Example Sheet 2: Solutions

(Several of these are modified from Hastie, Tibshirani, and Friedman)

1. How were the Bayes classification boundaries obtained for the the figure in lecture 5, page 15?

   The Bayes classifier is $\hat{G}(X) = \max_{g \in G} \Pr(g|X = x)$. To find the optimal boundary, we just need to find where

   $$\Pr(g = 0|X = x) = \Pr(g = 1|X = x) = 1/2,$$

   which is $\Leftrightarrow$

   $$\Pr(X = x|g = 0)\Pr(g = 0) = \Pr(X = x|g = 1)\Pr(g = 1)$$

   Since the generating density $\Pr(X = x|g)$ is known, prior $\Pr(g)$ is based on our prior knowledge, for example, we can set $\Pr(g = 1) = \Pr(g = 0) = 1/2$, we can calculate the boundary exactly.

2. Suppose you have done a cubic polynomial regression, fitting the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$. You would like to set a 95% confidence band on the resulting regression function.

   Compare the following two approaches:

   a At each point x_0, set a 95% confidence interval on the linear function $\sum_{i=0}^{3} \beta_j x_0^j$.

   b Set a 95% confidence region on $(\beta_0, \ldots, \beta_3)$ as

   $$\{\boldsymbol{\beta} \ : \ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq \hat{\sigma}^2 \chi^2_{4,.95}\}$$

   and use this to set the confidence band.

   Which method contains the other, and why?

- A 95% CI for $a^T\beta = \sum_{j=0}^{3} \beta_j x_0^j$ is:

$$\sum_{j=0}^{3} \hat{\beta}_j x_0^j \pm 1.96 \cdot \text{MSE}$$

, which depends on the number of observations we have for the regression. Since $\mathbf{E}(a^T\beta - a^T\hat{\beta}) = 0, \mathbf{Var}(a^T\beta - a^T\hat{\beta}) = \mathbf{MSE}(a^T\hat{\beta})$

- A 95% CI for β, as in (3.15), is $C_\beta = \{\beta | (\hat{\beta} - \beta)^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi^2_{p+1}(1-\alpha)\}$, which generates a CI for the function $f(x_0) = x_0^T\beta$, namely $\{x_0^T\beta | \beta \in C_\beta\}$. The upper and lower bounds depend not only on the predicted value but also on the specific value of x_0.

The CI from the second approach is much wider when x is far from the sample mean.

3. Show how to solve the generalized eigenvalue problem $\max \boldsymbol{a'Ba}$ subject to $\boldsymbol{a'Wa} = 1$ by transforming to a standard eigenvalue problem.

Applying the Lagrange multiplier, define

$$l(\lambda) = a^T B a - \lambda(a^T W a - 1)$$

Take derivative with respect to a, we get

$$\frac{\partial l}{\partial a} = (B + B^T)a - \lambda(W + W^T)a$$

Let it equal to 0, then we have

$$[(W + W^T)]^{-1}(B + B^T)a = \lambda a$$

, which is the standard eigenvalue problem. Assumming W and B symmetric, we have $W^{-1}Ba = \lambda a$.

4. Suppose one wants to use logistic regression to classify two groups based an observed covariate $x \in \mathbb{R}$. Describe the maximum likelihood estimates of the slope and intercept parameters if the training sample for the two groups is completely separated by a point $x_0 \in \mathbb{R}$.

Generalize this result to (a) $\boldsymbol{x} \in \mathbb{R}^p$ and (b) more than two classes.

For a two-class logistic regression with $x \in \mathbb{R}$, the loglikelihood can be written as

$$l(\beta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; p)$$

The decision boundary occurs at $\beta_0 + \beta_1 x = 0$, and the samples can be seperated by a single point x_0. If we make $\Pr(g = 1|x) = 1$ when $x < x_0$, and $\Pr(g = 1|x) = 0$, when $x > x_0$,then $l(\beta) = \sum_{i=1}^{N} \log(1)$, which achieves the maximum, while we model $\Pr(g = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$. Clearly, when $\hat{\beta}_0 = \infty$, we'll get the maximum likelihood. Similar analysis can be done when (a)$x \in \mathbb{R}^p$ and (b)more than two classes. The MLE of the intercept parameters for each class/group go to infinity.

5. Consider a nearest-neighbor regression algorithm in $\mathbb{R}^1$ that puts equal weight on the response for the second and third nearest x values, but no weight upon the closest value. Describe the "smoothing" matrix, indicate why it is unusual, check the fine print about smoothing matrices, and diagnose the problem. Also, describe the implications of this procedure for leave-one-out cross-validation.

   The smoothing matrix has zeros on the diagonal, and thus appears to have no degrees of freedom. But smoothing matrices must have all eigenvalues in the open interval (0,1), and this does not (since the trace is zero, then the eigenvalues must be zero or have some negative values). In fact, this is matrix is not at all a smoother—it is extremely unsmooth.

   For loo cross-validation, the full data estimate for a response will be the same as the loo estimate. Thus cross-validation provides no additional information about the predictive error.

6. Write code that implements boosting to improve a tree classifier. Apply this to the dataset on classifying Internet advertisements at the University of California at Irvine Machine Learning repository. Compare the results to those obtained by applying the tree classifier by itself and to the results from a Random Forests analysis.

7. Apply both Random Forests and two Support Vector Machines to the regression problem of predicting home value in the Boston Housing Data. Use both a linear kernel and the radial basis function kernel for the SVM analysis. Compare the results.

