in concrete choices

Many backoff 1 std error

Spam Example

10-fold cross validation

break training data into 10 sets randomly

train on 90%, validate 10%

do 10 times - one for each

so get 10 misclassify error

Can find std error

(5)

Must build a model of emails

↳ ie text file w/ just text

Can make better w/ random forest

will build 100 trees

Then majority rules

Good + Bad of trees

Sensitive to small

bagging bootstrap aggregation

Sample w/ replacement

Some repeats, Some missing

And have a lot of trees

Perturbing data through bootstrapping

(6)

Lec 6 Regression

How to select variables?

Multiple comparison

↳ lots of tests on data w type I error

↳ 5%

So end up w/ model that is nonsense

Want to predict + understand relationship btw factors

Then see if model makes sense.

Or do we not care if model fits if it predicts?

(7)

Simple linear data

If bootstrap data could change
review this

ε includes everything else

Note $E(Y_i | X_i) = B_0 + B_1 X_1$
given by

Error vertically is only 1 possible choice
can do ⊥ perpendicular → TOTL
Ti missed

Must plot data + look at
to understand what dealing w/

Noise is ind - not true w/ time series
that is difficult w/ data mining
even w/ boot strapping
but can't stir up if time gives

8

Homoskedasticity = same σ

Normal
↳ Prof: but not all data is normally distributed

Example: Questionaire for supervisors

Coeffs = $\hat{\beta}$'s

Std error = uncertainty

t stat = $\frac{\text{coeff}}{\text{st error}}$

p value = if coeff is 0 vs 1
level of significance
low → reject hyp coeff is 0
.026

Collinearity → tells us same thing
A way to do variable significance

(p)

p-value tells us strength of evidence

depends - how you want to structure significance

(I misread this last time)

5% → like 95% CI

Analyzing output

St error of coeffs

Coeffs of Determination

fitting hiere model

$$Y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = TSS$$

= total sum of squares

$$R^2 = 1 - \frac{SSR}{TSS}$$

TSS is fixed

SSR comes from model

(10)

SSR can go down, not up
But you can overfit!

How high depends on your field
High means along line

Slope 0 → explanatory variable has no response

Can make high w/ more xs

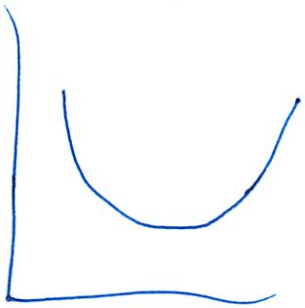
P penalizes through complexity
on training data
that is adjusted R^2

(11)

Correlation coeff measures linear relationship b/w

good to compute for bivariate data

note linear correlation only



← not linear correlation

Outlier screws it up big time

Inside the Secret World of the Data Crunchers Who Helped Obama Win

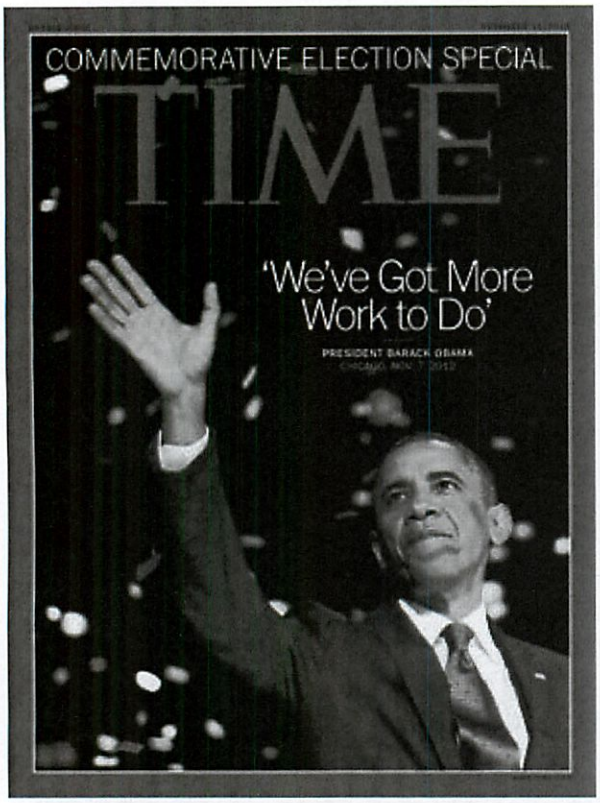
DANIEL SHEA FOR TIME

"The cave" at President Obama's campaign headquarters in Chicago

In late spring, the backroom number crunchers who powered Barack Obama's campaign to victory noticed that George Clooney had an almost gravitational tug on West Coast females ages 40 to 49. The women were far and away the single demographic group most likely to hand over cash, for a chance to dine in Hollywood with Clooney — and Obama.

