Business Statistics

Tommaso Proietti

DEF - Università di Roma 'Tor Vergata'

An introduction to Statistical Learning and Data Mining

Why Business Statistics? I

This course is about data and mining big data.

- Statistics has been defined as the art (or science) of learning from data. "Statistics concerns what can be learned from data" (A.C. Davison, Statistical Models, CUP, 2003).
- Data Mining as the "process of seeking interesting or valuable information within large data sets" (D. Hand et al., Statistical Science, 2000).
- "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". (Hand, Mannila & Smyth, Principles of Data Mining, MIT Press, 2001).

Why Business Statistics? II

- The advances in information technology have made available very rich information data sets, often generated automatically as a by-product of the main institutional activity of a firm or business unit. Most organizations today produce an electronic record of essentially every transaction in which they are involved. Firms collect terabytes data over operating periods (transactions data, e.g. credit cards). Most often these data are collected as secondary data, with no specific sampling design or research question on top.
- Data Mining deals with inferring and validating patterns, structures and relationships in data, as a tool to support decisions in the business environment.

Why Business Statistics? III

- The course offers an insight into the main statistical methodologies for the visualisation and the analysis of business and market data, providing the information requirements for specific tasks such as credit scoring, prediction and classification, market segmentation and product positioning. Emphasis will be given to empirical applications using modern software tools (Rstudio, Matlab, SAS).

Why Business Statistics? IV

- The aim of the course is to provide students with the most relevant analytical tools that are useful in the business world. This involves being able to extract information from the large volume of data readily available from the business environment, as well as being able to analyse this data in a way that leads to useful models and methods for the prediction of future outcomes.

- Lecturer:
  Tommaso Proietti
  Office: Room 47 Dipartimento di Economia e Finanza (B Building, 2nd Floor)
  Office hours: typically Monday 2-4 p.m.
  (tommaso.proietti@uniroma2.it)

Syllabus I

1. Introduction to data mining. Tools for data analysis, visualisation and description.
2. The linear regression model.
3. Model selection and evaluation: bias-variance trade-off, model complexity and goodness of fit. Cross-validation. Selection using information criteria.
4. Regularization and shrinkage methods: rigde regression, lasso, forward stagewise regression.
5. Linear methods for classification: Bayes Classification Rule. Discriminant analysis. Canonical variates.

Syllabus II

6. Linear methods for classification: logistic regression.
7. Semiparametric regression: Regression splines and smoothing splines.
8. Kernel smoothing methods: Local polynomial regression. Density estimation. Nearest neighbour classification.
9. Additive Models, tree-based methods. GAM, Regression and classification trees. Boosting.

Assessment

- 30% Weekly Assignments
- 70% Final Exam (Date: to be appointed)

Where to study

The main references for the course are

- G James, D Witten, T Hastie, and R Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, Springer Series in Statistics, 2013. Dowloadable at http://www-bcf.usc.edu/ gareth/ISL/
- T Hastie, R Tibshirani and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, Springer Series in Statistics, 2009. Website: http://www-stat.stanford.edu/ElemStatLearn/.

Slides, readings, datasets and supplemental material will be made available in the course website.

Statistical learning and data mining

We distinguish two main learning problems.

- Supervised learning. Data are grouped or ordered by some response. We wish to predict an outcome variable (output, dependent variable) from a set of features or characteristics (inputs).
  - Classification (prediction of a binary or multinomial outcome)
  - Regression (prediction of a quantitative outcome)
- Unsupervised learning: no outcome is available (cluster analysis, multidimensional scaling, principal components)

A typical dataset has at least two dimensions: individuals and variables (measurements). They lend themselves to the representation as **data matrices**. Three dimensional arrays (where the third dimension is time or space) are also common. It is important to distinguish variables types and their scale of measurement.

Measurement scales

1. Qualitative (Categorical)
   - Nominal (binary, multinomial). The measurement deals with the allocation of a case to a response category. Examples: sex, marital status, solvency. We can compare measurements according to the identity principle (equal or different).
   - Ordinal. Response categories are ordered. We can state which category is higher or lower.
2. Quantitative
   - Interval scale. The origin is arbitrary. ($\{$Temperature in degrees Celsius$\}$). We can compare two measurements using the algebraic sum (e.g. y_1 and y_2 differ by $y_1 - y_2$)
   - Ratio scale. The measurements have a natural zero (e.g. sales, n. of customers). We can compare measurements using their ratio.

Supervised learning: the prediction problem

Glossary and notation

- Target variable: Y (categorical or quantitative)
- Inputs: X
- Prediction model or method: $f(X)$
- Prediction error: $Y - f(X)$
- Loss function: $L(Y, f(X)) = L(Y - f(X))$. A function of the prediction error. Ex. quadratic loss $L(Y, f(X)) = (Y - f(X))^2$.
- The expected loss is the value of L averaged over the distribution of Y, X.
- The expected conditional loss is the value of L averaged over the distribution of $Y|X$.

- Given a loss function, we select the prediction method that yields the minimum expected loss. For instance, under squared loss, $f(X) = \mathrm{E}(Y|X)$. In fact, $\mathrm{E}(Y|X)$ minimises $\mathrm{E}[(Y - f(X))^2|X]$.
- The prediction method depends on a set of unknown parameters. These can be estimated using data (y, x) to give $\hat{f}(x)$. This generalizes to $\hat{f}(X)$.
- In a data-rich environment we (randomly) divide the dataset in 3 nonoverlapping parts:
  1. training set: used to fit the models and to construct prediction rules.
  2. validation set: used for model selection, to estimate the prediction error for model selection
  3. test set: assessment of the final chosen model and assessment of generalization error

- We define the *training error*: the average loss over the training sample

$$\overline{err} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- We define the *test error* or *generalization error*

$$Err_{\mathcal{T}} = \mathrm{E}[L(Y, \hat{f}(X))]$$

  For a fixed training set the model is estimated and the prediction rule established. The test error is the expected loss over an independent test sample taken from the population. The expected test error is $\mathrm{E}[Err_{\mathcal{T}}]$ where the expectation is taken wrt all the possible training sets.

- We wish to decide on the goodness of a prediction method on the basis of the test error.
- It turns out that the training error is a poor estimate of the test error.