STAT 3690 Homework 4

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Due at Apr 25 11:59 pm (Central Time)

Answers must be submitted electronically via Crowdmark. Please enclose your R code trunks (if applicable) as well.

1. We have information on $n = 138$ samples of Canadian hard red spring wheat and the flour made from these samples. The 5-dimensional vector $\mathbf{X}$ contains **standardized** wheat measurements on: (X_1) kernel texture, (X_2) test weight, (X_3) famaged kernels, (X_4) foreign material and (X_5) crude protein in the wheat. The 4-dimensional vector $\mathbf{Y}$ contains **standardized** flour measurements: (Y_1) wheat per barrel of flour, (Y_2) ash in flour, (Y_3) crude protein in flour, and (Y_4) gluten quality index. We are only given the sample correlation matrices:

$$R_X = \begin{pmatrix} 1.000 & 0.754 & -0.690 & -0.446 & 0.692 \\ & 1.000 & -0.712 & -0.515 & 0.412 \\ & & 1.000 & 0.323 & -0.444 \\ & & & 1.000 & -0.334 \\ & & & & 1.000 \end{pmatrix},$$

$$R_Y = \begin{pmatrix} 1.000 & 0.251 & -0.490 & 0.250 \\ & 1.000 & -0.434 & -0.079 \\ & & 1.000 & -0.163 \\ & & & 1.000 \end{pmatrix},$$

$$R_{XY} = \begin{pmatrix} -0.605 & -0.479 & 0.780 & -0.152 \\ -0.722 & -0.419 & 0.542 & -0.102 \\ 0.737 & 0.361 & -0.546 & 0.172 \\ 0.527 & 0.461 & -0.393 & -0.019 \\ -0.383 & -0.505 & 0.737 & -0.148 \end{pmatrix}.$$

   a. Use sequential tests (with the Holm-Bonferroni procedure) to determine the number of significant canonical correlations at level $\alpha = .05$.
   b. Compute sample canonical directions corresponding to the significant canonical correlations.

2. Consider the situation where you have two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. We observe independent samples $X_{1,1}, \ldots, X_{1,n_1} \sim N(\mu_1, \sigma_1^2)$ and $X_{2,1}, \ldots, X_{2,n_2} \sim N(\mu_2, \sigma_2^2)$ with means $\bar{X}_1$ and $\bar{X}_2$, respectively. We plan to use the following rule

   R : Classify a new X as coming from population 2 if $X > (\bar{X}_1 + \bar{X}_2)/2$ and population 1 otherwise.

   Assuming priors $\Pr(X \sim N(\mu_1, \sigma_1^2)) = \Pr(X \sim N(\mu_2, \sigma_2^2)) = 1/2$, please express the misclassification rate of rule R, i.e.,

   $$\mathrm{err}(X) = \Pr(X > (\bar{X}_1 + \bar{X}_2)/2 \text{ and } X \sim N(\mu_1, \sigma_1^2)) + \Pr(X \le (\bar{X}_1 + \bar{X}_2)/2 \text{ and } X \sim N(\mu_2, \sigma_2^2)),$$

   in terms of n_1, n_2, μ_1, μ_2, σ_1, σ_2 and the standard normal cumulative distribution function $\Phi(\cdot)$.

3. Suppose there is a binary classification task: one would like to predict labels of n subjects, say $Y_1, \ldots, Y_n$, according to their independent p-dimensional observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. The two potential populations are assumed to be $MVN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $MVN_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, i.e., $\mathbf{X}_i \mid Y_i = y_i \sim MVN_p(\boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma})$, $y_i = 1, 2$. Meanwhile, let $\Pr(Y_i = k) = \pi_k$ for all $k = 1, 2$ and $i = 1, \ldots, n$.
   a. Applying the linear discriminant analysis (LDA) to this problem, write down the mathematical expression of error rate in terms of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$, π_1, π_2 and the standard normal cumulative distribution function $\Phi(\cdot)$.
   b. There is a banknote authentication dataset (see below for the data import), where $n = 1,372$ data points consisted of features extracted (via the wavelet transformation) from images that were taken from genuine and forged banknotes. Specifically, the features are "variance" (the variance of wavelet-transformed image), "skewness" (the skewness of wavelet-transformed image), "curtosis" (the curtosis of wavelet-transformed image), and "entropy" (the entropy of image), all continuous. The authentication of banknote is indicated by "class" (0 for authentic and 1 for inauthentic). Figure out a parametric estimate for error rate of LDA by plugging estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$, π_1 and π_2 into the expression obtained in Q3a.
   c. Apply LDA to the dataset in Q3b and utilize 5×8-fold cross validation to estimate the resulting error rate. Report this error rate.
   d. Make a comment with one single sentence after comparing estimates given by Q3b and Q3c.

```r
bn_df = read.table(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt",
  sep = ","
)
names(bn_df) = c("variance", "skewness", "curtosis", "entropy", "class")
```