So as they did with all the other data collected, stored and analyzed in the two-year drive for re-election, Obama's top campaign aides decided to put this insight to use. They sought out an East Coast celebrity who had similar appeal among the same demographic, aiming to replicate the millions of dollars produced by the Clooney contest. "We were blessed with an overflowing menu of options, but we chose Sarah Jessica Parker," explains a senior campaign adviser. And so the next Dinner with Barack contest was born: a chance to eat at Parker's West Village brownstone.

(**MORE:** Four More Years: Obama Wins Re-election)



For the general public, there was no way to know that the idea for the Parker contest had come from a data-mining discovery about some supporters: affection for contests, small dinners and celebrity. But from the beginning, campaign manager Jim Messina had promised a totally different, metric-driven kind of campaign in which politics was the goal but political instincts might not be the means. "We are going to measure every single thing in this campaign," he said after taking the job. He hired an analytics department five times as large as that of the 2008 operation, with an official "chief scientist" for the Chicago headquarters named Rayid Ghani, who in a previous life crunched huge data sets to, among other things, maximize the efficiency of supermarket sales promotions.

Exactly what that team of dozens of data crunchers was doing, however, was a closely held secret. "They are our nuclear codes," campaign spokesman Ben LaBolt would say when asked about the efforts. Around the office, data-mining experiments were given mysterious code names such as Narwhal and Dreamcatcher. The team even worked at a remove from the rest of the campaign staff, setting up shop in a windowless room at the north end of the vast headquarters office. The "scientists" created regular briefings on their work for the President and top aides in the White House's Roosevelt Room, but public details were in short supply as the campaign guarded what it believed to be its biggest institutional advantage over Mitt Romney's campaign: its data.

On Nov. 4, a group of senior campaign advisers agreed to describe their cutting-edge efforts with TIME on the condition that they not be named and that the information not be published until after the winner was declared. What they revealed as they pulled back the curtain was a massive data effort that helped Obama raise $1 billion, remade the process of targeting TV ads and created detailed models of swing-state voters that could be used to increase the effectiveness of everything from phone calls and door knocks to direct mailings and social media.

(**Election 2012**: Photos From the Finish Line)

How to Raise $1 Billion

For all the praise Obama's team won in 2008 for its high-tech wizardry, its success masked a huge weakness: too many databases. Back then, volunteers making phone calls through the Obama website were working off lists that differed from the lists used by callers in the campaign office. Get-out-the-vote lists were never reconciled with fundraising lists. It was like the FBI and the CIA before 9/11: the two camps never shared data. "We analyzed very early that the problem in Democratic politics was you had databases all over the place," said one of the officials. "None of them talked to each other." So over the first 18 months, the campaign started over, creating a single massive system that could merge the information collected from pollsters, fundraisers, field workers and consumer databases as well as social-media and mobile contacts with the main Democratic voter files in the swing states.

The new megafile didn't just tell the campaign how to find voters and get their attention; it also allowed the number crunchers to run tests predicting which types of people would be persuaded by certain kinds of appeals. Call lists in field

offices, for instance, didn't just list names and numbers; they also ranked names in order of their persuadability, with the campaign's most important priorities first. About 75% of the determining factors were basics like age, sex, race, neighborhood and voting record. Consumer data about voters helped round out the picture. "We could [predict] people who were going to give online. We could model people who were going to give through mail. We could model volunteers," said one of the senior advisers about the predictive profiles built by the data. "In the end, modeling became something way bigger for us in '12 than in '08 because it made our time more efficient."

Early on, for example, the campaign discovered that people who had unsubscribed from the 2008 campaign e-mail lists were top targets, among the easiest to pull back into the fold with some personal attention. The strategists fashioned tests for specific demographic groups, trying out message scripts that they could then apply. They tested how much better a call from a local volunteer would do than a call from a volunteer from a non–swing state like California. As Messina had promised, assumptions were rarely left in place without numbers to back them up.

MORE: TIME Staff: Live Twitter Reactions

The new megafile also allowed the campaign to raise more money than it once thought possible. Until August, everyone in the Obama orbit had protested loudly that the campaign would not be able to reach the mythical $1 billion fundraising goal. "We had big fights because we wouldn't even accept a goal in the 900s," said one of the senior officials who was intimately involved in the process. "And then the Internet exploded over the summer," said another.

