Homework 1

Each problem is worth 20 points, except the last one, which is worth 30 points. Each optional part is worth 10 points.

1. 2.8, with the linear regression replaced by the (multinomial) logistic regression.

   Using the training and test data for the **3**s, **5**s and **8**s only, compare the performance of multinomial logistic regression and k-NN with $k = 1, 3, 5, 7, 15$ in terms of training error and test error. (Hint: *You can call* `multinom` *in R package* `nnet` *to fit a multinomial logistic regression.*)

2. 4.3. You can either do a computer experiment or prove the conclusion.

   (Hints for the theoretical part: (i) Comparing the discriminant functions shows that proving $B(B^T\Sigma B)^{-1}B^T\mu_k = \Sigma^{-1}\mu_k$ is enough. (ii) Let $U = X(X^TX)^{-1/2}$, $V = Y(Y^TY)^{-1/2}$. It suffices to show the conclusion for such column-normalized predictors and responses. (iii) Notice that $\Sigma = (U - P_YU)^T(U - P_YU) = U^T(I - P_Y)U$, and $\mu_k = U^TV\alpha \in span(U^TV)$. Use the SVD of U^TV.)

3. 4.9. Do not use any available `qda` function. The multi-class vowel recognition problem (with $p = 10$ and $K = 11$) is notoriously difficult.

4. 13.7 (1). Part 2 and part 3 are optional.

5. Recall that the LDA discriminant function $\delta_k(x)$ (Eq. (4.10), page 109) can be written as $\delta_k(x) = x^T\Sigma^{-1/2}\beta_k + \alpha_k$, where $\beta_k \in \mathbb{R}^p$ and $\alpha_k \in \mathbb{R}$. $\Sigma, \beta_k, \alpha_k$ can be easily estimated from the data. Consider a *reduced* k-NN classifier as follows.

   Calculate all pairwise differences between β_k ($1 \le k \le K$) to form a matrix N of size $p \times K(K-1)/2$. Let $r = rank(N)$. Project all data points onto the range (column space) of N. The transformed predictor matrix is denoted by $Z \in \mathbb{R}^{n\times r}$. Apply k-NN on the z-samples.

   (a) Compare the practical performance (prediction error) of LDA, k-NN and the LDA-reduced k-NN on the `zipcode` data. (Alternatively, you may perform a simulation experiment in the spirit of Problem 4. Set p somewhat large.)

(b) What if we normalize all observations first before doing the projection?

Hint: In implementation, it suffices to form $N \in \mathbb{R}^{p\times(K-1)}$ using the differences between β_k and β_K, $1 \le k \le K-1$. Let $N = UDV^T$ be the (reduced form) SVD, where D is an $r \times r$ diagonal matrix. Then $Z = XU$. (More preferably, QR decomposition can be used.)