STAT5P87 Assignment II

Brock University
STAT5P87 - Winter 2025
Assignment II

Due Date: Sunday February 16

- **Instructions:**
  - Assignments should be submitted electronically as an **single** R script (.R file)
    - Please be sure to use R version 3.6.1 or greater to ensure random number generation consistency.
    - Solutions may use the `glmnet` package, but no additional packages.
    - Written parts of the solutions should be included in the script as comments.
    - I will make sure that R can find and load the required data sets, otherwise the code should run as submitted.
  - The weight of each question is indicated in bold [#].
  - Coding style and efficiency will account for [**4**] marks on the assignment.
  - Style Requirements:
    - comments describing the input and output of defined functions
    - appropriate use of white space (spaces, line breaks)
    - indentation of functions, for loops, etc.

---

1) [**3**] A *mixture* distribution occurs when observations are randomly drawn from one of a set of possible distributions. Write a function that takes as input two vectors $\mu = (\mu_1, \mu_2)$ and $\sigma^2 = (\sigma_1^2, \sigma_2^2)$, a probability ($p$) and a sample size ($n$), and simulates n observations from a mixture of two Normal distributions, $\mathrm{N}(\mu_1, \sigma_1^2)$ with probability p and $\mathrm{N}(\mu_2, \sigma_2^2)$ with probability $1 - p$.

2) [**8**] Write a function to generate the folds for K-fold cross-validation. The function should have four inputs: for the first input, the user should be able to specify either a sample size or a vector of outputs, the second input should be the number of folds, the third input an option for whether the folds should be stratified (default `FALSE`) and the fourth input a seed for the random number generator (default 0). The stratified option should only be used for categorical inputs, if it is set as `TRUE` for continuous inputs, then it should be deactivated and give the user a warning.

3) The Bayes classifier is the optimal classifier when the underlying probability model is known,

$$\delta_k = \mathrm{P}(y = k \mid \boldsymbol{x}) = \frac{\mathrm{P}(\boldsymbol{x} \mid y)\,\mathrm{P}(y = k)}{\mathrm{P}(\boldsymbol{x})}.$$

In this question we will perform a simulation study to compare the Bayes optimal classifier with LDA. In other words, we investigate the loss of precision due to incorrectly assuming $\boldsymbol{x}$ follows a normal distribution. For this simulation study, we will have the following **known** probability model: y is a binary categorical variable with $\mathrm{P}(y = 0) = 0.4$ and $\mathrm{P}(y = 1) = 0.6$ and x is a univariate input where $\mathrm{P}(x \mid y = 0)$ is an equal-probability mixture of Normal distributions with $\mu = (0.2, 0.6)$ and $\sigma^2 = (0.04, 0.09)$, and $\mathrm{P}(x \mid y = 1)$ is an equal-probability mixture of Normal distributions with $\mu = (0.5, 0.8)$ and $\sigma^2 = (0.04, 0.01)$.

a) **[3]** Describe the decision boundary for the Bayes classifier of the above model (e.g., for which x values does the classifier predict $y = 0$ or $y = 1$).

b) **[4]** Simulate training ($n = 200$) and testing ($n = 1000$) data from the above model. Use the training data to fit an LDA classifier (no regularization) and compare the accuracy of LDA and Bayes classifiers on the testing data.

4) **[4]** This is a classification problem, where the response variable is one of $K = 11$ 'stable vowel sounds'. The input $\boldsymbol{x}$ consists of $p = 9$ input variables derived from an audio recording of a speaker making the appropriate sound. For this problems, we will use a subset of the vowel data, available in 'a2-vowel-data.csv' on Brightspace. Use regularized logistic regression to create a classifier, and 5-Fold cross-validation to evaluate the model. Report the maximum value of accuracy and corresponding value of λ.