A large portion of the cash raised online came through an intricate, metric-driven e-mail campaign in which dozens of fundraising appeals went out each day. Here again, data collection and analysis were paramount. Many of the e-mails sent to supporters were just tests, with different subject lines, senders and messages. Inside the campaign, there were office pools on which combination would raise the most money, and often the pools got it wrong. Michelle Obama's e-mails performed best in the spring, and at times, campaign boss Messina performed better than Vice President Joe Biden. In many cases, the top performers raised 10 times as much money for the campaign as the underperformers.

Chicago discovered that people who signed up for the campaign's Quick Donate program, which allowed repeat giving online or via text message without having to re-enter credit-card information, gave about four times as much as other donors. So the program was expanded and incentivized. By the end of October, Quick Donate had become a big part of the campaign's messaging to supporters, and first-time donors were offered a free bumper sticker to sign up.

(**PHOTOS:** Election 2012: Photos from the Finish Line)

Predicting Turnout

The magic tricks that opened wallets were then repurposed to turn out votes. The analytics team used four streams of polling data to build a detailed picture of voters in key states. In the past month, said one official, the analytics team had polling data from about 29,000 people in Ohio alone — a whopping sample that composed nearly half of 1% of all voters there — allowing for deep dives into exactly where each demographic and regional group was trending at any given moment. This was a huge advantage: when polls started to slip after the first debate, they could check to see which voters were changing sides and which were not.

It was this database that helped steady campaign aides in October's choppy waters, assuring them that most of the Ohioans in motion were not Obama backers but likely Romney supporters whom Romney had lost because of his September

blunders. "We were much calmer than others," said one of the officials. The polling and voter-contact data were processed and reprocessed nightly to account for every imaginable scenario. "We ran the election 66,000 times every night," said a senior official, describing the computer simulations the campaign ran to figure out Obama's odds of winning each swing state. "And every morning we got the spit-out — here are your chances of winning these states. And that is how we allocated resources."

Online, the get-out-the-vote effort continued with a first-ever attempt at using Facebook on a mass scale to replicate the door-knocking efforts of field organizers. In the final weeks of the campaign, people who had downloaded an app were sent messages with pictures of their friends in swing states. They were told to click a button to automatically urge those targeted voters to take certain actions, such as registering to vote, voting early or getting to the polls. The campaign found that roughly 1 in 5 people contacted by a Facebook pal acted on the request, in large part because the message came from someone they knew.

(**MORE:** Why the Importance of Early Voting Is Here to Stay)

Data helped drive the campaign's ad buying too. Rather than rely on outside media consultants to decide where ads should run, Messina based his purchases on the massive internal data sets. "We were able to put our target voters through some really complicated modeling, to say, O.K., if Miami-Dade women under 35 are the targets, [here is] how to reach them," said one official. As a result, the campaign bought ads to air during unconventional programming, like *Sons of Anarchy*, *The Walking Dead* and *Don't Trust the B—- in Apt. 23*, skirting the traditional route of buying ads next to local news programming. How much more efficient was the Obama campaign of 2012 than 2008 at ad buying? Chicago has a number for that: "On TV we were able to buy 14% more efficiently ... to make sure we were talking to our persuadable voters," the same official said.

The numbers also led the campaign to escort their man down roads not usually taken in the late stages of a presidential campaign. In August, Obama decided to answer questions on the social news website Reddit, which many of the President's senior aides did not know about. "Why did we put Barack Obama on Reddit?" an official asked rhetorically. "Because a whole bunch of our turnout targets were on Reddit."

That data-driven decisionmaking played a huge role in creating a second term for the 44th President and will be one of the more closely studied elements of the 2012 cycle. It's another sign that the role of the campaign pros in Washington who make decisions on hunches and experience is rapidly dwindling, being replaced by the work of quants and computer coders who can crack massive data sets for insight. As one official put it, the time of "guys sitting in a back room smoking cigars, saying 'We always buy *60 Minutes*'" is over. In politics, the era of big data has arrived.

PHOTOS: Last Days on the Road with Obama

15.062

11/18

(2 min late)

Final project → individual

& Regression + Selection of Variables
& Logistic Response

He has to grade → but possibly extension

Reviewing Regression

(missed)

Correlation

note tricky!

Outlyers + robustness

Make a ~~plot~~ plot so you can see it

Statistical inference See patterns

try to apply patterns to new cases

Most data arises from a sample

Attach some measure of uncertainty

(7)

Too many variables is bad
lots of issues

(study degrees of freedom)

Too ~~many~~ few is bad as well

B_1 should be unbiased $E(B_1) = B_1$

R_{12} is correlation b/w

Variance inflation → ~~one~~ blows up variance
when two variables too similar explanation

So T^2 turns out too small

Forward Selection

---

he likes
explanatory variables > # obs

(3)

Stops before 50

But searching all 200

how much better will it make the model?

