Homework #4 (BST 263, Spring 2019)

1. Consider the LDA model. Suppose we know all the parameters: that is, we know π_k and μ_k for each k, as well as the covariance matrix C. Starting from the formula

$$\mathbb{P}(Y_0 = k \mid x_0) = \frac{\mathcal{N}(x_0 \mid \mu_k, C)\,\pi_k}{\sum_{j=1}^{K} \mathcal{N}(x_0 \mid \mu_j, C)\,\pi_j},$$

show that choosing k to maximize $\mathbb{P}(Y_0 = k \mid x_0)$ is equivalent to

$$\underset{k}{\text{argmax}} \left(a_k + x_0^{\mathsf{T}} b_k\right)$$

where $a_k = -\frac{1}{2}\mu_k^{\mathsf{T}} C^{-1}\mu_k + \log(\pi_k)$ and $b_k = C^{-1}\mu_k$. Notation: $\mathcal{N}(x|\mu, C)$ is the density of $\mathcal{N}(\mu, C)$ at x.

2. Write an R function to estimate the LDA model parameters (i.e., the π's, μ's, and C) using the formulas from the slides, and compute a_k and b_k for each k using the formulas in problem 1 above. The inputs should be `x` ($d \times n$ matrix of training points), `y` (length n vector of training point classes in $\{1, \ldots, K\}$), and `K` (number of classes). The outputs should be `a` (length K vector where `a[k]` $= a_k$) and `b` ($d \times K$ matrix where `b[,k]` $= b_k$). (R code tip: You can return multiple arguments from a function by using `return(list(a=a, b=b))`.)

3. Write an R function to implement the LDA prediction rule from problem 1 above. The inputs should be `x0` (length d vector at which to predict the class), and `a` and `b` from your function in problem 2. The output should be `y0` (predicted class in $\{1, \ldots, K\}$).

4. Run your LDA algorithms (from problems 2 and 3) on the training and test data from problem 4 of Lab 1. Compare the test performance of LDA and KNN with $K = 9$ nearest neighbors in two cases: (a) $d = 2$ and (b) $d = 20$. Make plots of the LDA and KNN predictions as in problem 6 of Lab 1, for each case (a) and (b). Write 2 or 3 paragraphs discussing the numerical results and plots. Your discussion should touch on things like how well each model matches the true data generating process, flexibility and bias-variance tradeoffs, and interpreting the numerical results using the plots to try to explain what is happening.

5. ISL chapter 4, problem 10. You can use the R packages for logistic regression, LDA, QDA, and KNN; see ISL section 4.6 for details on using these packages.