Stats 202 Practice Problems

August 7, 2015

1. We fit a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ to some data. Suppose we change the units of the predictors X_i, to obtain a new set of predictors $Z_i = cX_i$. Then, we fit the same data to the model: $Y = \alpha_0 + \alpha_1 Z_1 + \ldots + \alpha_p Z_p$.

   (a) What is the relationship between the least squares coefficients $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_p)$ and $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$? Provide a proof.

   (b) What is the relationship between the fitted values in the two models?

2. Your colleague fitted a multivariate linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ and found all but three p-values are significant in the t-test. He decides to drop those three variables and keep all the remaining predictors. What do you think of your colleague's method?

3. Explain the purpose of an F-test for multiple linear regression.

4. True or false: The variance of a regression estimator $\hat{f}$ in the bias-variance decomposition can be written:

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{f}(x_i) - m)^2,$$

   where

$$m = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_i),$$

   and $x_1, \ldots, x_n$ are the inputs of the training data.

5. Suppose the data $(\mathbf{X}, \mathbf{y})$ are well fit by a linear model, how would you diagnose if the data point (x_i, y_i) is an outlier or a high leverage point?

6. Suppose we have a dataset with N observations, and each observation consists of three values:

   - y: binary variable that is 1 if a student passed and 0 if a student failed the exam
   - x_1: the number of hours spent studying for the exam
   - x_2: a binary variable indicating whether or not the student passed the previous exam.

   Suppose upon fitting a logistic regression of the y on x_1, x_2, and an intercept, the estimates for $\beta = (\beta_0, \beta_1, \beta_2)$ are

$$\begin{aligned}\hat{\beta}_0 &= -1.2\\ \hat{\beta}_1 &= 0.3\\ \hat{\beta}_2 &= 1.2\end{aligned}$$

Now suppose instead of using the number of hours spent studying, we used the number of minutes spent using for the exam. Can you identity what the new β_0, β_1, and β_2 would be? Why or why not? Justify your claim.

7. Suppose we have a classification problem with a binary response Y and a p-dimensional predictor variable $X = (X_1, \ldots, X_p)$. Logistic regression is fitted to a set of n samples. Then, logistic regression is fitted again to the same observations, where we include one additional predictor, such that:

$$X = (X_1, \ldots, X_p, X_{p+1}).$$

Explain how the training error, test error, and coefficients change in each of the following cases:

   (a) $X_{p+1} = X_1 + 2X_p$.

   (b) X_{p+1} is a random variable independent of Y.

8. What are the key differences between LOOCV and k-fold cross validation for estimating the true test error? Is one better than the other, and if so, why?

9. Consider the dataset

x	y
-2	'slow'
5	'fast'
-1	'slow'
10	'fast'
4	'fast'

Suppose we used logistic regression to fit this model: that is, if y is a binary variable that is either 'fast' or 'slow', we wish to fit the model

$$\mathbb{P}(y_i = \text{fast}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \qquad \mathbb{P}(y_i = \text{slow}) = \frac{e^{-(\beta_0 + \beta_1 x_i)}}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

for all $i = 1, \ldots, 5$. What value(s) of β would maximize the likelihood (and thus be the estimates returned from fitting this model)?

10. Suppose that $\mathbf{X}$ is an $n \times p$ matrix of predictors and $\mathbf{y}$ is a quantitative response. Suppose that $p > n \geq 1000$. You want to fit a linear model that helps you make predictions with new data. Explain which of the following methods could be applied, and the advantages of each one.

    (a) Least squares

    (b) Lasso

    (c) Ridge regression

    (d) Backward stepwise selection

    (e) Forward stepwise selection

(f) Best subset selection

11. Consider selecting subsets of predictors in the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

We run best subset selection, forward stepwise selection, and backward stepwise selection using RSS as the criterion and get the list of models:

$$\text{Best subset selection} : \mathcal{M}_0^{(bs)}, \mathcal{M}_1^{(bs)}, \cdots \mathcal{M}_p^{(bs)}$$
$$\text{Forward stepwise selection} : \mathcal{M}_0^{(forward)}, \mathcal{M}_1^{(forward)}, \cdots, \mathcal{M}_p^{(forward)}$$
$$\text{Backward stepwise selection} : \mathcal{M}_p^{(backward)}, \mathcal{M}_{p-1}^{(backward)}, \cdots, \mathcal{M}_0^{(backward)}$$

For example, $\mathcal{M}_k^{(bs)}$ is the 'best' model among all the models with k predictors and $\mathcal{M}_k^{(forward)}$ is the 'best' model among the $p-k$ candidate models after choosing $\mathcal{M}_{k-1}^{(forward)}$.
Let $RSS_{\mathcal{M}}$ the training error fitted by model $\mathcal{M}$. Show that

(a) $RSS_{\mathcal{M}_p^{(forward)}} = RSS_{\mathcal{M}_p^{(backward)}}$.
(b) $RSS_{\mathcal{M}_1^{(forward)}} \leq RSS_{\mathcal{M}_1^{(backward)}}$.