SSR = sum of squares of residuals

$S + \{i\}$ = add variable

Repeat for adjusting that have variables in the model

Backward

Start will all explanitory in model

Then take out worst (least impact)

can't start w/ more explanitory than data pts

Stepwise

Forward + Backward together

Will converge if $F_{out} < F_{in}$ (missed)

just works through all possible subsets

(4)

Look at a few models at end & find best

Don't "overfit"

Can bootstrap process

└ Small perturbations in data

~~All Might~~

All models might predict fairly well

Must check w/ validation data

And think about

But what figures of merit to use?

Can prune tree

Remove certain subsets

Branch + bound method

└ Can do all iterations up to 30

But what criteria to use?

(5)

R^2

adjusted R^2 w/ the penalty

C_p = Sum of squares of residuals

Example: Measuring Acidity in Marsh Grass

C_p goes to minimum

till = # of variables

(i always)

Then takes some at

Only ones that it can take out

L7 Logistic Regression

0 or 1

Convert categorical to separate variables

Usually dicotomous Yes or No

(6)

Often a 3rd → no data

but don't consider now

1. Estimate prob of belonging to each group
2. Use cutoff to classify 0 or 1

Just Least sqrs could yet any output

We only want 0 or 1

0 or 1 → Binomial → Bernolli

Let p be prob of belonging to group 1

So use logistic response function

$$p = \frac{e(B_0 + B_1 X_1 + \cdots)}{1 + e(B_0 + B_1 X_1 + \cdots)} = \frac{1}{1 + e^{-(B_0 + \cdots)}}$$

(7)

Curve looks like a CDF

note $\text{odds} = \frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \cdots}$

$\log(\text{odds}) = \text{logit} = \beta_0 + \beta_1 x_1 + \cdots$

(I stated this for Prob
Did I do HW too early?)

What happens if change one of the variables?

$x_1 \uparrow 1$ unit

$= \exp(\beta_1)$

So increases odds by the multiplicative factor $\exp(\beta_1)$

(8)

One parameter model

Find parameters that max the prob of getting the data we got

Iterative computation

Is optimization going on in bg

(mini max)

Deviance - measure of quality of fit

~~Note low D0 ret b~~

note $R^2 = \frac{D_0 - D}{D_0}$

So low deviance is high R^2

test if change in odds

9

Residuals

$$Y_i - \hat{p}_i$$

or normalize

$$\frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

Example: WSJ

people subscribe or not

Base accuracy → all in 1 category

Value is change in odds

get estimated prob for each
then divide 0 or 1

(10)

Fills 74.1 % of diff b/w 67.9% and 100%

15.0762 Recitation 11/20

Need to redo HW2
last recitation I attended
~15/60 students attend

Today: Classification + Regression Trees

Example

Classification tree

Input	Output
$X_1 \dots X_7$	Y
	Guess = 1

Normalize = to put on same scale

(2)

Max # of records in terminal node

Max # of levels to display

Classify as 1 or 0

Can see training score

lift chart = did gain predictive power

Went to score new data in workstep

Can see tree rules

(forgot this method's specifics — should have brought my book

(3)

Regression

Similar

but values 0 → 1 inclusive

Same procedure

but on predict tab, not classify

lift chart - hard to describe here

where it is continuous response

predicted value

WP: these types are pretty similar

(9)

lift chart Gini sorts probabilities for 1, 0

had to explain for continous

(kinda glad I haven't gone to recitation)

WP: CART algorithm

decide when to split

bagging / bootstrap aggregation

machine learning meta algorithm

generates new training sets

randomly (w/ replacement)

build a tree for each sample

get result for that

then run for 100 trees

(9)

And pick the most common response

Not automatic in XL Miner

If does not lead to 1 final tree answer!

(over 35 min early)

11/21

15.062

7 Logistic Regression
8 Discriminant Analysis

Missed Class due to Travel

15.062

11/26

L8 Discriminant Analysis
Neural Nets

(Oh HW 2 due today! :)

- I thought Wed
  - Hmm Opps
  - I think I am going to ~~Drop~~ Listen this class
  - I've just been too busy for this class

German Credit Case Contrast

(Why was I not more accurate at looking at this class
Starting #12 is hard!)

Still want to go to class

Should try hw too...

