Assignment 2: Bayesian Decision Theory and Parametric Estimation

Submission: Tuesday May 7th
Groups of maximum 3 students

Prof. Fabio A. Gonzalez
Machine Learning - 2019-I
Maestr a en Ing. de Sistemas y Computacion

Load the Iris dataset from sklearn and create a dataset using only the Petal Width and Sepal Width attributes. Plot the resulting dataset assigning a di erent color to each class.

1. (2.0)

   (a) Assume that each class is generated by a bivariate Gaussian distribution with the a covariance matrix $= I$, where is a scalar, shared by all the classes. This is, the distribution for each class has a di erent mean but the same covariance matrix. Calculate the parameters of the probability distribution functions for the three classes.

   (b) Write a di erent Python functions that calculate the discriminant function for each class.

   (c) Draw a plot, where the regions corresponding to the di erent classes are shown with di erent colors. A region corresponding to a class is the set of points where the particular class discriminant function is maximum (decision regions, [Alp14] Sect. 3.4).

   (d) The boundary between class regions must be a line. Calculate the equation of these lines clearly explaining the deduction process. Draw the lines along with the regions.

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

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

   (a) Write again Python functions that calculate the discriminant function for each class, but now, taking into account the possibility of rejection with a cost and cost 1 for misclassi cation ([?] Eq. (3.10)). Look for values of that produce a rejection region easily distinguishable from the other regions.

3. (2.0) Repeat the previous item, but this time:

   (a) The covariance matrix could be an arbitrary matrix (not diagonal) and di erent for each class.

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

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

The assignment must be submitted as a Jupyter notebook through the following Dropbox le request, before midnight of the deadline date. The le must be named as `ml-assign2-unalusername1-unalusername2-unalusername3.ipynb`, where `unalusername` is the user name assigned by the university (include the usernames of all the members of the group).