Sta 414/2104 S

http://www.utstat.utoronto.ca/reid/414S10.html

STA414S/2104S: Statistical Methods for Data Mining and Machine Learning

January - April, 2010 , Tuesday 12-2, Thursday 12-1, SS 2105

Course Information

This course will consider topics in statistics that have played a role in the development of techniques for data mining and machine learning. We will cover linear methods for regression and classification, nonparametric regression and classification methods, generalized additive models, aspects of model inference and model selection, model averaging and tree based methods.

Prerequisite : Either STA 302H (regression) or CSC 411H (machine learning). CSC108H was recently added: this is not urgent but you must be willing to use a statistical computing environment such as R or Matlab.

Office Hours : Tuesdays, 3-4; Thursdays, 2-3; or by appointment.

Textbook : Hastie, Tibshirani and Friedman. *The Elements of Statistical Learning*. Springer-Verlag. http://www-stat.stanford.edu/~tibs/ElemStatLearn/index.html

Course evaluation :

- Homework 1 due February 11: 20%,
- Homework 2 due March 4: 20%,
- Midterm exam, March 16: 20%,
- Final project due April 16: 40%.

Tentative Syllabus

- Regression: linear, ridge, lasso, logistic, polynomial splines, smoothing splines, kernel methods, additive models, regression trees, projection pursuit, neural networks: Chapters 3, 5, 9, 11
- Classification: logistic regression, linear discriminant analysis, generalized additive models, kernel methods, naive Bayes, classification trees, support vector machines, neural networks, K-means, k-nearest neighbours, random forests: Chapters 4, 6, 9, 11, 12, 15
- Model Selection and Averaging: AIC, cross-validation, test error, training error, bootstrap aggregation: Chapter 7, 8.7
- Unsupervised learning: Kmeans clustering, k-nearest neighbours, hierarchical clustering: Chapter 14

Some references

- Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S (4th Ed.)*. Springer-Verlag. Detailed reference for computing with R.
- Maindonald, J. and Braun, J. (). *Data Analysis and Graphics using R*. Cambridge University Press. Gentler source for R information.
- Hand, D., Mannila, H. and Smyth, P. (2001). *Principals of Data Mining*. MIT Press. Nice blend of computer science and statistical methods.
- Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press. Excellent, but concise, discussion of many machine learning methods.

Some Courses

- `http://www.stat.columbia.edu/~madigan/DM08`
  Columbia (with links to other courses)
- `http://www-stat.stanford.edu/~tibs/stat315a.html`
  Stanford
- `http://www.stat.cmu.edu/~larry/=sml2008/`
  Carnegie-Mellon

Data mining/machine learning

- large data sets
- high dimensional spaces
- potentially little information on structure
- computationally intensive
- plots are essential, but require considerable pre-processing
- emphasis on means and variances
- emphasis on prediction
- prediction rules are often automated: e.g. spam filters

- Hand, Mannila, Smyth “Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”

Applied Statistics (Sta 302/437/442)

- small data sets
- low dimension
- lots of information on the structure
- plots are useful, and easy to use
- emphasis on likelihood using probability distributions
- emphasis on inference: often maximum likelihood estimators (or least squares estimators)
- inference results highly ‘user-intensive’: focus on scientific understanding

Some common methods

- regression (Sta 302)
- classification (Sta 437 – applied multivariate)
- kernel smoothing (Sta 414, App Stat 2)

- neural networks
- Support Vector Machines
- kernel methods
- nearest neighbours
- classification and regression trees
- random forests
- ensemble learning

Some quotes

- HMS[1]: “Data mining is often set in the broader context of knowledge discovery in databases, or KDD” `http://www.kdnuggets.com/`
- Hand et al.: Data mining tasks are: exploratory data analysis; descriptive modelling (cluster analysis, segmentation), predictive modelling (classification, regression), pattern detection, retrieval by content, recommender systems
- Data mining algorithms include: model or pattern structure, score function for model quality, optimization and search, data management
- Ripley (pattern recognition):“Given some examples of complex signals and the correct decision for them, make decision automatically for a stream of future examples”

[1] Haughton, D. and 5 others (2003). A review of software packages for data mining. *American Statistician*, **57**, 290–309.