Statistics → Data Science

(missed last class - have not reviewed)

Part 8 cont today

Discriminate analysis

(2)

Linear has more assumptions
multivariate
Logistic have less

Could check marginal distribution of variables

Support Vector Machines have a margin
What matters are the info on the banding

Logistic regression weights things by binomial
variance

So some robustness in logistic you don't get
in linear

Use Ensemble methods
└ boosting + bagging

makes things better
Combine methods
Sometime better result

(3)

But how to make those methods interable
- Regression tree

But UOx Regression Tree is no longer interable

Can add computing time

Combine several

Interpability can be important

(4 Neural Nets

black box for many

but can be highly effective

good in complex situations

inc Machine Learing

Want to prevent overfitting

Base in code

Not in prob

9

Modeled on the brain

Neurons w/ connections

No specific relationship b/w response + prediction

Had to visualize inputs

Called multilayer feed-forward networks

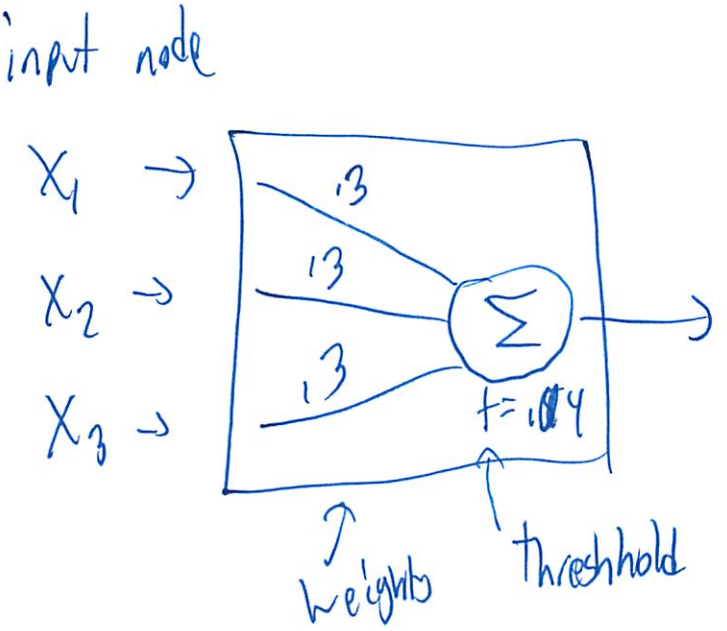
Categorical inputs

in this case

0 or 1 output in categorization

Other times some value

Black box



if sums > .4 then 1
otherwise 0) output

Perceptron Model

$$Y = I\left(\sum_i w_i X_i - t\right)$$

or

$$Y = \text{sign}\left(\sum_i w_i X_i - t\right)$$

(6)

People don't generally react linearly

So more complex

- multiple layers
- fancy activation function

Connections die out if are not used
esp at age 0-3
and again at puberty
neurons grows rapidly 9-14 age

Tough to get people interested in Sci then

Does not need to be to every node
So immediately you get a variable selection process
Have weights → inc small weights
So may use L-citeron
abs value

(7)

more nodes = tighter fit (but might not generalize)

SW usually auto defaults to 3 → 10

So try + check w/ how many nodes?

└which is similar to pruning w/ CART

---

intercept term is called the bias

Weights are set randomly up front

Then we tune them

Prof: in humans I think some are preset

transfer function g is monotone

logistic → linearly in middle range

$$g(v) = \frac{1}{1+e^{-sv}}$$

can be linear or exponential

Small squashing effect

Can modify to make sharper or more tapered

Q: Do we try to update the model?

As we get new data

(Real time

Or at night)

Q Do we time data

Weight new data more

Dives out old data

Some function to have those decline

$$\text{Output}_j = g\left(\theta_j + \sum_{i=1}^{p} w_{ij} x_i\right) = \frac{1}{1 + e^{-\left(\theta_j + \sum_{i=1}^{p} w_{ij} x_i\right)}}$$

(9)

Output layer often only has 1 node
but may have several

Like a hidden layer, but can see output layer

Ex: classify new product
fat + salt on user acceptance
weights + threshhold (biases) are estimated
using updating fn

Can get logistic regression from this!
from a simple neural net
but learns via max likelihood

Often scale data so $0 \to 1$
4 level interval for 4 level variable

(10)

Nominal are converted to dummies

~~careful~~ ← no numeric (like Gender)

Useful to take log of highly skewed data

SW can alert you about this

↳ good sw should

but most sw today doesn't

How does it learn?

