UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Revision - Week 6

1 Where to study

G James, D Witten, T Hastie, and R Tibshirani and J Friedman. *An Introduction to Statistical Learning with Applications in R*, Second Edition. Springer, Springer Series in Statistics, 2021.

- Generalized Additive Models: section 7.7.
- Classification and regression trees, random forests: chapter 8, up to page 345.

2 Exam questions

1. The prediction of a binary outcome (e.g. whether a client will default or not), also known as the classification problem, is one the most relevant problems in supervised learning.

   Explain how you perform classification using decision trees, addressing the following points:

   (a) Define the classification tree and describe the operational steps by which it is grown.

   (b) Describe how you select the optimal size of the tree.

2. What are the undesirable features of the histogram as an estimator of the density of a quantitative variable?

3. Illustrate the classification method known as k-Nearest Neighbour. What regulates the bias-variance trade-off?

4. In classification trees a popular measure of "node impurity" is the Cross-entropy or deviance:

$$-\sum_{k=1}^{K} \hat{p}_{mk} \ln \hat{p}_{mk},$$

   where $\hat{p}_{mk}$ denotes the fraction of observation in the rectangular region R_m belonging to class k.

   (a) What is the meaning of the measure and what is the range of values that the index can take?

   (b) What other measures of "node impurity" are you familiar with?

5. One of the limitations of classification trees is their volatility, which is related to the hierarchical structure of the splitting process. Small changes in the training sample produce a different sequence of splits. A possible solution is *bagging*. Provide a brief description.