12. Imagine starting to grow a 2-class classification tree. We have 80 points total, with 3 possible values of x. In class 1, 30 of the points are at $x = 0$, 10 points are at $x = 1$, and no points are at $x = 2$. In class 2, 10 of the points are at $x = 0$, 10 points are at $x = 1$, and 20 points are at $x = 2$. There are two potential splits on x: split between $x = 0$ and $x = 1$, or split between $x = 1$ and $x = 2$.

Compute the the misclassification error and the Gini index for the two splits. Which criterion produces a pure region?

13. We build a classification tree using the predictors $X_1, X_2, \ldots, X_p$. We build a second tree using the predictors $f(X_1), f(X_2), \ldots, f(X_p)$, where f is monotone:

$$f(x) > f(y) \text{ if and only if } x > y.$$

Prove that the two trees produce the same partition of the training data.

14. The standard method for fitting a decision tree involves:

- Growing the tree split by split. We maximize the reduction of the training error at each step until there are at most 5 samples per region.
- Pruning the tree to obtain a sequence of trees of decreasing size.
- Selecting the optimal size by cross-validation.

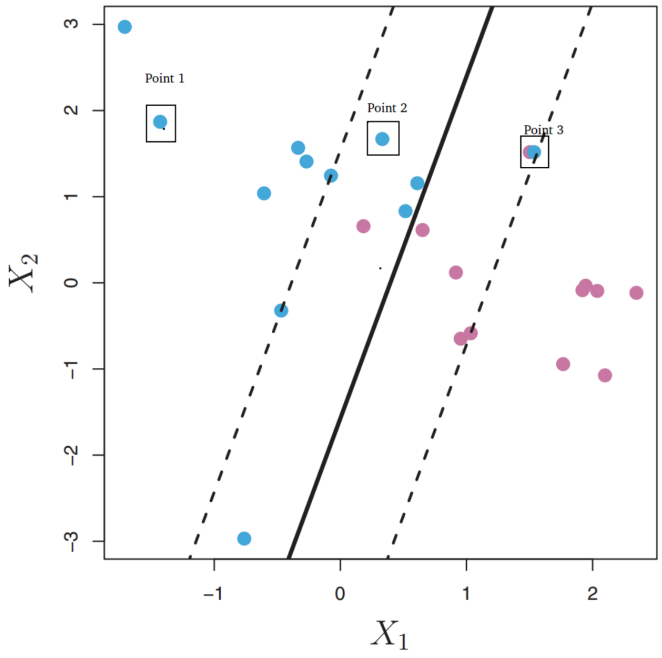
Consider the following alternative approach. Grow the tree split by split until the reduction in the training error produced by the next split is smaller than some threshold. This approach may lead to bad results because it is possible to make a split which does not decrease the error by much, and then make a second split which reduces the error significantly.

Draw an example dataset where this happens with 2 predictors X_1 and X_2, and a binary categorical response.

15. When we apply Bagging to decision trees, or when we construct Random Forests, each tree fit to a different bootstrap replicate of the data can be grown "deep", i.e. to a level where there are very few training samples per leaf. The pruning step is skipped. Explain why this doesn't lead to overfitting.

16. The plot below, from ISLR, shows the separating hyperplane (solid line) and margin (dotted lines) resulting from fitting the SVM optimization

$$\text{maximize } M \tag{1}$$
$$\text{over } \beta, \epsilon \text{ such that,} \tag{2}$$
$$\beta_0 + \beta_1^2 + \beta_2^2 = 1 \tag{3}$$
$$y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right) \geq M \left(1 - \epsilon_i\right) \tag{4}$$
$$\epsilon_i \geq 0 \tag{5}$$
$$\sum_{i=1}^{n} \epsilon_i \leq C. \tag{6}$$



This problem is about the interpretation of these parameters.

(a) Suppose we decrease C. What happens to the separating hyperplane? What happens to the margin?

(b) For each of the boxed blue points, specify whether $\epsilon_i = 0, \epsilon_i \in (0, 1]$, or $\epsilon_i > 1$.