Data Mining: Supervised Methods

Fabrizio Cipollini

Version: June 13, 2016

1 Introduction

We present some selected topics concerning Statistical Supervised Methods, focusing on regression and classification problems. We apply them to real data using scripts written in `R`.

2 Background

- `R` Intro: Zhao (2012, ch. 1-3); lessons by prof. Stefanini
- Linear and Logistic Regression models: James *et al.* (2014, p. 59-102, 127-138); lessons by prof. Rampichini

3 Topics

- Continuous vs categorical dependent variables: regressions vs classification. James *et al.* (2014, Ch. 2)
- Focus on parameters and their estimators (Statistics) vs focus on Prediction and corresponding error measures (Statistical Learning). James *et al.* (2014, Ch. 2)
- Mean Squared Error (MSE), MSE decomposition and Bias/Variance trade-off, Training vs Test MSE, the idea of penalization. James *et al.* (2014, Ch. 2)
- Too many predictors? Automatic variable selection:
  - Subset selection: best subset; stepwise (forward, backward, hybrid)
  - Shrinkage (regularization) methods: Ridge Regression, Lasso Regression, Elastic Nets

  James *et al.* (2014, Ch. 6 and 5), Hastie and Qian (2014)
- Possible non-linearities? Basis expansion and regularization (again)
  - Regression splines vs Smoothing splines (and hybrids)
  - The concept of degrees of freedom
  - Additive models (GAM)

  James *et al.* (2014, Ch. 7), Wood (2006, ch. 3, 4)
- Even more complex formulations? Tree based methods:

  - Regression trees
  - Multivariate Adaptive Regression Splines (MARS) (**Skip**)

  James *et al.* (2014, Ch. 7), Hastie *et al.* (2013, p. 295-332)

- Classification:
  - Introduction
  - From probability predictions to classification: classification rules
  - Confusion matrix and related statistics
  - The ROC Curve and related prediction measures
  - Further performance indices
  - Classification trees

4 Software

- `R`, `RStudio`
- Packages: `MASS`, `ROCR`, `mgcv`, `glmnet`, `rpart`, `earth`
- Scripts: provided by the teacher

5 Datasets

- `http://www.bee-viva.com/competitions`

References

Hastie, T. and Qian, J. (2014). Glmnet vignette. Technical report, Stanford University.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2013). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. CRC Texts in Statistical Science. Chapman & Hall, 1 edition.

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies*. Academic Press, Elsevier.