Statistics 315a

Midterm Exam

4:15-5:30pm, February 29, 2012.

You may use the class text, notes and calculators, but not computers or any device that connects to the Internet. The questions below require fairly short answers and are of equal value. They are in no particular order. No one is expected to answer all of the questions. but everyone is encouraged to try them. All answers should be written in the space provided.

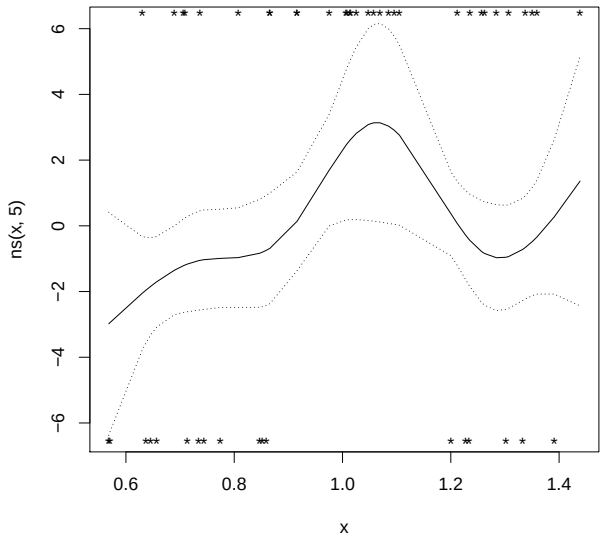
1. Your biology collaborator has the gene-expression values for 5000 genes on 150 biological samples, half of which exhibit stunted growth. He suspects a handful of genes are responsible, and asks you to build a classifier to help him uncover the mechanism. Which of the following method(s) are likely to be the most useful: (a) ridge regression (b) all subsets regression (c) nearest-shrunken centroids (d) nearest neighbor classification (e) elastic-net penalized logistic regression? Give reasons.

   (c) and (e) will be most useful. In (c), we explicitly focus on a small subset of the genes, and in (e) the elastic net (like the lasso) does variable selection. This way we dont get hurt by the curse with all the redundant genes. (a) would suffer from bias, since massive shrinkage would be needed to dull the effect of all the spurious genes. (b) would be reasonable except p is too large for best subset. (d) uses all the variables, and so would create massively wide neighborhoods.

2. A software package does supervised learning, and produces a prediction function $\hat{f}(x_1, x_2, \ldots, x_p)$. It reports the importance of predictor j as $\frac{1}{N}\sum_{i=1}^{N} |\frac{\partial \hat{f}}{\partial x_j}(\mathbf{x}_i)|$. What is wrong with this measure. Draw analogies with linear regression. Can you think of a better way to measure the importance of each predictor?

   This measure is not scale invariant, for starters. In linear regression it would be the coefficient of X_j. It also doesnt account for correlations. If two variables are very correlated, their coefficients can be wild and compensatory. A better approach would be to refit the model with the variable omitted, and use the change in the training error as a measure of importance.

3. You have some binary response and a single X variable, so you fit a logistic regression using a natural cubic spline with 4 interior knots to model the logit. Below is the fitted logit with pointwise standard errors. Included in the plot are *s indicating the locations of the zeros and ones (The upper points represent ones, the lower points zeros). You expect to see the standard error bands widening at the boundary of the data, but they appear to do so in the interior around $x = 1.1$ as well. Suggest why this might be so. Would you expect the same phenomenon for a linear fit?



The X values are more or less uniformly distributed. However, there is a pure region of 1s near the middle of the range of X. With localized basis functions like splines, there is the ability of fitting these local pure regions exactly — i.e. with probabilities approaching 1 in that region. This requires certain coefficients to grow very large in size, and this accounts for the wide standard error bands (the weights in IRLS get really small, so the inverse Hessian in that region explodes). A linear logistic fit would be stable, because no point separates the two classes.

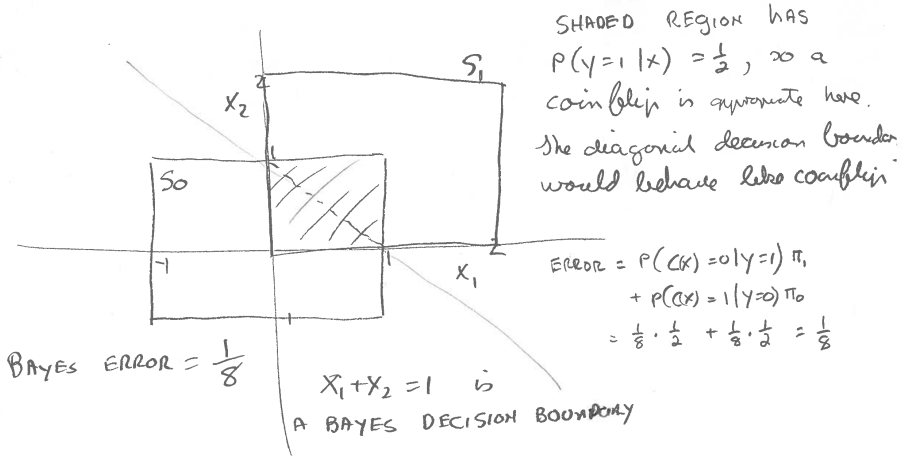
4. Your are fitting a regression model with 100 observations and 500 features. The data have been carefully curated and are complete. However, you are told that when the model is to be used to make predictions, you can expect about 25% of the values to be missing at random at each evaluation point (and not the same 25% each time). Suggest at least two methods that might be appropriate for this problem, and explain why. How will you deal with the missing values?

