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

Midterm, 03/26/2021 **Net ID (Print):** _____

Time Limit: 2PM - 3PM

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. 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 Mar 26. 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	35	
2	15	
3	10	
4	15	
5	25	
Total:	100	

1. General questions

   (i) (5 points) What was the most frequently mentioned concept during the first half of this course?

   A. Degrees of freedom
   B. Bias-variance trade-off
   C. Difference between linear and nonlinear models
   D. Difference between different optimization algorithms

   (ii) (5 points) Which of the following models is suitable for handling high-dimensional data in general?

   A. Best subset selection
   B. Ordinary least squares estimator
   C. Lasso and Ridge regression
   D. KNN

   (iii) (5 points) Tuning parameters in the models we introduced may be used to

   A. Select the number of variables
   B. Detect outliers
   C. Alter the complexity of the model
   D. Reduce bias

   (iv) (5 points) Which of the following models do not involve tuning?

   A. KNN
   B. Lasso
   C. Ordinary least squares estimation
   D. Best subset selection

   (v) (5 points) If the true underlying regression function is linear and sparse, and the number of variables is large, which of the following can achieve small bias?

   A. KNN
   B. Ordinary least squares estimator
   C. Ridge regression
   D. Lasso regression

   (vi) (5 points) In a genetic study of the risk of cancer using thousands of variables (genes), it is unlikely that only a few genes determines the risk. The genes are also likely to be highly correlated. In this case, which model is more suitable?

   A. kNN
   B. Ordinary least squares estimator
   C. Ridge regression
   D. Lasso regression

   (vii) (5 points) Which of the following models use all variables for prediction at a future target point x_0

   A. Ordinary least squares estimator
   B. Lasso
   C. Ridge
   D. All of the above, but depends on the target point

2. Numerical Optimizations

   (i) (5 points) Which of the following models requires knowing the entire training dataset when predicting a new target point?

   A. KNN
   B. Best subset selection
   C. Lasso
   D. Ridge

   (ii) (5 points) An analytic solution exists for which of the following models

   A. KNN
   B. Ridge regression
   C. Best subset selection
   D. Elastic Net

   (iii) (5 points) Which of the following procedures may lead to different results if two people are running them independently

   A. Lasso with 10 fold cross-validation to select the best λ
   B. Ridge with leave-one-out cross-validation to select the best λ
   C. k-means clustering
   D. Forward stepwise regression

3. Bias-variance trade-off

   (i) (5 points) Bias-variance trade-off can usually be achieved by

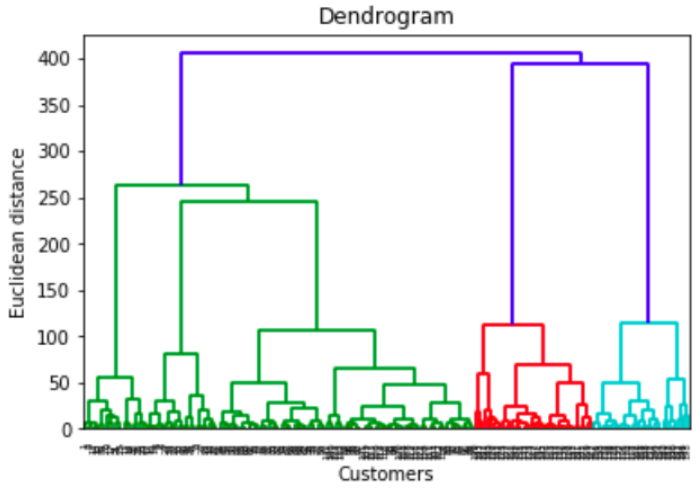
   A. Choosing the tuning parameter
   B. Change the optimization algorithm
   C. Choose a less/more complex model structure
   D. Look at a different target prediction point

   (ii) (5 points) Which of the following options will theoretically reduce bias?

   A. Increasing λ in Lasso and Ridge
   B. Increasing k in KNN
   C. Reducing λ in Lasso and Ridge
   D. Reducing k in KNN

4. Clustering

(i) (5 points) Based on the following figure, if a researcher is planning to construct three clusters, what should be done?



A. Cut at around 350
B. Cut at around 300
C. Cut at around 250
D. Cut at around 200

(ii) (5 points) Based on the above figure, which statement is true?

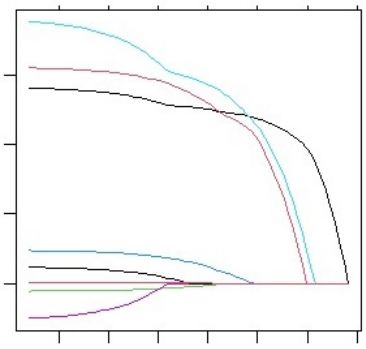
A. The red and light blue (the far right) clusters are (relatively) close to each other
B. The red and light blue (the far right) clusters are (relatively) far away from each other
C. If we want the overall within-cluster distance to be as small as possible, then we would let each observation to form its own cluster
D. The total amount of distance across all observations is about 400

(iii) (5 points) Which of the following can be solved using both coordinate descent and gradient descent?