Software

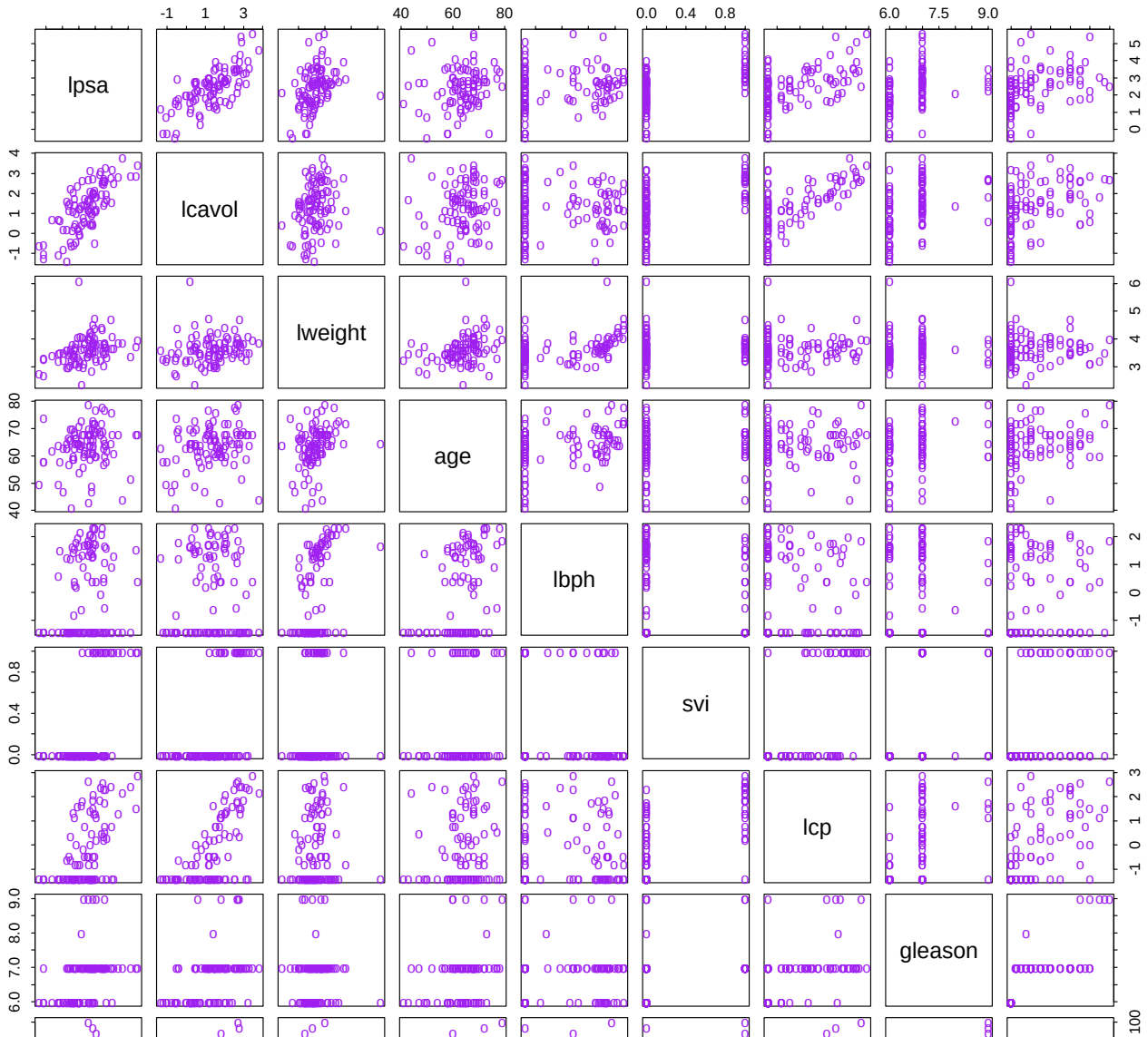
- Haughton et al., software for data mining: Enterprise Miner, XLMiner, Quadstone, GhostMiner, Clementine
Enterprise Miner	\$40,000-\$100,000
XLMiner	\$1,499 (educ)
Quadstone	\$200,000 - \$1,000,000
GhostMiner	\$2,500 – \$30,000 + annual fees
- `http://probability.ca/cran/` – Free!!
- You will want to install the package `MASS`, other packages will be mentioned as needed

Examples from text

- email spam 4601 messages, frequencies of 57 commonly occurring words
- prostrate cancer 97 patients, 9 covariates
- handwriting recognition data
- DNA microarray data 64 samples, 6830 genes

- "Learning" = building a prediction model to predict an outcome for new observations
- "Supervised learning" = regression or classification: have available sample of outcome (y)
- "Unsupervised learning" = clustering: searching for structure in the data

