HW2

(Due October 26, midnight)

Problem 1 (Ridge logistic regression): Let $\boldsymbol{X}$ be the $n \times p$ feature matrix for a logistic regression, with $p \gg n$, and $\boldsymbol{y}$ the response vector. Assume $\boldsymbol{X}$ has row-rank n. We consider the ridge logistic regression problem

$$\max_{\beta} \left\{ \ell(\boldsymbol{y}, \beta_0 + \boldsymbol{X}\beta) - \lambda \beta^T \beta \right\}$$

where ℓ is the binomial log-likelihood function.

1. Show that the solution is equivariant with respect to an orthogonal $p \times p$ transformation Q: if $\hat{\beta}$ solves the original problem, then $\tilde{\beta} = Q^T \hat{\beta}$ solves the problem with $\tilde{\boldsymbol{X}} = \boldsymbol{X}Q$.

2. Consider the full QR decomposition of $\boldsymbol{X}^T$:

   $$\boldsymbol{X}^T = QR = (Q_1, Q_2) \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

   where Q_1 is $p \times n$ and hence R_1 is $n \times n$, and non-singular. Show that one can solve the logistic regression with R_1 rather than $\boldsymbol{X}$. How would you recover the original solution?

3. Assuming the data are separable, what happens to the solution $\hat{\beta}_\lambda$ as $\lambda \to 0$ [Qualitatively, the next two questions focus on the precise behavior].

4. Set up a small simulation with separable data (you don't need to have p > n for this part). Fit the SVM optimal separating hyperplane (for example, use function svm() in package e1071 with a large value for the cost parameter), and extract the unit vector β_{svm} normal to the separating hyperplane. Now fit a series of ridged logistic regression models with λ decreasing toward 0 (use a fine grid on the log scale). (The package glmnet might be useful here, with alpha=0, standardize=FALSE, and perhaps providing your own sequence of values for lambda). For each of your solutions, compute

   $$\theta_\lambda = \frac{\hat{\beta}_\lambda}{\|\hat{\beta}_\lambda\|_2}$$

   What is your conclusion? (Both empirical or theoretical arguments are accepted.)

Problem 2 (Derivation of smoothing splines): ESL 5.7

Problem 3 (Semi-parametric Linear Model): Consider an additive model $y_i = x_i^T \beta + f(z_i) + \epsilon_i$, $i = 1, \dots, n$. Here, $x_i \in \mathbb{R}^p$ is a vector of p predictors (including the element 1 for the intercept), and z_I is an additional confounding predictor. You plan to fit this model by penalized least squares, using a smoothing spline to control the roughness of f (β will be unpenalized).

1. Write down the penalized least-squares criterion for fitting this model.
2. Show that the minimizers $\hat{\beta}$ and $\hat{f}$ satisfy the following pair of estimating equations:

$$\boldsymbol{X}\hat{\beta} = \boldsymbol{H}(\boldsymbol{y} - \hat{\boldsymbol{f}})$$
$$\hat{\boldsymbol{f}} = \boldsymbol{S}_\lambda(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

   where $\hat{\boldsymbol{f}}$ is the vector of n fitted values for the function $\boldsymbol{f}$, and $\boldsymbol{S}_\lambda$ is the appropriate smoothing spline operator matrix. What is $\boldsymbol{H}$?

3. Show that you can solve these equations explicitly for $\hat{\beta}$, and hence for $\hat{\boldsymbol{f}}$ as well.
4. Show that the fitted vector $\hat{\boldsymbol{y}}$ is linear in $\boldsymbol{y}$, i.e. $\hat{\boldsymbol{y}} = \boldsymbol{M}\boldsymbol{y}$. Give an expression for $\boldsymbol{M}$.
5. Assume the errors ϵ_i are iid with mean 0 and variance σ^2. Give the expression for the conditional covariance matrices of $\hat{\boldsymbol{f}}$ and $\hat{\beta}$.

Problem 4 (Application) In this problem, you will experiment with different methods on the vowel data set. We have encountered this data set before, and the linear methods can achieve approximated 50% of accuracy. Can we have better performance with the non-linear methods? You decide to try some methods that you have learned from BIS 555 so far and compare them (see Table 12.3 from ESL). The methods you will compare are

- LDA.
- Multiclass logistic regression.
- Multiclass logistic regression with radial kernels.
- Generalized additive model for multinomial data.

Do you observe improved prediction accuracy for the nonlinear methods? (Note that you will get full credits for this question as long as you have tried out different methods.)