A. $\widehat{\beta}^{\text{ols}}$
B. $\widehat{\beta}^{\text{lasso}}$
C. $\widehat{\beta}^{\text{ridge}}$

5. Practical Questions

(i) (5 points) The following plot is possibly produced by which model/algorithm?

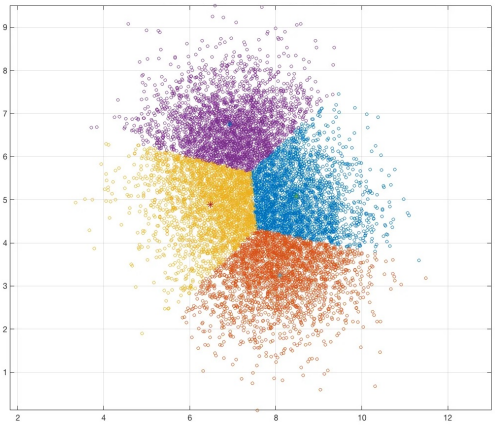


A. Lasso
B. Ridge
C. Backward step regression
D. Best subset regression and plot across all model sizes

(ii) (5 points) When running regression models on a few variables, the researcher decided to standardize a variable to 1 standard deviation. What is the impact of this data preprocessing step?

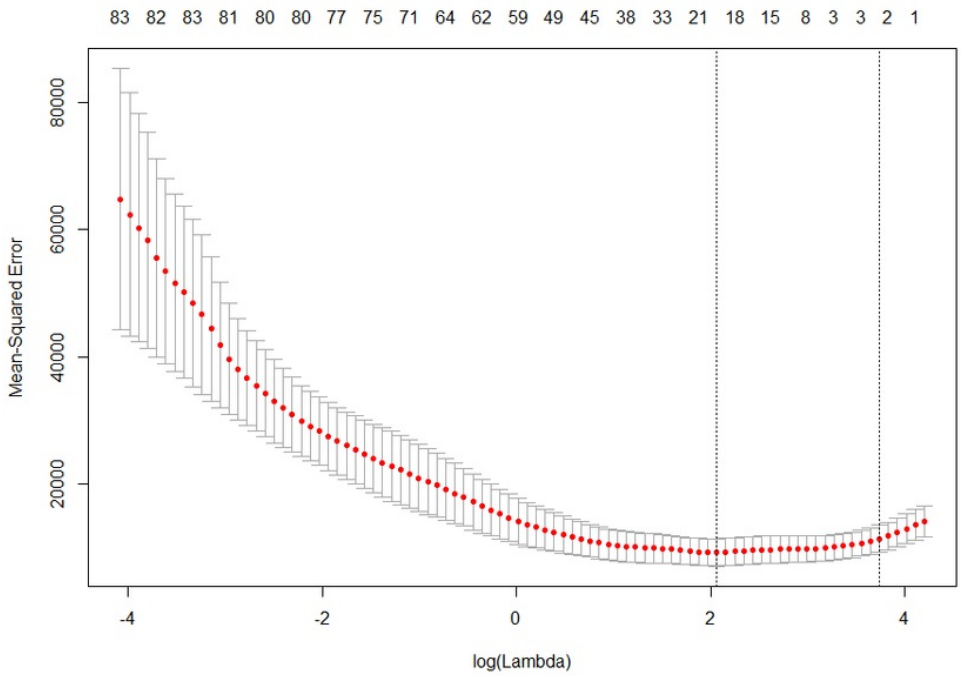
A. If fitting a linear regression model, this will slightly change the fitted value $\hat{y}$ on the training data, but it will not affect the testing data
B. If fitting a linear regression model, this will not change the fitted value $\hat{y}$ or the prediction on the testing data
C. If fitting a KNN, this may change the fitted regression line
D. If fitting a KNN, this will not change the fitted regression line

(iii) (5 points) A researcher performed clustering based on a two-dimensional data and obtained the following results. Which clustering algorithm was used?



A. k-means
B. hierarchical clustering
C. Self-organizing map
D. Spectral clustering

(iv) (5 points) The following plot is obtained. Which statement(s) must be true?



A. This is a cross-validation plot from Ridge regression
B. This is using 10-fold cross-validation
C. The best lambda value is around 2
D. There is a severe overfitting towards the left-hand side of this plot

(v) (5 points) Which method was invented earlier?

A. Lasso
B. Ridge
C. PCA
D. k-means