STA521 **2019**

Midterm II

```
> knitr::opts_chunk$set(echo = TRUE)
> library(ISLR)
```

NOTES:

1. You may use two sheets of notes and a calculator. If you do not have a calculator work out expressions as far as you can.
2. Distributions are at the end of the exam.
3. Most parts do not depend on earlier parts of the problem, so if you are stuck please move on to the next section. Partial credit will be given if an answer does depend on an earlier part that was incorrect and the later problem was worked correctly after accounting for the previous error.
4. You do not have to re-derive well known results unless asked to, but if you do use specific results from class or Theorems please state them where used in your explanations.
5. The amount of space is not always an indication of the expected length of an answer. In general brief answers to all questions are better than more detailed responses to half the problems, so please use your time wisely.
6. Partial credit will be given where appropriate, although no points will be given for simply restating the problem.
7. Please erase any work that you do not want to be graded, so that there is one solution. DO NOT USE THE BACK of a page for any work you wish to be graded.
8. In signing your name below, you agree to abide by the Duke Community Standard and will not receive or give help from/to anyone else. If you notice violations of the Duke Community Standard, you should report this.

PRINT NAME (first last):_____

Print NETID:_____

Signature:_____

1. **Circle True or False** (no credit for writing T or F elsewhere near the problem!)

   (a) True or False: Best subset selection cannot be performed if p is too large.

   (b) True or False: We can use Box-Cox to find the best transformation of $\mathbf{Y}$ for Poisson or Negative Binomial regression.

   (c) True or False: The model which has the smallest RMSE in the test data will also have the best coverage probability for confidence or credible sets.

   (d) True or False: The MSE from a test data set provides an appropriate estimate of the irreducible error.

   (e) True or False: lasso more or less shrinks all coefficients to zero by the same amount, and sufficiently small coefficients are shrunken all the way to zero.

   (f) True or False: Using many trees in boosting avoids over-fitting as the depth is typically limited.

   (g) True or False: Using the lasso to select a model and then using OLS with the selected variables provides unbiased estimates of the β's.

   (h) True or False: Bayesian Model Averaging cannot be implemented if p is too large to enumerate all of the models.

   (i) True or False: With enough data the highest posterior probabilty will select the model that is "closest" to the true model, even if the true model is not one of the models in BMA.

   (j) True or False: A limitation of Generalized Additive Models (GAMs) is that they are additive in the variables and hence cannot capture interactions.

(k) True or False: Using AIC and backwards selection, the predictors in the best k variable model are a subset of the predictors in the best $k+1$ variable model.

(l) True or False When using Zellner's g-prior one should always center and rescale the predictors so that prior variances of the β's are equal to avoid overshrinking when the predictors have different units of measurement.

(m) True or False: In the Bayesian lasso, the posterior mean of $\boldsymbol{\beta}$ can be exactly zero, leading to variable selection.

(n) True or False: Random Forests tends to out-perform Bagging as the selection of additional trees that go into estimate of the regression function depends heavily on the previous trees.

(o) True or False: Ridge regression, lasso, and BMA all attempt to prevent over-fitting while using all p predictors by shrinking coefficients to zero and significantly reducing their variance.

(p) True or False: Using a test set for validation or cross validation provides a direct estimate of the test error for selecting a model.

(q) Using AIC and best subset selection, the predictors identified in the best k-variable model are a subset of the predictors in the best $k+1$ variable model.

(r) True or False: Averaging single tree models tends to reduce the expected bias over single tree models.

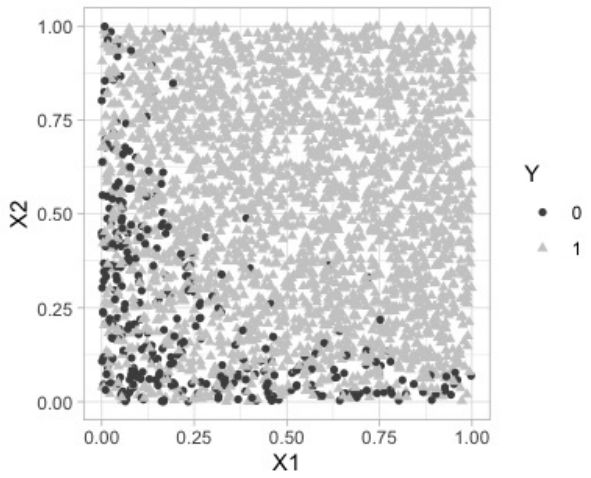
(s) True or False: An advantage of GAMs is that they can model nonlinear relationships that we might miss with PowerTransform or Box-Tidwell in OLS.

(t) True or False: A disadvantage of GAMs is that they cannot easily handle qualitative variables and numeric variables together in the same model.

(u) True or False: Using a large depth for trees in Bagging or Random Forests will lead to over-fitting.

(v) True or False: lasso will always have a smaller out of sample prediction error compared to ridge regression because it is able to shrink some coefficients all the way to zero.

2. Consider the following plot of two predictors, $X1$ and $X2$, where the color/shape of the points is designated by the response Y for 3000 observations where the log odds that Y equals one is given by $25 * X1 * X2$.



