Assignment 2: Bayesian Decision Theory and Parametric Estimation

Submission: Wednesday August 31st
Groups of maximum 2 students

Prof. Fabio A. González
Machine Learning - 2016-II
Maestría en Ing. de Sistemas y Computación

---

1. (2.5) Download the dataset from this link. The dataset is a text file with a number of data samples, one per line. Each line has the following structure:

$$x_i \, y_i \, C_i,$$

where $(x_i, y_i) \in \mathbb{R}^2$ and $C_i \in \{0, 1, 2\}$.

   (a) Use the data for classes 1 and 2 to estimate the parameters of a bivariate Gaussian distribution for each class. Assume that the covariance matrix is the same for both classes.

   (b) Write the parameters of the probability distribution functions for both classes.

   (c) Write a function that calculates the discriminant function for each class.

   (d) Draw a plot, where the regions corresponding to the different classes are shown with different colors. A region corresponding to a class is the set of points where the particular class discriminant function is maximum (decision regions, [Alp10] Sect. 3.4).

   (e) The boundary between both class regions must be a line. Calculate the equation of this line clearly explaining the deduction process. Draw the line along with the regions.

   (f) What happens with the boundary line if we change the prior probabilities of the classes? Illustrate with a graphical example.

2. (1.0) Repeat steps (a) to (d) from previous item, but this time:

   (a) Use data from the three classes.

   (b) Estimate a different covariance matrix for each class.

3. (1.5) Repeat the previous item, but this time:

   (a) Use only a portion of the dataset (80% of the samples) to estimate the parameters of the probability distribution functions of each class.

   (b) Write a function that calculates the discriminant function for each class, taking into account the possibility of rejection with a cost λ and cost 1 for misclassification ([Alp10] Eq. (3.10)). Look for values of λ that produce a rejection region easily distinguishable from the other regions.

   (c) Classify the rest of the dataset that was not used for estimation (20%), using a classifier based on the discriminant functions. Evaluate the results using a confusion matrix.

4. The assignment must be submitted as an IPython notebook through the following Dropbox file request, before midnight of the deadline date. The file must be named as `ml-assign2-unalusername1-unalusername2.ipynb`, where `unalusername` is the user name assigned by the university (include the usernames of all the members of the group).

References

[Alp10] Alpaydin, E. 2010 Introduction to Machine Learning, 2Ed. The MIT Press.

[DHS00] Duda, R. O., Hart, P. E., and Stork, D. G. 2000 Pattern Classification (2nd Edition). Wiley-Interscience.