STAT 432, Spring 2021 **Name (Print):** _____

Midterm II, 05/04/2021 **Net ID (Print):** _____

Time Limit: 9:30AM - 11AM

This exam contains 6 pages (including this cover page) and 5 problems. Please read the following descriptions and requirements carefully.

- You do not need to submit a hard copy of this exam. Instead, write all of your answers (clearly labeled) on a single file (MS word or txt) and submit it to Compass2g. Make sure to also write your name and NetID in the file. Your file should look like the following:

  Name: Ruoqing Zhu
  NetID: rqzhu
  Q1: ABC
  Q2: C
  Q3: BD
  ...

- There are 20 questions. Each question worth 5 points. All questions may have multiple correct answers or even no correct answers. For each wrongly selected item (correct but not selected, or incorrect but selected), you lose one point. For example, if the correct answer is AD, and your answer is AC, then you will lose two points, for not selecting D and wrongly selecting C.

- This is an open-book exam and you can use your class notes, homework, calculator, PC, etc. or even google search.

- **You are NOT allowed to discuss** the content of this exam to anyone else (except the instructor) until the end of May 4th. This includes posting any related questions on online discussion forums or social media during and after the exam. A violation of this policy will lead to an **immediate F** as your final score of this course!

Section	Points	Score
1	20	
2	20	
3	25	
4	25	
5	10	
Total:	100	

1. Classifications

   (i) (5 points) Which of the following is true regarding linear and quadratic discriminate analysis

   A. They both assume normal distributions
   B. They are based on the Bayes theorem
   C. LDA assumes all covaraince matrix to be the same
   D. LDA give a linear decision rule, which is the same as Logistic regression

   (ii) (5 points) Which of the following is true regrading a logistic regression

   A. If we need to make a hard classification rule based on the logistic regression, it would be based on a linear function of X
   B. Based on a logistic regression, if $\beta_1 = 0.1$ then for each unit increase of X_1, the probability of Y being 1 will increase by 0.1.
   C. Logistic regression can be solved using gradient descent
   D. Logistic regression can be solved using coordinate descent

   (iii) (5 points) ROC and AUC

   A. ROC curve is a more sensible measure when one of the class labels dominates
   B. ROC curve is a more sensible measure when class proportions are similar
   C. AUC is a summary of the ROC curve
   D. ROC curve cannot be applied to a hard classification rule

   (iv) (5 points) Which of the following choices will affect the bias-variance trade-off (or the complexity) of the model

   A. Choosing LDA vs QDA
   B. Add a Lasso penalty on logistic regression
   C. Add a Ridge penalty on logistic regression
   D. Add a diagonal matrix to the covariance matrix in QDA

2. Splines

(i) (5 points) The following R code is an example of

```
lmfit <- lm(Y ~ splines::bs(X, degree = 2, knots = c(1,2,3)))
```

A. Linear spline
B. Piecewise linear
C. Quadratic spline
D. Piecewise quadratic

(ii) (5 points) A researcher is fitting a univariate regression using cubic spline model with 3 knots. How many degrees of freedom this model has?

A. 4
B. 5
C. 6
D. 7

(iii) (5 points) The main advantage of natural cubic spline compared with cubic spline is

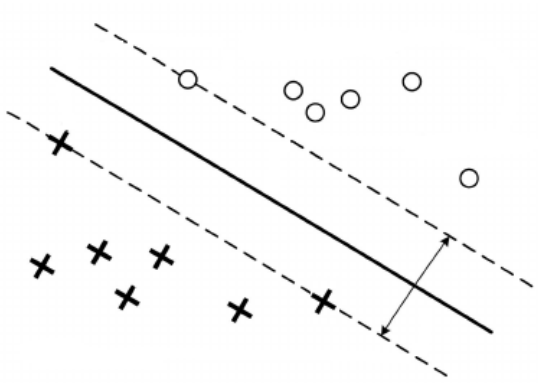
A. More flexible
B. Easier to compute
C. Less overfitting at the boundary
D. Less overfitting at the interior

(iv) (5 points) The main steps of fitting a multi-variate spline model involves

A. Creating the spline basis and fitting a kernel regression
B. Creating the spline basis and fitting a linear regression
C. Creating the neighboring weights and fitting a weighted linear regression
D. Creating the neighboring weights and fitting a weighted kernel regression

3. Support Vector Machines.

(i) (5 points) The following figure is an illustration of a fitted SVM. Which setting best describes this model?



