FEUP 2021/2022 (1ST SEMESTER)

Assignment 02

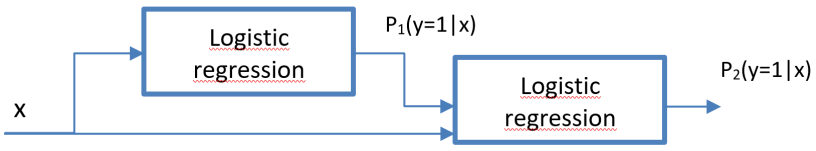
To be solved in groups of at most three elements

Submit by November 18, 2021, 23h59 in TEAMS. Submit a SINGLE zip file, named with the name of the students. Only a member of the group submits the work. The other member(s) of the group only submit a txt file stating "joint submission with [Colleague Name]".

1. Consider the data in 'heightWeightData_short.txt'. The first column is the class label (1=male, 2=female), the second column is height, the third weight.
   a) Write a Python function to model each class data as follows: assuming that height and weight are independent given the class, model the height using a histogram with bins breakpoints at every 10 cm $(10, 20, 30, \cdots, 170, 180, \cdots)$ and the weight with the parzen window method using a Gaussian kernel with the bandwidth set to 1.5. You can use suitable functions in Python like matplotlib.pyplot.hist and sklearn.neighbors.KernelDensity. The function should receive as input the training data and the test data, making prediction (male/female) for the test point.
   b) Use the previous function to make predictions (male / female) for the following test points: $[165\ 80]^t$, $[181\ 65]^t$, $[161\ 57]^t$ and $[181\ 77]^t$.
   c) What's the estimated $p([181\ 65]^t|(female))$?

2. In a two-dimensional classification problem, class C_1 has mean $\mu_1 = \begin{bmatrix} a \\ 0 \end{bmatrix}$ and covariance $\Sigma_1 = \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix}$, and class C_2 has mean $\mu_2 = \begin{bmatrix} c \\ 0 \end{bmatrix}$ and covariance $\Sigma_2 = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix}$. The variables a, b, c, and d are all greater than 0. Prove that the MICD decision boundary for this problem is a circle, except in special cases, and determine its centre and radius. Note: the MICD classifier is defined by the following decision rule:
   $x \in C_1$ iff $(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) < (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)$

3. Consider the logistic regression model for classification and the cascaded model in the figure. In the following use you own implementation of the logistic regression or the provided implementation with the Newton optimization method and without regularization.



   a) What kind of boundaries is a logistic regression model able to express?
   b) Is the cascaded model able to express boundaries more complex than those expressed by the logistic regression model?
   c) Repeat last item assuming that in the first block we have a linear regression model.
   d) Train the cascaded model and a simple logistic regression model in the first 80% of the data (dataset.txt, where last column is the class) and evaluate the quality in the remaining 20%. To train the cascaded model, train first the first block, then, with the first block fixed, train the second. What's the estimated accuracy of each?
   e) What's the prediction of each model in the point $x = [87, 41]^t$?
   f) Is the optimization adopted in item d) for the cascaded model the best approach? Can you propose another?