Pattern Analysis and Machine Intelligence

Matteo Matteucci, Davide Eynard

14/09/2015

1 Statistical learning (8 points)

In statistical learning theory, Test and Training set Mean Squared Errors are related by the so called Bias-Variance trade-off:

(a) Write and comment the formula representing the Bias-Variance trade off for the *Expected Prediction Error* in Regression

(b) The previous formula does not hold for Classification, but a useful result exists for the *Classification Error Rate*; write and comment what statistical learning theory states about the minimum achievable average test error rate.

(c) Describe in detail how the previous result for the optimal classifier is used to derive Linear Discriminant Analysis (LDA) classifier providing a detailed description of the underlining model and the shape of its decision boundary (derive it from the model).

(d) Train a LDA classifier using the data provided in the table and provide the classification error achieved by this classifier on the training data

X	class
0	A
1	A
2	A
1	A
3	A
1	B
2	B
3	B
4	B
4	C
6	C
6	C
6	C

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.

3 Classification (8 points)

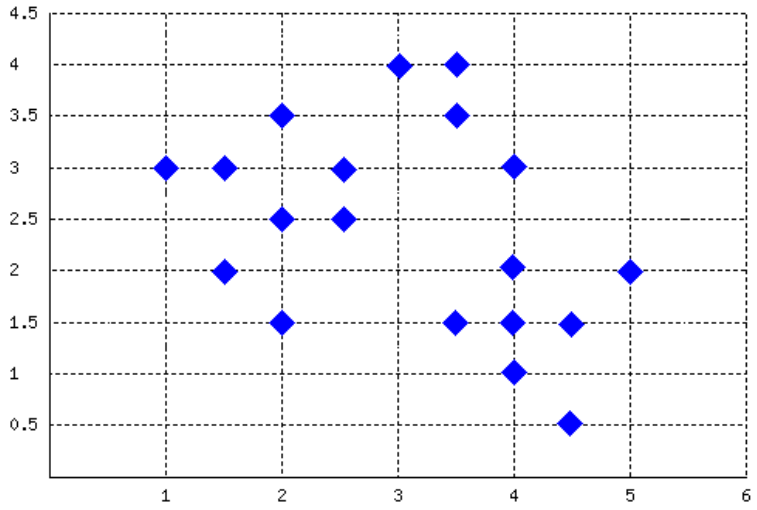
(a) What is the difference between Discriminative and Generative methods for classification? Provide one example for each category explaining why it can be considered as Discriminative/Generative.

(b) Provide a detailed description of the Generative method you previously introduced, the underlining model, the training algorithm, and derive the shape of its decision boundary.

(c) Provide a detailed description of the Discriminative method you previously introduced, the underlining model, the training algorithm, and describe the shape of its decision boundary.

4 Clustering (8 points)

Given the dataset shown in figure, execute the steps of a hierarchical (agglomerative) algorithm using the single linkage technique, showing the new links you create at each step of the algorithm (labeling them with a number) and stopping when you obtain two clusters.



Then, calculate and show the different steps of a K-Means algorithm run on the same dataset with the starting centroid positions $C_1 = (1, 1)$ and $C_2 = (5, 3.5)$, in the following way:

- at each step, specify the initial positions of the centroids
- without actually calculating it (unless it is needed to verify distances you cannot tell apart at a glance), for each step specify which centroid the various dataset points belong to
- after you have assigned points to the different centroids, calculate their new positions and proceed to next step

Are there any differences between the two algorithms? If so, how could you explain their different behavior?