Supervised so it does know the truth

learns through example

knows if it makes a mistake

at each level it adjusts weights to see if it can do better → back propagation

Adjusts weights at end

then moves ← one layer, trying to ↓ error rate

(11)

Then decides at some point to stop

Greedy alg

Since gradient method

Weighting the error

$$err_j = \hat{y}_j(1-\hat{y}_j)(y_j - \hat{y}_j)$$

bernoli variance

Error adjusted by squashing fn

Optimization:

take old value and add a adjustment factor by error

Fancier: Newtons' method

L w/ 2nd deriv

(12)

Sw doesn't let ya fiddle w/ ℓ

ℓ is weight decay parameter

Does involve the explanitory variable

For Regression → we minimized sum of square of errors

there → gradient descent

easy to use parallel computing

How do we improve?

take deivitives

$$W_{ij}^{new} = W_{ij}^{old} + \ell \delta_j x_i$$

$$\delta_j = -\frac{\partial R}{\partial v_j} (y_i - g_i(v_i)) g_i(v_i)$$

Called the generalized delta rule

(can read about this on WP - need to study)

(13) Case one as each case goes through
more accurate
batch faster
not as good

lots of iterations through

So output

Similar to how we saw before

get actual final model for network..

---

if know something about problem

ie writing #s on check

(didn't get this)

When do ya stop?

When rates don't change much

Most SW has some stopping pt

(14)

(summary)

Do some sensitivity analysis to see how results change

Pros

Good predictive petomance

Models very complex

Noise tolerant

Cons

Black box - doesn't answer why?

Bt local optimum

and easy to over fit

Poor Man's version of seeing if a variable matters: take it out!

(missed example)

- Passing through an obs is fast

(15)

Ex Classify Handwritten #s

16 x 16 greyscale images

first ~~training~~ normalize

- remove slant
- size scale

then training data

Tried 5 diff networks

c) patch connected – since handwriting has certain connections b/w

d) Could say slope same, but y-intercept changes

fixed term

bias changes

(16)

If think about problem -> you think about how network explots the structure of how people write

The Boston Globe

Lifestyle

A new love affair, by the numbers ; Thanks to Nate Silver, statisticians get a second look

Beth Teitell

By Beth Teitell Globe Staff
550 words
22 November 2012
The Boston Globe
BSTNGB
G.14
English

Throughout history, "statistician" has not typically been one of the sexier job titles. But now -- thanks in part to Nate Silver, who correctly predicted the presidential election -- that may be changing.

Although he's yet to hear anyone use "Wanna go home and crunch some numbers?" as a pickup line, MIT professor Erik Brynjolfsson says the field's allure is growing.

"Statisticians have become sexy just the way geeky Internet nerds became sexy in the 1990s, and I suppose investment bankers were in the 1980s," said Brynjolfsson, director of the MIT Center for Digital Business. "Things that drive the economy give people power, and I guess that's sexy."

"There are over 100 billion Internet searches every month," he said. "That's a staggering amount of data. People like Nate Silver are now very much in demand because they have the tools for looking at all this data."

Indeed, a recent study by the McKinsey Global Institute predicts that the US will need between 140,000 and 190,000 more professionals with expertise in statistical methods by 2018.

The new heartthrob stature of statisticians was captured in the Nov. 19 issue of The New Yorker, in an imagined love letter to Silver, whose FiveThirtyEight political calculus blog runs in The New York Times.

"I can't stop thinking about how you study polls and create probability models and predict elections and how you're always right, which I think is so unbelievably cute," a fictional smitten 11-year-old wrote, "and I keep imagining you saying to me, 'Emma, I think that there's a 93.7% chance of me falling in love with you.' " (Paul Rudnick penned the piece.)

So pronounced was the post-election statistician bump, that the American Statistical Association put out a press release that both reveled in the field's high profile during the 2012 election -- and pointed out that statisticians are enabling advances in other fields, too. Among them: medicine, economics, public health, agriculture, business analytics, law enforcement, and weather forecasting.

No mention was made of Boston's most famous statistician, Bill James, who coined the term "sabermetrics" to describe the specialized analysis of baseball through objective evidence. (The term is derived from an acronym for Society for American Baseball Research.)

As baseball -- and Brad Pitt -- fans no doubt recall, statistics got a pre-Nate Silver glamour boost when Pitt played the stats- using Oakland Athletics' general manager, Billy Beane, in the 2011 movie adapted from the book "Moneyball."

Meanwhile, with the International Year of Statistics just over a month away, in 2013, the statistical association's incoming president reflected on the field's change.

"People used to think of us as actuaries," said Marie Davidian, a statistics professor at North Carolina State University in Raleigh.