Which method would you recommend as providing the best estimator for the mean function for Y for this data? (circle only one)

(a) lasso

(b) boosting

(c) BART

(d) Random Forests

(e) Bagging

(f) BMA with binomial (Bernoulli) family

(g) Support Vector Machines

Explain:

3. Circle True or False: Non-linear methods such as boosting, GAMs and SVM relative to OLS, are

   (a) True or False: more flexible and hence will give improved prediction accuracy when their increase in bias is less than its decrease in variance.

   (b) True or False: more flexible and hence will give improved accuracy when their increase in variance is less than its decrease in bias.

   (c) True or False: less flexible and hence will give improved prediction accuracy when their increase in bias is less than its decrease in variance.

   (d) True or False: less flexible and hence will give improved prediction accuracy when their increase in variance is less than its decrease in bias.

4. Suppose we use the following hierarchical model for the data

$$Y_i \mid \sigma^2, \boldsymbol{\beta} \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \sum_j^p x_{ij}\beta_j, \sigma^2) \tag{1}$$

$$\beta_j \mid \sigma^2, \lambda \overset{\text{iid}}{\sim} \mathsf{DE}(0, \lambda/(2\sigma^2)) \tag{2}$$

$$p(\beta_0, \sigma^2) \propto 1/\sigma^2 \tag{3}$$

where you may assume that the mean of $\mathbf{x}_j$ is zero and that they have been scaled so that $\sum x_{ij}^2 = 1$. The predictors, however are not uncorrelated.

(a) Circle the best response. For a fixed value of λ, estimating the β_j's using the posterior mode is equivalent to

i. the posterior mean with Zellner's g-prior with $g = 1/\lambda$
ii. a penalized likelihood estimate using the Lasso
iii. a penalized likelihood estimate using Ridge Regression
iv. the posterior mode using the Jeffreys- Zellner Siow Cauchy Prior
v. using BIC to select a model if $\lambda = n$
vi. None of the above

(b) This model is equivalent to using a penalized loss

i. $\sum_{i=1}^n (Y_i - \beta_0 - \sum_j^p x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^p |\beta_j|$
ii. $\sum_{i=1}^n (Y_i - \beta_0 - \sum_j^p x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^p |\beta_j|^2$
iii. $\sum_{i=1}^n (Y_i - \beta_0 - \sum_j^p x_{ij}\beta_j)^2 + p\log(n)$
iv. None of the above

(c) Circle the best response. As λ goes to zero, the estimates of the β_j's will be

i. equivalent to using OLS
ii. all go to zero

(d) The largest value of the MSE $\equiv$ SSE/n for the training data corresponds to which value of λ?

(e) The smallest value of the MSE $\equiv$ SSE/n for the training data corresponds to which value of λ?

(f) As we increase λ from 0 to ∞ we expect the irreducible error in the test data to

i. Increase initially and then eventually start decreasing in an inverted U shape.
ii. Decrease initially and then eventually start increasing in an inverted U shape.
iii. Steadily increase
iv. Steadily decrease
v. Remain constant
vi. Non of the above

(g) As we increase λ from 0 to ∞ we expect the mean squared error in the test data to

i. Increase initially and then eventually start decreasing in an inverted U shape.
ii. Decrease initially and then eventually start increasing in a U shape.
iii. Steadily increase
iv. Steadily decrease
v. Remain constant
vi. Non of the above

Useful Distributions

Multivariate Normal

$$\mathbf{Y} \mid \boldsymbol{\mu}, \mathbf{V} \sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{V})$$
$$p(\mathbf{Y}) = (2\pi)^{-n/2}|V|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(Y-\boldsymbol{\mu})\} \quad \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^n, \mathbf{V} > 0$$

Poisson

$$Y \mid \lambda \sim \text{Poi}(\lambda)$$
$$p(y) = \frac{y^\lambda e^{-\lambda}}{y!}$$

Bernoulli

$$Y \mid \lambda \sim \mathsf{Ber}(\pi)$$
$$p(y) = \pi^y(1-\pi)^{1-y}$$

Gamma

$$Y \mid a, b \sim \mathsf{Gamma}(a, b)$$
$$p(y) = \frac{b^a}{\Gamma(a)}y^{a-1}e^{-by} \text{ for } a > 0, b > 0, y > 0$$
$$\mathsf{E}[Y] = a/b \qquad \mathsf{Var}[Y] = a/b^2$$

Double Exponential

$$Y \mid \lambda \sim \mathsf{DE}(\mu, \lambda)$$
$$p(y) = \frac{\lambda}{2}\exp\left(-\lambda|Y-\mu|\right) \qquad y \in \mathbb{R}, \lambda > 0, \mu \in \mathbb{R}$$

Student-t

$$Y \mid \mu, \sigma^2 \sim \mathsf{St}(\nu, \mu, \sigma^2)$$
$$p(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\sigma^2\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} \qquad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0$$

Multivariate Student-t

$$\mathbf{Y} \mid \boldsymbol{\mu}, \mathcal{S} \sim \mathsf{St}(\nu, \boldsymbol{\mu}, \mathcal{S})$$
$$p(\mathbf{Y}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu\pi)^{p/2}|\mathcal{S}|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{1}{\nu}(\mathbf{Y}-\boldsymbol{\mu})^T\mathcal{S}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right)^{-(\nu+p)/2} \qquad \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^p, \mathcal{S} > 0$$