`Book.Figures`


lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45

Some of the handwriting data

```
9 -1 -1 -1 -1 -1 -0.948 -0.561 0.148 0.384 0.904 0.29 -0.782 -1 -1 -1 -1 -1 -1 -1 -1
-0.748 0.588 1 1 0.991 0.915 1 0.931 -0.476 -1 -1 -1 -1 -1 -1 -0.787 0.794 1 0.727
-0.178 -0.693 -0.786 -0.624 0.834 0.756 -0.822 -1 -1 -1 -1 -0.922 0.81 1 0.01 -0.928 -1
-1 -1 -1 -0.39 1 0.271 -1 -1 -1 -1 0.012 1 0.248 -1 -1 -1 -1 -1 -0.402 0.326 1 0.801
-0.998 -1 -1 -0.981 0.645 1 -0.687 -1 -1 -1 -1 -0.792 0.976 1 1 0.413 -0.976 -1 -1
-0.993 0.834 0.897 -0.951 -1 -1 -1 -0.831 0.14 1 1 0.302 -0.889 -1 -1 -1 -1 0.356 0.794
-0.836 -1 -0.445 0.074 0.833 1 1 0.696 -0.881 -1 -1 -1 -1 -1 -0.368 0.955 1 1 1 1 0.905
1 1 -0.262 -1 -1 -1 -1 -1 -1 -1 -0.507 0.451 0.692 0.692 -0.007 -0.237 1 0.882 -0.795
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 0.155 1 0.436 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-0.991 0.703 1 -0.025 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.833 0.959 1 -0.629 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.6 0.998 0.841 -0.932 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -0.424 1 0.732 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.908 0.43 0.622 -0.973 -1 -1
-1 -1 -1
6 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.783 -0.973 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.364 0.789 -0.371 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-0.467 0.963 0.609 -0.986 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -0.875 0.605 0.96 -0.351 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 0.05 1 0.096 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-0.582 0.97 0.532 -0.922 -1 -1 -1 -1 -1 -1 -0.602 0.307 0.718 0.718 -0.373 -0.998 0.723
1 -0.431 -1 -1 -1 -1 -1 -0.817 -0.136 0.808 1 1 1 0.697 -0.67 0.965 0.659 -1 -1 -1 -1
-1 -0.512 0.738 1 0.839 -0.336 -0.977 0.433 0.878 0.161 1 -0.102 -1 -1 -1 -1 -0.643
0.87 0.97 0.264 -0.971 -1 -0.399 1 0.117 0.835 0.968 -0.701 -1 -1 -1 -1 0.198 1 0.052
-1 -1 -0.291 0.876 0.79 -0.819 0.392 0.962 -0.461 -1 -1 -1 -0.948 0.82 1 -0.168 -0.475
0.28 0.968 0.88 -0.613 -1 -0.551 0.854 0.98 0.498 0.324 0.324 0.328 0.998 1 0.97 0.995
0.976 0.25 -0.642 -1 -1 -1 -0.64 0.661 0.971 1 1 1 0.95 0.774 0.774 0.302 -0.522 -1 -1
-1 -1 -1 -1 -1 -0.663 -0.606 -0.606 -0.606 -0.688 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

Some basic definitions

- input X (features), output Y (response)
- data $(x_i, y_i), i = 1, \ldots N$
