**STA521** **2017**

# 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 cross out any work that you do not want to be graded, so that there is one solution. If you need additional space, you may use the backs of pages, but please indicate that you have done so.
8. In signing your name below, you agree to abide by the Duke Community Standare and will not receive or give help from/to anyone else. If you notice violations of the Duke Community Standard, you should report this.

NAME:__________________________

SCORE:__________________________

1. Consider the model

$$Y_i \mid \beta_0, \beta_1, \ldots \beta_p, \sigma^2 \overset{\text{ind}}{\sim} \mathsf{N}\left(\beta_0 + \sum_{j=1}^{p} \frac{(x_{ij} - \bar{x}_j)}{\mathsf{SSX}_j}\beta_j, \sigma^2\right) \qquad \mathsf{SSX}_j = \sum_i (x_{ij} - \bar{x}_j)^2$$

$$\beta_j \mid \lambda, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{DE}(0, \lambda/\sigma)$$
$$p(\beta_0, \sigma) \propto 1/\sigma$$

(a) Which of the following estimators of $\beta_j$ under the above model may lead to estimates of coefficients being exactly zero?

i. posterion mean
ii. posterior mode
iii. posterior median
iv. all of (i), (ii), and (iii)
v. none of the above as Bayes estimators (i), (ii), and (iii) under this prior are never exactly zero.

(b) If $\alpha_j = \beta_j/\mathsf{SSX}_j$, find the density of $\alpha_j \mid \lambda, \sigma$.

(c) Explain (briefly) why variables in lasso, ridge regression and the Bayesian analogs are usually scaled so that $\sum_i (x_{ij} - \bar{x}_j)^2 = 1$.

2. Suppose we estimate regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\frac{(x_{ij}-\bar{x}_j}{\mathsf{SSX}_j})\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

(a) This is equivalent to using the posterior mode under which prior distribution on the $\beta_j$? (specify name and all parameters)

(b) The smallest value of the MSE for the training data corresponds to which value of $\lambda$?

(c) Choosing $\lambda$ to minimize the posterior variance of the coefficients leads to estimates that have optimal credible intervals. True or False? (circle one)

(d) (Justify your selection) As we increase $\lambda$ from 0 to $\infty$ we expect the squared bias 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

(e) (Justify your selection) As we increase $\lambda$ from 0 to $\infty$ 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

(f) (Justify your selection) As we increase $\lambda$ from 0 to $\infty$ we expect the residual sum of squares in the training 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

3. Circle True or False (no credit for writing T or F next to answer; if you are unsure add justification)

True or False Bagging with trees is a special case of Random Forests

True or False Boosting is a special case of Random Forests where all the trees that are aggregated have a depth of $d$.

True or False Using the lasso to select the non-zero coefficients and then using OLS with the selected variables leads to unbiased estimates of the $\beta$'s with smaller variances.

True or False Bayes factors are independent of any prior hyperparameters and can be used to express the evidence in the data in favor of a model.

True or False Random forests are parsimonious even when the true model is linear in the predictors.

True or False By using more flexible models like lasso, ridge, (and Bayesian analogs), trees, random forests, boosting, bart, etc we can eliminate the irreducible error for predicting $Y$ that is present with OLS.

True or False By using more flexible models like lasso, ridge, (and Bayesian analogs), trees, random forests, boosting, bart, etc we can reduce bias for predicting $Y$.

True or False If $\beta_j \mid \kappa, \sigma^2$ are independent and identically distributed as $\mathsf{N}(0, \sigma^2/\kappa)$ for $j = 1, \ldots, p$ and $\kappa \mid \sigma^2 \sim \mathsf{Gamma}(1/2, 1/2)$, then $\beta_j \mid \sigma^2$ have independent Cauchy distributions.

True or False If a marginal posterior inclusion probability is less than 0.5 then we can conclude that the corresponding variable is likely independent of the response.

True or False Bayesian Model Averaging cannot be used if we cannot enumerate all models, i.e. if the number of predictors is greater than 25-35.

True or False Flexible models like random forests, boosting, and bart are good for reducing the error in estimating $f(x)$ and avoid overfitting.

True or False Ridge regression results in unbiased estimates with smaller variance than OLS

True or False In the Bayesian lasso, the posterior maximum a priori (MAP) estimate of$_neverleadstovariableselection,$

True or False Unlike Bayesian estimates that are typically biased, the lasso leads to estimators that are unbiased.

4. Explain (briefly) why bagging is expected to have a smaller variance or mean squared error for estimating the unknown mean function $f(x)$ than using a single tree model. Why is random forests expected to improve over bagging?

5. Consider the hierarchical model

$$Y \mid \lambda \sim \text{Poi}(\lambda) \tag{1}$$

$$\lambda \mid \mu, \theta \sim \mathsf{Gamma}(\theta, \theta/\mu) \tag{2}$$

$$\tag{3}$$

Two students were discussing how to choose a prior distribution on $\theta$. One argued that since he had little knowledge about the problem that using the "non-informative" prior $\theta \sim \mathsf{Gamma}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon} = 0.1$, would be sensible for the problem.

(a) The $\mathsf{Gamma}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$ corresponds to the plot in

i.

ii.

iii.

iv. none of the plots shown

(b) Based on similar data, the other student thought that the variance of $Y$ would be at most about twice the variance of the Poisson model. Would any of the above distributions do a good job at representing that belief? Explain. How would you go about providing a prior distribution for this problem?

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

$$\begin{aligned} Y_i \mid \lambda_i, \sigma^2, \underline{,}\gamma_1, \dots \gamma_p &\overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \sum_j^p x_{ij}\beta_j, \sigma^2/\lambda_i) \\ \lambda_i \mid \underline{,}\sigma^2, \gamma_1, \dots \gamma_p &\overset{\text{iid}}{\sim} \mathsf{Gamma}(\alpha/2, \alpha/2) \\ \beta_j \mid \beta_0, \sigma^2, \kappa_j, \gamma_j &\overset{\text{ind}}{\sim} \mathsf{N}(0, \sigma^2\gamma_j/\kappa_j) \\ \kappa_j \mid \beta_0, \sigma^2\gamma_j &\overset{\text{iid}}{\sim} \mathsf{Gamma}(1/2, 1/2) \\ \gamma_j \mid \pi, \beta_0, \sigma &\overset{\text{iid}}{\sim} \mathsf{Ber}(\pi) \\ p(\beta_0, \sigma) &\propto 1/\sigma \end{aligned}$$

(a) Derive the distribution of $Y_i$ given $\beta_0$, $\beta_1, \dots \beta_p$, $\gamma_1, \dots \gamma_p$, $\sigma^2$ and $\alpha$. (provide the name and all hyperparameters)

(b) Derive the distribution $\lambda_i$ given $Y_i$, $\beta_0$, $\beta_1, \dots \beta_p$, $\sigma^2$, $\gamma_1, \dots \gamma_p$ and $\alpha$. (provide the name and all hyperparameters)

Useful Distributions

Multivariate Normal

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

Poisson

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

Bernoulli

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

Gamma

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

Double Exponential

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

Student-t

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

Multivariate Student-t

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