A. linearly separable SVM
B. linear SVM with slack variables
C. Nonlinear separable SVM
D. Nonlinear SVM with slack variables

(ii) (5 points) How many support vectors do we have in the solution?

A. 2
B. 3
C. 4
D. 5

(iii) (5 points) When calculating β_0 of this solution, we should use

A. All the samples
B. The support vectors
C. All samples except the support vectors
D. All support vectors from just the positive (or negative) side

(iv) (5 points) Which of the following statements is true regarding a nonlinear SVM problem? All notations follow our lecture notes.

A. The calculation of inner product $\langle \Phi(X_i), \Phi(X_j) \rangle$ is faster if we use the kernel trick
B. We often choose a basis expansion $\Phi(X_i)$ first, then pick the corresponding kernel function $K(\cdot)$ to match that
C. If the true decision rule is $X_1^2 + X_2^2 < 1$, then a polynomial kernel can fit the data well
D. Radial basis function kernel corresponds to nonlinear decision rules

(v) (5 points) The bise-variance trade-off in SVM maybe controlled by

A. Choosing the kernel function
B. Choosing the penalty level
C. Choosing the loss function to approximate the objective function
D. Choosing the basis expansion

4. Tree-based models

(i) (5 points) CART is

A. a tree model
B. a type of linear model
C. likely to underfit if split until there is only 1 observation in each terminal node
D. likely to overfit if split until there is only 1 observation in each terminal node

(ii) (5 points) Which of the following choices may reduce the computational cost of random forests

A. Reducing mtry
B. Reducing nodesize
C. Reducing ntrees
D. All of the others

(iii) (5 points) Which of the following choices may lead to a higher risk of overfitting of random forests

A. Increasing mtry
B. Increasing nodesize
C. Increasing ntrees
D. All of the others

(iv) (5 points) Which of the following is true about random forests

A. It provides classification rules that can be expressed as linear combinations of features
B. It provides variable importance measures using permutation of out-of-bag data
C. It is able to handle high-dimensional data without convergence issues
D. All of the others

(v) (5 points) When we are always able to choose a base learner that gives good classification error (less than 1/2), the adaboost algorithm is guaranteed to

A. Give 0 training error eventually
B. Give 0 testing error eventually
C. Terminate within a certain number of iterations
D. Be exactly the same as a fine-tuned CART model

5. Other questions

(i) (5 points) A researcher is analyzing an education data. Each student was randomly provided one of the two online learning courses that teach the same topic. The class label for each student is denoted as T_i's. Their performances (denoted as Y_i's) are evaluated using a same set of exam questions. The researcher is interested in knowing which course provides a better education and recommend that to a new student. Which approach is appropriate for such a task?

A. Since the course is randomly provided, they are equally good

B. A t-test that compares the mean exam performance of the exam scores from these two courses and select the larger one.

C. A linear regression based on the model $\texttt{lm}(Y \sim T)$. Then for a new student, suggest the course label that has the larger predicted outcome.

D. A logistic regression based on the model $\texttt{glm}(T \sim Y, family = "binomial")$. Then for a new student, suggest the course label that has the larger predicted probability.

(ii) (5 points) A researcher is analyzing an education data. Each student was randomly one of the two online learning courses that teach the same topic. The class label for each student is denoted as T_i's. Their performances (denoted as Y_i's) are evaluated using a same set of exam questions. In addition, the students' personal background information (such as age and previous course taken) is also collected, and denoted as X_i's. The researcher is interested in knowing which course provides a better education and recommend that to a new student with a particular background X^*. Which approach is appropriate for such a task?

A. A linear regression based on the model $\texttt{lm}(Y \sim T)$. Then for a new student, suggest the course label that has the larger predicted outcome.

B. A linear regression based on the model $\texttt{lm}(Y \sim T + X)$. Then for a new student, suggest the course label that has the larger predicted outcome on X^*.

C. A linear regression based on the model $\texttt{lm}(Y \sim T + X + X * T)$. Then for a new student, suggest the course label that has the larger predicted outcome on X^*.

D. A virtual twin model