Pattern Analysis and Machine Intelligence

Matteo Matteucci, Davide Eynard

09/09/2016

Please write Ex1 and Ex3 on one sheet and Ex2 and Ex4 on a different one. Indicate clearly which exercise and question you are answering in you manuscript.

1 Statistical learning (8 points)

(a) Both classification and regression problems can be solved by simple but restrictive methods as well as more complex but flexible ones. Explain which ones are better:

1. when doing *inference* or *prediction*;
2. when the irreducible error is extremely high;
3. when the number of observations is very large and the number of predictors is small;
4. when the function we need to estimate is highly non-linear.

(b) What are the *Bayes classifier* and the *Bayes error rate*? Define them and explain why the latter is defined as analogous to the irreducible error.

2 Linear regression (8 points)

(a) What is the *standard error* and how is it used to calculate a confidence interval? For instance, what does it mean to have a 95% confidence interval on the parameter β_1 of a linear regression?

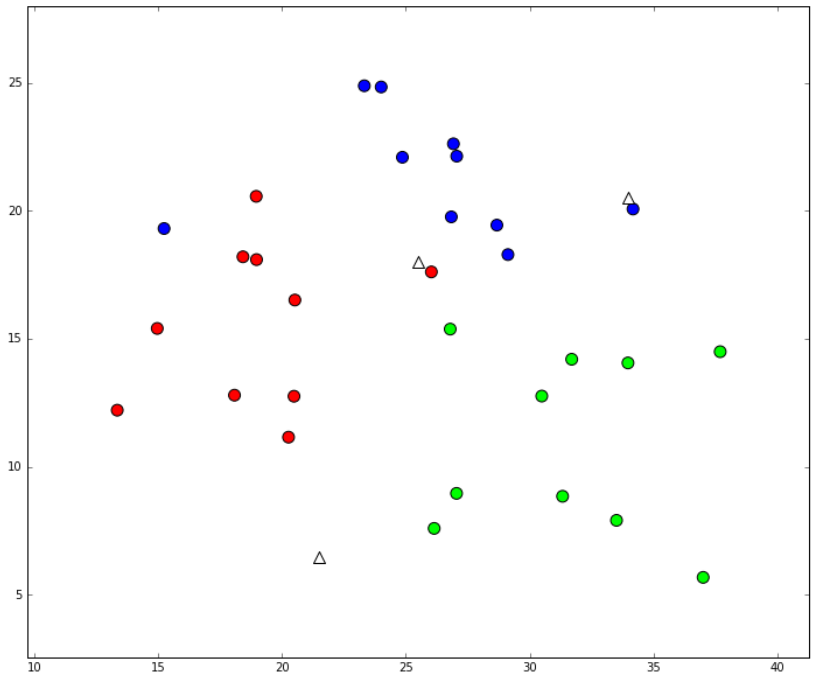
(b) Explain what the *null hypothesis* is in the context of linear regression and how it is verified.

(c) What is the *additive assumption* in a linear regression model? Show how you would detect a possible interaction between variables and how you would model it. Finally explain, with an example, how the new model would take this interaction into account.

3 Classification (8 points)

1. Given the dataset in figure, classify the three points identified with white triangles (at coordinates $(21.5, 6.5)$, $(25.5, 18)$, and $(34, 20.5)$ respectively), using the KNN algorithm with $k = 2, 3, 5$. Note: if your point has the same amount of neighbors for each class, you can assign it the class of the closest one.

2. Explain what the "curse of dimensionality" is. How would you address this problem?



4 Clustering (8 points)

1. Hierarchical clustering is not a single algorithm but rather a family of different clustering algorithms. Explain (1) how this family is composed, (2) how these algorithms work, and (3) what metrics exist to measure the distance between clusters.

2. Invent a clustering problem (for example, clustering of students according to their grades, news articles according to the words they contain, or images according to their visual descriptors). Describe the problem in detail, specifying e.g. what kind of application you are doing clustering for, the dataset size and dimensionality, what problems you might have while clustering, and so on. Then choose any two of the algorithms we have studied, and try to "sell" us one of the two, describing the characteristics of both and explaining why using one is better than the other (for instance, in terms of speed, quality of results, etc.).