COLORADO SCHOOL OF MINES

MATH437/537 **Fall, 2022**

Homewrok 6. Due November 21

1. For this question you will need the data sets (in CANVAS) `hmk6q1.txt` (training set) and `hmk6q1test.txt` (testing set) consisting of values of variables X_1 and X_2 for classes $C = 1$ and $C = 2$. Assume $p_1 = p_2$, $c(1|2) = 2c(2|1)$.

   (a) Determine the equation of the line

   $$\boldsymbol{a}_*^\top \boldsymbol{x} = \boldsymbol{a}_*^\top (\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2)/2 + \log\left(\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right)$$

   defined by the sample linear discriminant function, where $\boldsymbol{a}_* = \boldsymbol{S}_{\text{pooled}}^{-1}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)$.

   (b) Make a scatterplot of the training set using points of different colors for the different classes. Draw the line from (a) on the same plot.

   (c) Make a scatterplot of the testing set using points of different colors for the different classes. Draw the line from (a) on the same plot.

   (d) Determine the APER (you can get this from the plot).

   (e) Determine the 1-cross-validation estimate of the AER using only the training set (you will have to write your own code using the algorithm described in class).

2. Consider two classes C_1 and C_2 with densities

   $$\begin{aligned} f_1(x) &= 1 - |x| & -1 \le x \le 1 \\ f_2(x) &= 1 - |x - 1/2| & -1/2 \le x \le 1/2 \end{aligned}$$

   (a) Draw the two densities on the same plot.

   (b) Identify the optimal classification regions for the case $p_1 = p_2$, $c(1|2) = c(2|1)$.

   (c) Identify the optimal classification regions for the case $p_1 = 0.2$, $c(1|2) = c(2|1)$.

3. Consider a classification problem with four classes $C_1, \ldots, C_4$ with prior probabilities

   $$p_1 = 0.2, \quad p_2 = 0.2, \quad p_3 = 0.3, \quad p_4 = 0.3$$

   and densities $f_1, \ldots f_4$. Suppose you are given a data point $\boldsymbol{x}$ such that

   $$f_1(\boldsymbol{x}) = 0.3, \quad f_2(\boldsymbol{x}) = .36, \quad f_3(\boldsymbol{x}) = 0.8, \quad f_4(\boldsymbol{x}) = 0.7.$$

   Determine the posterior probabilities. To which class would you assign $\boldsymbol{x}$?

4. Consider a classification problem with three classes C_1, C_2 and C_3 with the following costs and probabilities:

$$c(2|1) = 250,\ c(3|1) = 50,\ c(1|2) = 5,\ c(3|2) = 25,\ c(1|3) = 25,\ c(2|3) = 100,$$

$$p_1 = 0.05,\ p_2 = 0.60,\ p_3 = 0.35.$$

You are given a data point $\boldsymbol{x}$ whose density values are $f_1(\boldsymbol{x}) = 0.01$, $f_2(\boldsymbol{x}) = 0.85$ and $f_3(\boldsymbol{x}) = 1$. To what class would you assign $\boldsymbol{x}$?