You've come a long way . . . mathematicians?

Beth Teitell can be reached at bteitell@globe.com. Follow her on Twitter @bethteitell.

Caption: Clockwise from above: Nate Silver, author of the political blog FiveThirtyEight; sabermetrics guru Bill James; Brad Pitt (with Jonah Hill) as Billy Beane in "Moneyball." Robert Gauldin/ Associated Press Melinda Sue Gordon/Columbia Pictures Wendy Maeda/ Globe Staff/file 2011

Globe Newspaper Company, Inc.

Document BSTNGB0020121122e8bm0003g

15.062

L10 Cluster Analysis

1/28

(6 min late) 24 Neural Nets

Reviewing Neural Net (L9)

Don't overfit

Check reading

Want less # links + # weights

Using special structure

but not bring up degrees of freedom

---

are some better black boxes

but can we communicate bette

(2) 40

Cluster Analysis

Forget training data

Just get some data

do observations look similar to others?

Common qu to ask of data

---

want close clusters (intra)

but far away from others (inter)

---

need some distance measure

How many clusters?

(3)

hirarchical vs partitional

Can nest piles

Will talk abat both

produces nested clusters

dendrogram

goes botton up

puts pairs together

including single objects and previous pairs

Can identify atlyers somewhat eaisly

(4)

Don't need to tell in advance how many clusters we want

Better if have domain knowledge

Agglomerative techniques

1. Min distance
2. Max distance
3. Grp avg
   - Center of mass

Each gives ya diff results
- Not about a single alg or look at data
  - does looking at something tell ya about it

(5)

Cutting → limited # of clusters

but normally don't say how many clusters

example Public Utility Data

Cluster objects
Do deep analysis on 1
Then investigate rest of set

need distance matrix

Lhe is being vague how this is found

Can be complex to keep in memory
for large data sets

multiple options
don't pick the one ya like - unethical
try to find one that explains data best

prev assumption

(6)

* watch the scale *

This is why we explore data / do data mining

Slice + dice → looking for patterns

When it doesn't look right, investigate

DNA Microarrays

Sample 1 Sample 2 S3 ...

gene 1

gene 2

⋮

gene n

cluster genes (rows) → diff tiss pts

cluster samples (cols) → expression level of thousands of genes

(better explained than 7.012)

(7)

Why are these results interesting?

Validating Clusters

Easy to do w/ supervised learning

harder here

- interpretability
- make sense to subject matter experts
- other features
- stability w/ more data

  or break in half randomly

  are clusters still similar?

Limitations

Computational cost of $O(n^2)$ distance matrix

1 pass through data

Sensitivity = stability

(8)

k-means

the hard way

non hierarchical

try all possible clusters #s

└ like k-NN

Combinatorial: all possible subsets of all things

approx method

(leave 25 min early)

15.062

LII Affinity Analysis

12/3

Missed class

15.062 12/5

12 Ensemble

~~13 Time Series Forecasting~~

(6 min late)

Ensemble Methods

Majority rules

Try it through a bunch

Difference is through the bootstrapped samples

Not changing the data

Does building 100 trees help?

Assuming each is independent

Binomial 35% each

Guess is making a mistake

But get majority of rules to pick wrong thing

13 of 25 (6% now)

(2)

Bagging = Bootstrap Aggregation

Sample w/ replacement

Same # of observations

average the results of 100 predictions

each tree might have diff features + nodes

* Sample mean is more precise than single sample

But its much harder to show ya the tree!

Active Research Area

** For scatter plot → if we have 100

~~least → most extreme~~

show 0, 25, 50, 75, 100 — or extreme on both ends & — mid →

(3)

Simulated example

Only looks at 1st feature

Small data perturbation makes huge impact on tree

11 trees

Large drop in error

Consensus bounces around

Or look at probability

Bagging a good classifier makes it better but for a worse data makes it worse

interpretability can ya explain it to some one

Boosting

Really got people excited

Esp trying to understand why it works

Out of lets try something + see why it works

Doesn't matter is asy correct

Initially all N records assigned = weights

Then update weights at end

What does it mean to put weights on observations?

Some are more important than others

Each model is based on previous weights

Now modify training data set

to reflect which ones we think should show up.

(5)

How do ya up + down weight the observations?

