Suggestions - Problem Set 4

#1: Justify the discussion on p. 114 regarding the variable $Z = a^T X$.

This justifies a formula of Fisher for projecting an LDA problem into one dimension. Let your feature vectors $X = x \in \mathbb{R}^p$, and make the usual LDA assumption, i.e., of groups $G_1, \ldots, G_k$, for each of which the feature vectors $x$ have their own distribution. The distribution of vectors $x$ from group $k$ is as usual $\mathcal{N}(\mu_k, \Sigma)$. The class centroid (center) of group $k$ is $\mu_k$. We wish to project the onto a single dimensional subspace in the direction of the unit vector $a \in \mathbb{R}^p$, chosen for maximal distance between the projections $a^T \mu_k$, and minimal variance of the projected points from their respective means. Show that if $W = \Sigma$ is the common covariance of the vectors $x \in \mathbb{R}^p$, then $a^T Wa$ will be the common covariance of the projected vectors $a^T x$. Look at the covariance structure of the vectors defining the $\mu_k$. Show this is an empirical covariance matrix $V(\mu)$ of the $K$ class centroid vectors $\mu_k$ given by

$$B = \frac{1}{K-1} \sum_{j=1}^{K} (\mu_j - \nu)(\mu_j - \nu)^T,$$

where $\nu = \frac{1}{K} \sum_k \mu_k$. Show we can also write $B = V(M^T)$, where $M^T = [\mu_1, \mu_2, \ldots, \mu_K]$ has column vectors $\mu_k$. By the variance of a matrix like $M^T$ we mean the empirical covariance of its columns. Show the variance of the projections of the $\mu_k$ onto $a$ is the empirical covariance of $a^T M^T$, projecting each $\mu_k$ onto the vector $a$. Show the empirical covariance $a^T \mu_k$ is $V(a^T M^T) = a^T V(M^T) a = a^T B a$. Thus show Fisher's goal is to find the direction $a$ so projecting onto $a$ maximizes the ratio of projected mean variance $a^T B a$ to the common projected covariance, or $a^T Wa$. Explain your reasoning in detail.

4.5. Consider the log likelihood in (4.20). Assume $x_i$ in class $y_i = 1$ are perfectly separated from those with $y_i = 0$ by a point $x_0$ in one dimension, or more generally by a hyperplane $L$ in $p$ dimensions. For all $\beta = (\beta_0, \beta_1)$, we have $\ell(\beta) \leq 0$ - why? (look at (4.20) — notice all the probabilities have to be 1 or less). Show that

$$p(x_i; \beta) = \frac{e^{\beta_0 + \beta_1^T x}}{1 + e^{\beta_0 + \beta_1^T x}}.$$

Logistic regression classifies the test point $x$ to be in class $y = 1$ if $p(x; \beta) > 1/2$, otherwise class $y = 0$ (why?). Show the separator between the regions is $\beta_0 + \beta_1^T x = 0$. Now show that for any $\beta$ with a separating hyperplane

$$L : \beta_0 + \beta_1^T x = 0$$
which perfectly separates these classes, you can multiply $\beta$ by a constant $a$, and will still have the same $L$. Now show this replacement as $a \to \infty$, has the property that for $x$, with $y_i = 1$, $p(x, a\beta) = \frac{1}{a} \to 1$, with an analogous limit if $y_i = 0$.

Show that $\lim_{a \to \infty} \ell(a\beta) = 0$. Conclude that for any initial $\beta$ for which $L$ separates the dataset perfectly, a better choice will be $a\beta$, with $a \to \infty$.

You can give a similar argument for $K$ classes, with parameters $\beta_1, \ldots, \beta_{K-1}$. If the corresponding separators separate the $K$ classes perfectly, then a similar replacement of $\beta_k$ will increase $\ell$ to its maximum as $a \to \infty$.

This is the problem with separable data — unless we regularize optimization of $\beta$, increasing it arbitrarily will give better and better likelihoods.

5.4. (a) Recall natural splines are piecewise cubic on $[a, b]$ with knots at $\{\xi_1, \ldots, \xi_K\}$ which are restricted as linear in $[a, \xi_1]$ and in $[\xi_K, b]$. Show linearity in the first interval is equivalent to $\beta_2 = \beta_3 = 0$. In $[\xi_K, b]$ how does $f(X)$ simplify? Use this fact to get the remaining conditions.

(b) For the basis functions in (5.4), you can prove they are a basis for natural splines by showing (i) they are in the space, (ii) there are $N$ of them, and (iii) that they are linearly independent. Why are these sufficient? For (i) recall the conditions in (a) above. For linear independence, why is it enough to show that each function in list (5.4) is linearly independent of the previous ones?

To show that each $N_k$ is linearly independent of the previous $N_1, \ldots, N_{k-1}$, note $N_1 = 1, N_2 = x$ while

$$N_3(x) = d_1(x) - d_{K-1}(x) = \frac{(x - \xi_1)^3}{\xi_K - \xi_1} - \frac{(x - \xi_K)^3}{\xi_K - \xi_{K-1}},$$

so $N_3$ has the property that it has a discontinuity in its third derivative at $\xi_1$, because of the term $(x - \xi_1)^3$. Since $N_1$ and $N_2$ do not have such discontinuities, it follows that $N_1$ cannot be a linear combination of $N_0, N_1$. Similarly, $N_4$ has a third derivative discontinuity at $\xi_2$, and so cannot be a combination of the previous functions, none of which have derivative discontinuities at $\xi_4$. Continue this way, until noting that $N_{K-1} = d_{K-3}$ and $N_K$ has a third derivative discontinuity at $\xi_{K-3}$, while none of the previous $N_k$ do, and $N_K$ has a third derivative discontinuity at $\xi_{K-2}$, while none of the previous ones do. Thus conclude each function is linearly independent of the previous ones. How does this complete the proof?