STAT 542, Spring 2022 **Name (Print):** _____

Midterm Exam, 04/08/2022 **Net ID (Print):** _____

Time: 3PM - 3:50PM

This exam contains 5 pages (including this cover page) and 20 questions. Each question worth 5 points. Please read the following descriptions and requirements carefully.

- Write all of your answers in the Table at Page 2.
- There are 20 questions. Each question worth 5 points.
  - For questions marked as **"Single Choice"**, there is only one correct answer. Selecting the wrong item loss all points.
  - For questions marked as **"Multiple Choices"** there must be more than one correct answer. Each wrongly selected item (correct but not selected, or incorrect but selected) cost 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 a closed-book exam.

Section	Points	Score
1	15	
2	15	
3	10	
4	15	
5	15	
6	15	
7	15	
Total:	100	

1. KNN

   (i) (5 points) [Single Choice] Which of the following models are similar to KNN?

   A. Penalized Linear regression
   B. Splines
   C. Kernel averaging ✓
   D. SVM

   (ii) (5 points) [Single Choice] Based on our understanding of the bias-variance trade-off of kNN, which of the following is closer to the prediction error of a 1NN model when the sample size is sufficiently large? (σ^2 is the variance of noise)

   A. σ^2
   B. $2\sigma^2$ ✓
   C. $3\sigma^2$
   D. $4\sigma^2$

   (iii) (5 points) [Multiple Choices] Two researchers fit kNN using two datasets generated based on the same physical mechanics. The two datasets are using the same design matrix $\mathbf{X}$ but independently observed outcomes Y conditioning on X. They use the same k to fit their models but observe quite different prediction errors on the same set of new testing data. This is a possible reflection of

   A. Large variance caused by small k ✓
   B. Large bias caused by large k
   C. Large random noise ✓
   D. Not tuning k properly ✓

2. Linear and penalized linear models

   (i) (5 points) [Multiple Choices] Which of the following is true about both Lasso and Ridge

   A. They can be solved using coordinate descent ✓
   B. They can be solved using gradient descent
   C. Larger λ leads to more shrinkage towards 0 ✓
   D. Larger λ leads to smaller variance ✓

   (ii) (5 points) [Single Choice] When we have orthogonal design matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X} = n\mathbf{I}$, and the linear model is true, the bias of a ridge regression estimator is

   A. $-\frac{\lambda}{1+\lambda}\boldsymbol{\beta}$ ✓
   B. $-\frac{1}{1+\lambda}\boldsymbol{\beta}$
   C. $-\frac{1}{1+\lambda^2}\boldsymbol{\beta}$
   D. $-\frac{\lambda}{1+\lambda}\|\boldsymbol{\beta}\|^2$

(iii) (5 points) [Single Choice] Marrow's C_p is motivated by

A. The difference between training and testing errors ✓
B. Reducing the cross-validation error
C. Correcting bias in a linear regression model
D. Reducing variance in a linear regression model

3. Numerical optimization

(i) (5 points) [Multiple Choices] Which of the following statements is true?

A. The coordinate-wise analytic solution exist for Lasso ✓
B. Newton-Raphson is based on first order Taylor expansion
C. Gradient descent assumes the Hessian matrix to be an identity matrix ✓
D. Lagrangian multiplier are used to solve constrained optimization ✓

(ii) (5 points) [Single Choice] For each loop of the coordinate-wise update of Lasso, where we update all variables once, the computational cost is

A. $\mathcal{O}(np)$ ✓
B. $\mathcal{O}(np^2)$
C. $\mathcal{O}(n^2p)$
D. $\mathcal{O}(n^2p^2)$

4. About spline

(i) (5 points) [Single Choice] Suppose you fit a piece-wise cubic regression that is continuous and continuous on the first derivative. If you have two pieces (1 knot), how many basis functions, i.e., free parameters, are invoved? (hint: how many constraints do you have?).

A. 5
B. 6 ✓
C. 7
D. 8

(ii) (5 points) [Single Choice] The roughness penalty involved in smoothing spline is penalizing

A. square of fitted function value
B. square of first-order derivative
C. square of second-order derivative ✓
D. square of third-order derivative

(iii) (5 points) [Single Choice] The degrees of freedom of a smoothing spline is

A. n
B. p
C. something less than n, depends on the tuning parameter ✓
D. something less than p, depends on the tuning parameter

5. Kernel Smoothing

(i) (5 points) [Multiple Choices] The λ parameter in a kernel density/regression model works as

A. Increasing λ increases bias ✓
B. Increasing λ decrease bias
C. Increasing λ increases variance
D. Increasing λ decrease variance ✓

(ii) (5 points) [Multiple Choices] Which of the followings are correct about Silverman's rule of thumb?

A. It is based on Gaussian assumptions of both target density and kernel function ✓
B. It works best in practice
C. It is in the rate of $\mathcal{O}(n^{-1/5})$ in a 1-dimension setting ✓
D. It is also optimal for Epanechnikov kernel

(iii) (5 points) [Multiple Choices] Which of the followings are correct about local polynomial regression

A. The performance depends heavily on the choice of kernel
B. The performance depends heavily on the choice of bandwidth ✓
C. It addressed boundary issues ✓
D. For each new testing data, we need to estimate a new set of parameters ✓

6. Classification models

(i) (5 points) [Single Choice] Given the p-dimensional Gaussian density function for each class $k \in \{1, 2\}$

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}_k)\right]$$

for $k = 1, 2$. Suppose two classes have equal number of training samples. What is the discriminate analysis decision boundary, as a function of x?

A. $2(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1} x + \widehat{\boldsymbol{\mu}}_1^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_2 = 0$
B. $2(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1} x - \widehat{\boldsymbol{\mu}}_1^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_2 = 0$ ✓
C. $-x^\top \widehat{\boldsymbol{\Sigma}}^{-1} x + 2(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1} x + \widehat{\boldsymbol{\mu}}_1^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_2 = 0$
D. $-x^\top \widehat{\boldsymbol{\Sigma}}^{-1} x + 2(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1} x - \widehat{\boldsymbol{\mu}}_1^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_2 = 0$

(ii) (5 points) [Multiple Choices] When $\boldsymbol{\Sigma}$ is not invertible in the previous question, we can consider

A. Use class-specific $\boldsymbol{\Sigma}_k$ instead
B. Add diagonal matrix to $\boldsymbol{\Sigma}$ ✓
C. Reduce the number of variables ✓
D. Make the sample size of each class more balanced

(iii) (5 points) [Multiple Choices] Which of the following statements is true regarding QDA?

A. It provides the Bayes rule when the data is generated based on this model, and we know the true parameters ✓
B. It is the same as logistic regression with quadratic terms
C. It is more flexible than LDA ✓
D. It is computationally more expensive than LDA ✓

7. Support vector machines

(i) (5 points) [Single Choice] In SVM, a slack variable $\xi_i = 0$ means this observation is

A. right on the margin
B. right on the decision line
C. On the margin or have larger correct gap with the decision line ✓
D. On the wrong side of the decision line

(ii) (5 points) [Single Choice] In the penalized version of SVM, which of the following component is changed to the loss part?

A. ξ_i's ✓
B. The $\boldsymbol{\beta}$ parameters
C. α_i's
D. None of the above

(iii) (5 points) [Multiple Choice] Which of the following can be solved using gradient descent?

A. Penalized version of SVM with logistic loss ✓
B. Penalized version of SVM with hinge loss
C. Penalized version of SVM with Modified Huber Loss ✓
D. All of the above