8. For 0-1 loss with $Y \in \{0, 1\}$ and $\mathbb{P}[Y = 1|x_0] = f(x_0)$, show that

$$\begin{aligned} \text{Err}(x_0) &= \mathbb{P}[Y \neq \hat{G}(x_0)|X = x_0] \\ &= \text{Err}_B(x_0) + |2f(x_0) - 1| \cdot \mathbb{P}[\hat{G}(x_0) \neq G(x_0)|x_0], \end{aligned}$$

where $\hat{G}(x) = I[\hat{f}(x) > .5]$, $G(x) = I[f(x) > .5]$ is the Bayes classifier, and $\text{Err}_B(x_0) = \mathbb{P}[Y \neq G(x_0)|X = x_0]$. The last expression is the irreducible Bayes error at x_0.

Let $A = \{X = x_0\}$. Since Err is the test error, Y and $\hat{G}(x_0)$ are independent given $X = x_0$.

$$\begin{aligned} \text{Err}(x_0) &= P[Y \neq \hat{G}(x_0)|A] \\ &= P[Y \neq \hat{G}(x_0), \hat{G}(x_0) = G(x_0)|A] + P[Y \neq \hat{G}(x_0), \hat{G}(x_0) \neq G(x_0)|A] \\ &= P[Y \neq G(x_0), \hat{G}(x_0) = G(x_0)|A] + P[Y = G(x_0), \hat{G}(x_0) \neq G(x_0)|A] \\ &= P[Y \neq G(x_0)|A] \left\{1 - P[\hat{G}(x_0) \neq G(x_0)|A]\right\} \\ &\quad + P[Y = G(x_0)|A] P[\hat{G}(x_0) \neq G(x_0)|A] \\ &= \text{Err}_B(x_0) + P[\hat{G}(x_0) \neq G(x_0)|A] \left\{P[Y = G(x_0)|A] - P[Y \neq G(x_0)|A]\right\} \\ &= \text{Err}_B(x_0) + |2f(x_0) - 1| \cdot P[\hat{G}(x_0) \neq G(x_0)|A] \end{aligned}$$

since

$$\begin{aligned} P[Y = G(x_0)|A] &= \begin{cases} P[Y = 1|A] & f(x_0) > 0.5 \\ P[Y = 0|A] & f(x_0) \leq 0.5 \end{cases} \\ &= \begin{cases} f(x_0) & f(x_0) > 0.5 \\ 1 - f(x_0) & f(x_0) \leq 0.5 \end{cases} \\ P[Y = G(x_0)|A] - P[Y \neq G(x_0)|A] &= \begin{cases} 2f(x_0) - 1 & f(x_0) > 0.5 \\ 1 - 2f(x_0) & f(x_0) \leq 0.5 \end{cases} \\ &= |2f(x_0) - 1| \end{aligned}$$

9. In the preceding problem, use the approximation

$$\hat{f}(x_0) \sim N(E[\hat{f}(x_0)], \mathrm{Var}(\hat{f}(x_0))$$

to show that

$$\mathbb{P}[\hat{G}(x_0) \neq G(x_0)|X = x_0] \approx \Phi\left(\frac{\mathrm{sign}(.5 - f(x_0))(\mathbb{E}[\hat{f}(x_0)] - .5)}{\sqrt{\mathrm{Var}(\hat{f}(x_0))}}\right).$$

Let $B = E[\hat{f}(x_0)]$ and $C = \mathrm{Var}[\hat{f}(x_0)]$. Let $\phi_{B,C}(x)$ be the normal density with mean B and variance C at point x. Rearranging the result from question 8 yields

$$\begin{aligned}
& |2f(x_0) - 1|P[\hat{G}(x_0) \neq G(x_0)|A] \\
= \; & P[Y \neq \hat{G}(x_0)|A] - P[Y \neq G(x_0)|A] \\
= \; & P[Y = 0|A]P[\hat{G}(x_0) = 1|A] + P[Y = 1|A]P[\hat{G}(x_0) = 0|A] - P[Y \neq G(x_0)|A] \\
& \text{using independence} \\
\approx \; & [1 - f(x_0)] \int_{0.5}^{\infty} \phi_{B,C}(z)dz + f(x_0) \int_{-\infty}^{0.5} \phi_{B,C}(z)dz - P[Y \neq G(x_0)|A] \\
& \text{using the given approximation} \\
= \; & \begin{cases} -[1 - f(x_0)] \int_{-\infty}^{0.5} \phi_{B,C}(z)dz + f(x_0) \int_{-\infty}^{0.5} \phi_{B,C}(z)dz & f(x_0) > 0.5 \\ [1 - f(x_0)] \int_{0.5}^{\infty} \phi_{B,C}(z)dz - f(x_0) \int_{0.5}^{\infty} \phi_{B,C}(z)dz & f(x_0) \leq 0.5 \end{cases} \\
& \text{expanding the two cases of } P[Y \neq G(x_0)|A] \\
= \; & |2f(x_0) - 1| \begin{cases} \int_{-\infty}^{0.5} \phi_{B,C}(z)dz & f(x_0) > 0.5 \\ \int_{0.5}^{\infty} \phi_{B,C}(z)dz & f(x_0) \leq 0.5 \end{cases} \\
= \; & |2f(x_0) - 1| \begin{cases} \int_{-\infty}^{(0.5-B)/\sqrt{C}} \phi_{0,1}(z)dz & f(x_0) > 0.5 \\ \int_{(0.5-B)/\sqrt{C}}^{\infty} \phi_{0,1}(z)dz & f(x_0) \leq 0.5 \end{cases} \\
= \; & |2f(x_0) - 1|\Phi\left(\frac{(0.5 - B)\mathrm{sgn}(f(x_0) - 0.5)}{\sqrt{C}}\right),
\end{aligned}$$

from which the desired result follows immediately.

10. Explain the role of shrinkage in high-dimensional nonparametric regression and classification. Give examples from methods we have discussed.