

M3/4S7: Project 1

Hand-out date: 29th Jan 2015

Hand-in date: 16th Feb 2015

Notes: SUBMISSION OF THIS PROJECT FORMALLY ENROLS YOU AND COMMITS YOU TO M3/4S7.

Hand in a tidy report that includes your answers, graphs and the R code you used.

Separately, e-mail me with the code you used attached as a *single* stand alone script file. I should be able to execute this, without making any modifications to the code. The subject of the e-mail must be in the format

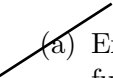



S7: YOUR NAME: PROJECT 1

..... and don't forget to attach the script file.

1. Consider a two-class classification problem, with priors $P(C_1) = 0.5$ and $P(C_2) = 0.5$ and class conditional densities given by

$$p(x|C_i) = f_i(x; \sigma_i) = \frac{x}{\sigma_i^2} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) \quad 0 \leq x < \infty$$

for $i = 1, 2$, where $\sigma_1 > \sigma_2 > 0$.

-  (a) Express the decision threshold $T_{\min} > 0$ for minimum error as a function of σ_1 and σ_2 .
-  (b) In terms of F_i , the cumulative distribution functions of the class conditional densities, obtain an expression for the error rate associated with some decision threshold, $T \geq 0$. For $\sigma_1 = 4$, and three separate cases of (i) $\sigma_2 = 0.5$, (ii) $\sigma_2 = 1$ and (iii) $\sigma_2 = 1.5$, plot the error rate as a function of T for $0 \leq T \leq 4$ using the same axes for all three plots.
-  (c) Give an expression for the Bayes error rate for general σ_1 and σ_2 . Your answer should be given in terms of $r = \sigma_1/\sigma_2$.
-  (d) For $\sigma_1 = 4$ and $\sigma_2 = 1$, compute the value of the minimum error threshold and the Bayes error rate to 2 decimal places. Construct a plot of the prior weighted class conditional densities and construct a plot of the posterior densities. Include a marker for the minimum error decision threshold in both plots.

Question 1 continues on next page

(e) Suppose we now associate the following costs with class allocation

		TRUE	
		C_1	C_2
PREDICTED	C_1	0	5
	C_2	1	0

For general $\sigma_1 > \sigma_2 > 0$ express in terms of σ_1 and σ_2 the Bayes decision threshold for minimum risk.

Using the values of $\sigma_1 = 4$ and $\sigma_2 = 1$, compute this threshold and place a marker for it in both the plots you made in part (d).

TOTAL [13]

2. Consider a two class classification problem with equal priors and bivariate normal class conditional densities

$$p(\mathbf{x}|C_1) \sim N\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \Sigma_1\right) \quad \text{and} \quad p(\mathbf{x}|C_2) \sim N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_2\right)$$

where

$$\Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}.$$

- (a) Sample a random vector, \mathbf{x}^* , uniformly from the rectangular region $(-6, 6) \times (-6, 6)$. Compute an appropriate discriminant score, $g_i(\mathbf{x})$, for \mathbf{x}^* for C_i , $i = 1, 2$. On the basis of these discriminant scores, to which class should \mathbf{x}^* be assigned?
- (b) Construct a plot that displays

$$g_1(\mathbf{x}) - g_2(\mathbf{x})$$

evaluated over a regular grid on $(-6, 6) \times (-6, 6)$. Add a marker for \mathbf{x}^* to the plot.

TOTAL [6]

3. Consider a 2-class p -dimensional problem (p will be ≥ 10 in this question) with $P(C_1) = 0.25$ and $P(C_2) = 0.75$, with **multivariate normal class conditional densities**. Mean vector $\boldsymbol{\mu}_1$, is a p -dimensional vector of zeros, and $\boldsymbol{\mu}_2$ is yet unspecified. Covariance matrix $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$, where \mathbf{I}_p is the p dimensional identity matrix. The diagonal elements of $\boldsymbol{\Sigma}_2$ are all equal to 0.5, and all off-diagonal elements are zero, except for $\sigma_{13} = \sigma_{31} = -0.3$, $\sigma_{23} = \sigma_{32} = 0.2$ and $\sigma_{48} = \sigma_{84} = 0.1$.
- Select an integer uniformly from the set $\{10, 11, 12, 13, 14, 15\}$ for p .
 - Select a random vector of length p uniformly from the region defined by $(0, 1)^p$, for $\boldsymbol{\mu}_2$.
 - For $n = 100, 200, 300, 400, 500, 600, 700, 800, 900$ and 1000 do the following;
 - Obtain a training sample of n observations, and a test sample of 4000 observations, in the same ratio as the prior probabilities.
 - Compute the out-of-sample empirical error rate, using the test set, for linear and quadratic discriminant functions estimated from the training data.
- Hint: Do steps (c)i. and (c)ii. within a single for-loop.*
- On the same set of axes plot the training set size on the x-axis and the out-of-sample empirical error rate on the y-axis for both your linear discriminant analysis and your quadratic discriminant analysis.
 - Comment on your results.
 - Compute an estimate of the Bayes error rate.

TOTAL [13]

4. Suppose we are interested in finding a minimum of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ on an interval (a, b) . Write an R function that uses the *Golden ratio bracketing* algorithm to determine the location of the minimum. The bracketing function should be written to take a target function f as an argument.

Test the R function using

$$f(x) = \exp(-x) \sin(x) + \sinh(x)$$

to find the minimum in the interval $(-4, 0)$.

TOTAL [4]

GRAND TOTAL [36]