- use data to assign a rule (function) taking $X \rightarrow Y$
- goal is to predict a new value of Y, given X: $\widehat{Y}(X)$

- regression if Y is continuous
  classification if Y is discrete
- Examples...

How to know if the rule works?

- compare $\hat{y}_i$ to y_i on data **training error**
- compare $\widehat{Y}$ to Y on *new* data **test error**
- In ordinary linear regression, the least squares estimates minimize

$$\Sigma(y_i - x_i^T \beta)^2$$

- and the minimized value is the sum of the squared residuals

$$\Sigma_i(y_i - \hat{y}_i)^2 = \Sigma_i(y_i - x_i^T \hat{\beta})^2$$

- test error is

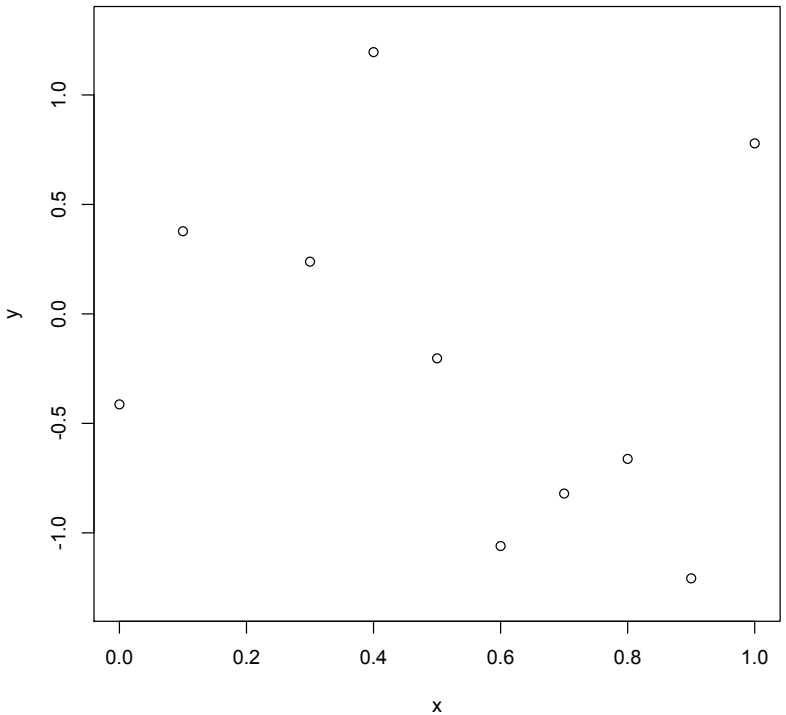
$$\Sigma_j(Y_j - x_j^T \hat{\beta})^2$$

... training and test error

- In classification: use training data (x_i, y_i) to estimate a classification rule $\widehat{G}(\cdot)$
- $\widehat{G}(x)$ takes a new observation x and assigns a class/category/target $g \in \mathcal{G}$ for the corresponding new y
- for example, the simplest version of linear discriminant analysis says
  $y_j = 1$ `if` $x_j^T \hat{\beta} > c$ `else` $y = 0$
- what criterion function does this minimize?
- test error is proportion of mis-classifications in the test data set
- in regression or in linear discriminant analysis, training error is always smaller than test error: why?

Linear regression with polynomials

- model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \epsilon_i$
- assume $\epsilon_i \sim N(0, \sigma^2)$ (independent)
- how to choose degree of polynomial?
- for fixed degree, find $\hat{\beta}$ by least squares



go to R

The Netflix Prize

http://www.netflixprize.com//index

Digital Doctoring

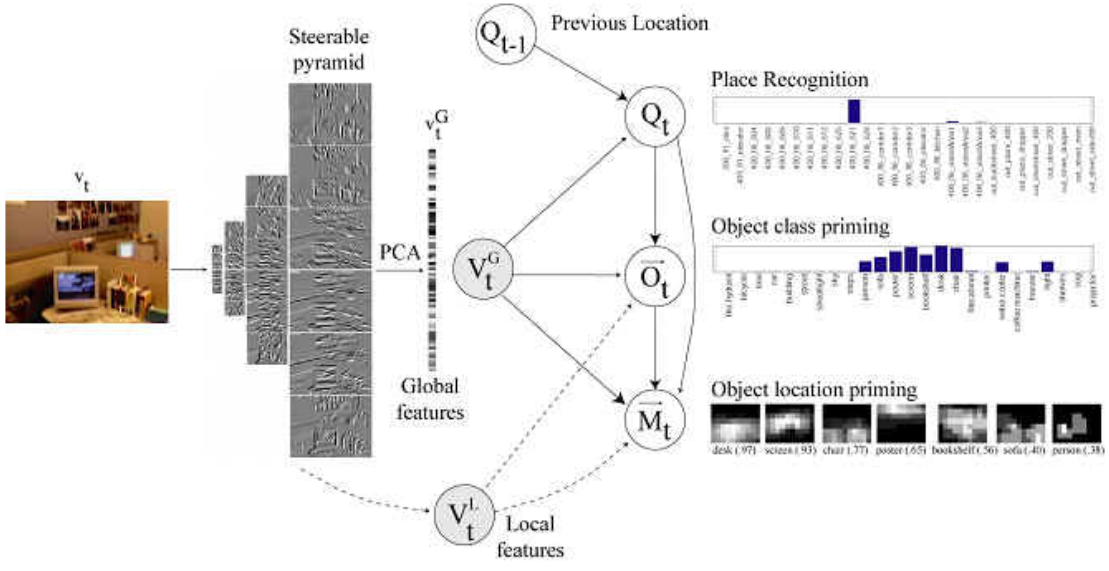
- "Digital doctoring: how to tell the real from the fake", H. Farid, *Significance* 2006.
- "Exposing Digital Forgeries by Detecting Inconsistencies in Lighting", M.K. Johnson, H. Farid, *ACM 2005*
- `http://www.cs.dartmouth.edu/farid/`

Figure 4: The lighting of Pitt and Jolie is inconsistent in this composite.

H. Farid, *Significance* 2006.

Automated Vision Systems

- `http://www.cs.ubc.ca/~murphyk/Vision/placeRecognition.html`
- "Context-based vision system for place and object recognition" Antonio Torralba, Kevin P. Murphy, William T. Freeman and Mark Rubin, ICCV 2003.




Kevin Murphy

For next class/week

- Thursday: Make sure you can start `R`
- Thursday: Overview of linear regression in `R`
- Tuesday: Read Chapter 1, 2.1–2.3 of HTF
- Tuesday: We will cover Ch. 3.1-3.3/4 of HTF