   Ridge regression or principal component regression would be suitable, while lasso would not. Lasso might use a subset of variables most of which could be missing at certain evaluations. We would replace missing variables by their mean in the training data when making predictions.

5. Suppose that $Pr(Y = 1) = Pr(Y = 0) = 1/2$, and

$$\begin{aligned} X|Y = 0 &\sim \text{Uniform on } S_0 \\ X|Y = 1 &\sim \text{Uniform on } S_1. \end{aligned}$$

   Here S_0 is the square region in R^2 with corners $(1, 1), (1, -1), (-1, -1), (-1, 1)$, while S_1 is is the square with corners $(0, 0), (2, 0), (2, 2), (0, 2)$. Find an expression for the Bayes classifier for Y. What is the Bayes risk?



6. A scientist fits a classifier to some high-dimensional microarray data with 2 classes. Since there are so many genes in the dataset, he first filters the genes, keeping only the 1000 genes having largest variance across the samples. With these he builds his classifier. To estimate the error rate of his classifier he applies cross-validation using these 1000 genes. The estimated CV error rate is 2% and he is very happy. Your are asked to be the statistics expert on the paper. Any objections to his approach, or suggestions?

   This is a bit of a trick question. There is no objection. Variance is unsepervised, so it is an allowable screen. It may be that the variance is due to class differences, which is the hedge you are hoping for. But of course you may miss some good genes too.

7. You fit a linear regression model to measure the effect on PSA of a number of lifestyle variables, such as exercise level, smoking, diet, alcohol intake and more — 11 in all. Your software prints out the summary statistics, and 4 of the 11 are significant at the 1% level, and the rest are not at all significant. So you drop the 7 weak variables, refit your model, and report 95% confidence intervals for each of the four coefficients. Is there anything wrong with this approach. Give details, and suggest a better one.

   There are several problems with this approach. Firstly, the "weak" variables might not all be weak; there could be correlated pairs where only one is needed, but drop both and the model suffers. Best to drop the variables one at a time. Secondly, since we have used a two-stage procedure, the confidence intervals would not have the right coverage. If we can automate our procedure, we could use the bootstrap to get an idea of the sampling distribution of the coefficients, including the proportion of times they are omitted.

8. You fit an optimal separating hyperplane using your sample of 150 0/1 observations and 3 predictors. Your misclassification rate estimated by cross-validation is 0.053. Just before you finish your paper, your collaborator provides you with 17 more observations. So you refit the model, and redo the cross-validation. To your astonishment, the fitted line has not changed at all, while the CV error rate is now 0.048. Is this possible? Offer some explanations.

   It is possible. The separating hyperplane might have as few as 4 support points, and your new samples might not occur in the margin region. If this is the case, the solution would remain the same, since the new points are not support points. The CV error could decrease, if the new points were easy to classify.

9. We want to fit a ridged linear regression model to our data and one of the predictors is a 5-level factor. We decide to represent its levels by five dummy variables (0/1) $Z_1, \ldots, Z_5$ with associated coefficients $\alpha_1, \ldots, \alpha_5$. They, along with all the other coefficients in the model get penalized by the same ridge penalty as in

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{5} z_{ij}\alpha_j - \sum_{\ell=1}^{L} x_\ell\beta_\ell)^2 + \lambda\sum_{j=1}^{5}\alpha_j^2 + \lambda\sum_{\ell=1}^{L}\beta_\ell^2.$$

   Notice that α_0 is not penalized. Show that the solutions for α_j satisfy $\sum_{j=1}^{5}\hat{\alpha}_j = 0$. What would be the case if we used a lasso rather than ridge penalty?

   Suppose α_0^* and $\alpha_j^*,\ j = 1, \ldots, 5$ are candidate solutions. Then as far as the loss is concerned, $\alpha_0^* + c,\ \alpha_j^* - c,\ j = 1, \ldots, 5$ will be just as good, because $\sum_{j=1}^{5} z_{ij} = 1\ \forall j$. Hence the optimal solution for c will be the one such that $\sum_{j=1}^{5}(\alpha_j^* - c)^2$ is minimized. This will be $\hat{c} = \frac{1}{5}\sum_{j=1}^{5}\alpha_j^*$. Hence the optimal $\hat{\alpha}_j$ will average 0. If we use the lasso instead, c will minimize $\sum_{j=1}^{5}|\alpha_j^* - c|$; this is the median, and hence the median of the $\hat{\alpha}_j$ will be zero.

10. You have 5000 gene-expression features and 270 samples for building a linear classifier in order to predict the five-year outcome after breast surgery. You randomly extract a test dataset of 100 samples, and train on the remaining 170. You run a lasso regression procedure with complexity parameter determined by cross-validation. You are pleased to see a test error of only 10% on your separate test dataset. You submit to a journal, and provide the data as well. A referee report claims that your analysis is wrong, because they ran your analysis and got a CV error rate of 17%. What has gone wrong? Is this a reasonable discrepancy? What sources of error are involved? What could you have done instead to avoid this situation.

    The referee probably got a different division into training and test, and a different set of random folds. The sources of error are the different training set/ test set divisions leading to a different trained model, different tuning parameter selection due to different random folds, and variations in test populations leading to different errors. Better to provide a random-number seed with your code, to make it completely reproducible. Since these sources of variance seem quite large here, you might run your whole procedure from beginning to end many times with different seeds, and report the variation that you observe.