To more weight where was doing less well
Less weight where doing better

Can only use actual weights in logistic regression
So use weights w/ input data

Committee more powerful classifier

Code as $\ominus 1$, $\oplus 1$

Average error - when make mistake
On training data

Something else on real data
(where don't know answer)

(6)

either up weight or down weight

just switch signs of output

Teasy fix

(missed & this should or shouldn't happen)

Adjust for how good classifier is given error rate

So add up all or classifiers

bad = 75% error

then an switches sign

This is called Ada-boost

Pretty basic concepts

No objective fn like an optimization alg

But Underlying this there is an objective fn

(1)

Is an ensemble method

---

ex: Can we turn lead into gold?

χ^2 distribution = Sum of ind Gaussian variables

More items we have in there
the more variability we'll see in practice

Just a little bit better than coin flipping

So boosting gets you to sweetspot w/ 100 trees

Since is an optimization problem

Note this assumes a lot of ind sample
if measure wrong on all then screwed

Trees are pretty stable too.

~~Also~~ robust against explanatory variable,
intrinsicly robust
not like regression

(8)

Generally speaking these methods help

But could hurt

↳ slide ~~###~~ 43

Random Forests

~~When~~ do we have big

# of rows ?

columns ?

is it sparse ?

how do we deal w/ those diff cases ?

~~###~~ (missed)

Popular tool

Very hard to explain

Way of putting tree together

(9)

Use validation data to see how many boosted steps is good

Double Randomization

Will eventually get all explanatory variables in

but eventually

Always have validation data

Is the data left outside of the bag on boosting

Check your alg on 50 data sets

No universal truth

Fall 2012 Data Mining: Finding the Data and Models that Create Value 15.062 (ESD.754J) (Welsch)

Homework #3

Due: Friday, December 7, 2012

Reading:

DMBI Chapters 11, 12, 13, and 14 for homework. In class we will discuss material from Chapters 4, 15, 16, and 17.

Problems (individual work unless otherwise noted)**:**

1. **Case** (up to two may work on together and submit one write-up):

German Credit case at the end of the book. Use the following methods on these data: discriminant analysis and neural nets. I would like for you to compare and contrast the results you obtained using the different methods including the four you used for this case on the last problem set.. To do this please modify part 2 of this case and divide the data into training, validation, and test data sets as follows: Train with 600, validate with 200, and test with 200. Please also let us know what you think your best model is. We will pick a random test set to compare the final models suggested by each of you.

2. 14.2

3. 13.3

11/17/2012

15.062 12/10

L 13 Time Series

(25 min late)
L slide 17

Remove seasonal

Could fit a line to data

Or seasonalize the trend

Classical time series forecasting

Could convert to freq domain w/ FFT
Could forecast w/ that

Linear trend model

How do people code it?
L Must think about!

Can't compare R^2 - diff units
w/ log

(2)

Must convert back before looking at $\hat{y}$

So why not go to regression model right away?

Use dummy variables for seasonal

remember drop last one
since it would be all 1s

How integrate trend models w/ seasonal
(not really following)

People boring $ now to pay dividends before taxes ?

Ex: Coca Cola

Dummy variables

regression w/ coefficients

(3)

Forecast w/ future

Tune w/ training data

Exponential growth is very difficult

↳ multiplicative growth

Not dynamic

can't adapt

but was good to see the classical method

Simple Exponential Smoothing

Common type of modeling

Optimization problem

How do ya pick α?

(9)

Winters' method
(missed)

Diff divisions use diff methods
Its esp useful to know how your forecast petered in the past

ARIMA

Box - Jenkins
Collection of Linear Statistical Models
if it works, it works
kinda like a black box
must choose model
must estimate parameters

Autoregressive process (AR)

Straightforward
Could fit w/ regression

(5)

Moving Average Process (MA)

same term, diff meaning as before ...

auto regression on error

Why was it done this way?

Can model any time series w/ enough terms and enough data

But may have lots of coefficients

Running up degrees of freedom

Introduce moving avg part

ARMA

AR and MA

how many lags to use?

(6)

Could measure unemployment over time

Can get upper + lower 95% limits

Must go back and build prob model for the error

Rt forecast limits never shared

Possible to do it w/ this model

Sad part of this profession

Pure Integrated (I) Process

Random Walk

Y is not expected to stay close to any long term mean value

Random Walk down Wall St

Stocks behave this way?

How do ya make $ in Wall St?

aka pure integrated process

are modeling the difference

(7)

ex Caterpillar Stock

difference is random

does have some diff

but no clear diff

auto correlation

runs test

correlation coefficients are quite small

Std error

1st diff series seems like nice series

Chart of 1st differences

except for spike

must adjust for dividend payment date

so must be cleaned up

Can we model the difference?

Zig zags don't get much bigger

Next time: Forecasting

15.062

12/12

L14 SVM

(skipped class)