Stat/CS 5525 Homework 3

(Due on October 24th, 2023)

1. Ex 4.2 (a), (b), (c) from textbook.

   Hint to (b): "the target coded as $-N/N_1$, N/N_2" means that $y_i = -N/N_1$ if the point belongs to class 1 and $y_i = N/N_2$ belongs to class 2. Arrange the response vector as

   $$\mathbf{y} = \begin{pmatrix} -\frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_2}\mathbf{1}_{N_2} \end{pmatrix}.$$

   Here $\mathbf{1}_M$ denote the vector of 1's with length M. Then the model is

   $$\boldsymbol{y} = \beta_0 \mathbf{1}_N + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

   Get the estimate for $\boldsymbol{\beta}$.

   Hint to (c): one can use the formula:

   $$(\mathbf{A} + \mathbf{b}\mathbf{b}^T)^{-1} = \mathbf{A}^{-1} - (1 + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b})^{-1}\mathbf{A}^{-1}\mathbf{b}\mathbf{b}^T\mathbf{A}^{-1}.$$

2. Use logistic regression to analyze the data "admit.txt".

   Data background: a researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution (rank), affect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

   Show the estimation of the coefficients as in Table 4.2 in the textbook, and write the log-ratio as in eq. (4.17) in the textbook.

3. Randomly partition the data in Problem 2 into two parts (with equal sample size), one as a training set, and the other as the test set. Compare the the performance of logistic regression and LDA in terms of the test error. Interpret your results.

4. [**For fun**] Try the l_1 and l_2 regularized logistic regression for the data in Problem 3. Compare the performance. (*Remark*: one can use *Glmnet* package if coding with R.)