Pattern Analysis and Machine Intelligence

Matteo Matteucci, Davide Eynard

03/02/2016

1 Statistical learning (8 points)

According to statistical learning theory, in regression we assume a relationship exists between an observed variable and a dependent variable in the form

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \sim N(o, \sigma^2).$$

1. What are the *two* sources of errors we have when estimating f from data and what are these errors due to?
2. According to statistical learning theory, Test and Training Mean Squared Errors are related by the Bias-Variance trade-off; write and comment the formula representing the Bias-Variance trade off for the Expected Prediction Error in Regression.
3. The previous formula does not hold for Classification, but a useful result exists for the Classification Error Rate as well. Write and comment what statistical learning theory states about the minimum achievable average test error rate.
4. Describe in detail how the previous result for the optimal classifier is used to derive the Logistic Regression classifier and derive the shape of the decision boundary for Logistic Regression.

2 Linear regression (8 points)

Given the following observations

$$x = \{45, 54, 41, 55, 52, 56, 49, 50, 46, 47\}$$

$$y = \{108, 121, 98, 124, 124, 122, 112, 114, 105, 107\}$$

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 is it just due to spurious correlations? You have to provide supporting computations and justifications for your answer.

3 Classification (8 points)

Let consider the Linear Discriminant Analysis (LDA)classifier and the following dataset:

$$x = \{3, -1, -4, 0, 2, 5, -1, -2, -2, -2\},$$

$$y = \{A, B, C, A, A, B, A, B, B, C\}.$$

- Describe the LDA model and its underlining assumptions. What is its relationship with the Bayes classifier?
- Derive the analytical form of the decision boundary defined by the LDA classifier in a single dimension setting
- Derive the analytical form of the decision boundary defined by the LDA classifier in a multidimensional setting
- Learn an LDA classifier from the provided dataset and compute its classification error on the training set

4 Clustering (8 points)

From the very definition of what clustering is, we learned that the main purpose of a clustering algorithm is to "group objects in classes, so that intra-class similarity is maximized and inter-class similarity is minimized". Answer the following questions, *providing examples* to support your claims.

1. What is the relationship between *similarity* and *distance*? Is it always possible to calculate one from the other?
2. Do all clustering algorithms just need similarities/distances to work or are there any other conditions to be met (or parameters to be specified)?
3. The quality of a clustering algorithm is often evaluated in terms of SSE. What is it? What value of SSE is a good value? Is this measure applicable across different datasets (i.e. is SSE=5 on dataset A always better than SSE=10 on dataset B)? Explain why.
4. Suppose you run the same clustering algorithm on 1000 datasets which were randomly generated within a given interval. The 1000 SSE values you calculate fit a normal distribution with mean μ. You then run the same algorithm on a real dataset (whose data span the same interval you used to generate the random data) and get a result whose SSE is much smaller than μ. What can you conclude from this?