Pattern Analysis and Machine Intelligence

Matteo Matteucci, Davide Eynard

09/02/2015

1 Statistical learning (8 points)

Answer the following questions

1. Describe what are the *bias*, *variance*, and *irreducible error* of a model, how are they related with its complexity, how they are related to the expected prediction error, and 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 variables $x = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and $y = \{3.3, 3.6, 5.2, 5.6, 7.4, 8.3, 8.7, 9.7, 11.2\}$

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. How can we compute if the trend identified by $\hat{\beta}_1$ is significant or it is just due to spurious correlations?

To ease your computation, you can follow the following steps:

- calculate the mean $\bar{x}$ of x
- calculate the mean $\bar{y}$ of y
- calculate x-$\bar{x}$ (a vector where each value is $x_i - \bar{x}$)
- calculate y-$\bar{y}$ (as above)
- calculate $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- calculate $\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$

3 Classification (8 points)

Given the two sets of samples from classes RED = $\{(1,3),(2,2),(3.5,1),(5,4),(1.5,4),(4,2)\}$ and GREEN = $\{(2,3),(3,0.5),(4,3),(3.5,2),(1,2),(2,1)\}$, and the three unclassified elements $a = (4,1)$, $b = (2,3.5)$, and $c = (3,4)$

1. Use the KNN approach to classify the unknown items for $k = 1,3,5$. Apply the Euclidean distance as a metric (note that you can skip the actual distance calculations if you can tell the nearest neighbours at a glance).

2. Describe in details the Linear Discriminant Analysis and Logistic Regression techniques for classification and discuss how these techniques perform, in general, with respect to the KNN classifiers.

4 Clustering (8 points)

Suppose you want to evaluate the results of some clustering algorithms using SSE and Accuracy.

1. Which of these measures is defined as "internal", which is "external", and what does this mean?

2. After running your function, you obtain the result SSE=21.5. How can you evaluate whether this is a good or bad result? What would you compare this result with?

3. One of the clustering algorithms allows you to choose the number of clusters in advance. You calculate SSE after different executions of this algorithm, using K=2,3,...,10. SSE for K=10 provides the lowest value: what can you deduce from this?

4. Now suppose you have ground truth for your dataset. You run two different clustering algorithms on the same data and obtain the following results:

	SSE	Accuracy
Algorithm1	115.3	87%
Algorithm2	1285	95%

What is the meaning of these results? Which algorithm is better?