Pattern Analysis and Machine Intelligence

Matteo Matteucci, Davide Eynard

30/09/2015

1 Statistical learning (8 points)

Answer the following questions

1. Describe (1) what are the *bias*, *variance*, and *irreducible error* of a model, (2) how are they related with its complexity, (3) how they are related to the expected prediction error, and (4) what is the meaning of "bias-variance tradeoff"?
2. Draw a plot of (1) bias, (2) variance, (3) training error, (4) test error, and (5) irreducible error curves as a function of increasing amount of flexibility in a statistical learning method. Explain the reason of their shapes and highlight the relationships among them.

2 Linear regression (8 points)

Given the following observations

$$x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$y = \{12, 11.2, 9.7, 8.7, 8.3, 7.4, 5.6, 5.2, 3.6, 3.3\}$$

1. Manually compute the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of a linear model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ which fits the given data
2. What is the value of MSE calculated between the values of y and the ones returned by the $\hat{y}$ function?
3. Is the trend identified by $\hat{\beta}_1$ significant or it is just due to spurious correlations? You have to provide supporting computations and justifications for your answer.

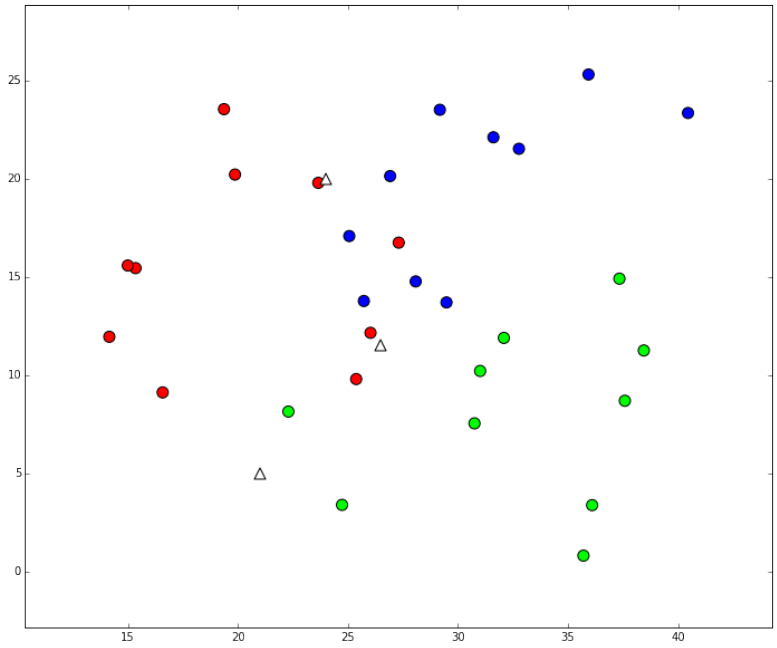
3 Classification (8 points)

(a) Given the dataset in figure, classify the three points identified with white triangles (at coordinates $(21, 5)$, $(24, 20)$, and $(26.5, 11.5)$ 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.

(b) A hospital collected data for a group of patients to study the relationship between Heart-attack Risk Index (HRI=X_1), weekly hours of physical activity (PHY=X_2), and the probability Y of having a heart attack. Roughly, for a heart attack probability to be low the HRI should be below 5, and the more hours one spends exercising the better it is. After fitting a logistic regression, the following coefficients were estimated: $\hat{\beta}_0 = -9.7$, $\hat{\beta}_1 = 1.05$, and $\hat{\beta}_2 = -0.29$.

- estimate the probability for a patient with HRI=5 and PHY=2 to have a heart attack;
- estimate how many hours of PHY a patient with HRI=7.5 should do to have that same probability.



4 Clustering (8 points)

a) 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.

b) 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.).