FIRST NAME: _____

LAST NAME: _____

STUDENT ID: _____

Summer 2023: STA 314H1S, Final Exam

Instructor: Ziang Zhang

Time allowed: 3 hours

Aid allowed: One non-programmable calculator

Instructions

- Fill out your name and student number both on the top of this page, and on the bubble sheet at the last page.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- If you possess an unauthorized aid during an exam, you may be charged with an academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If it is left in your pocket, it may be an academic offence.
- When you are done your exam, raise your hand for someone to come and collect your exam. Do not collect your bag and jacket before your exam is handed in.
- If you are feeling ill and unable to finish your exam, please bring it to the attention of an Exam Facilitator so it can be recorded before leaving the exam hall.
- In the event of a fire alarm, do not check your cell phone when escorted outside
- Students are expected to complete *all* the questions within the space provided. If the extra page is used, please indicate clearly which question is being answered.
- Write your answers clearly and show your detailed steps. If the examiner cannot read an answer or follow your steps, it will be marked as incorrect.
- You have to write your answers on the QR-coded pages. The non-QR coded pages can be used as scratch papers.
- For multiple choice questions, remember to fill your answers on the bubble sheet. You are suggested to fill the answers using pencils, so you can still change them.
- The total mark in this test is 100. Good luck!

Multiple Choices (20 pts): For each of these questions, select *the best* choice.

1. Which of the following is a non-parametric regression method?
   (a) Linear regression
   (b) KNN regression
   (c) Support vector machine classifier
   (d) There are more than one correct options above

2. Which of the following statement is true about a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$?
   (a) $\mathbf{x}^T \Sigma \mathbf{x} \geq 0$ for any $\mathbf{x} \in \mathbb{R}^p$
   (b) Σ must have a well-defined Eigen-decomposition.
   (c) Σ must have a well-defined singular value decomposition.
   (d) All the statements above.

3. Which of the following is a valid reason to use L^p norm regularization in a model?
   (a) By penalizing the magnitude of the regression coefficients, the training error could be improved.
   (b) By introducing an additional parameter λ, the flexibility of the model is improved.
   (c) For certain problem where the original MLE solution does not exist, adding regularization could ensure the existence of an unique solution.
   (d) None of the above.

4. The MAP estimate is obtained by maximizing which of the following quantities?
   (a) Prior
   (b) Likelihood
   (c) Posterior
   (d) Log likelihood

5. Assume $A, B \in \mathbb{R}^{p \times p}$, Which of the following statements is correct?

   (a) $|A + B| = |A| + |B|$

   (b) $\text{tr}(AB) = \text{tr}(A)\text{tr}(B)$

   (c) If both A and B are invertible, so will be $A + B$

   (d) None of the statement above is correct.

6. Which of the following is not necessarily true for a kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ that is positive definite?

   (a) For all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^p$, $K(c\mathbf{x}, \mathbf{z}) = cK(\mathbf{x}, \mathbf{z})$, where c is any scalar.

   (b) For all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^p$, $K(\mathbf{x}, \mathbf{z}) = K(\mathbf{z}, \mathbf{x})$.

   (c) There exists a function mapping ϕ such that for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^p$, $K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$.

   (d) For all $\mathbf{x} \in \mathbb{R}^p$, $K(\mathbf{x}, \mathbf{x}) \geq 0$.

7. Which of the following method involves the use of a greedy algorithm?

   (a) Forward selection method.

   (b) Backward selection method.

   (c) Recursive binary split.

   (d) All the methods above.

8. Which of the following statement correctly describes the gradient descend (GD) method?

   (a) All the GD methods can only use the first order information of the function.

   (b) The simple GD method cannot be directly applied to the loss function of the elastic net.

   (c) For the simple GD method, decreasing its learning rate α always make its convergence faster.

   (d) All the statement above are correct.

9. Which of the following metric should be used to describe node impurity of a classification tree?

   (a) Gini index

   (b) RSS

   (c) AUC

   (d) Tree depth

10. Which of the following is an example of the Ensemble method?

    (a) Bagging.

    (b) Random Forest.

    (c) Boosting.

    (d) All the choices above.

11. Which of the following method should be used for clustering, if you do not have information on the total number of clusters.

    (a) K-means clustering.

    (b) Hierarchical clustering.

    (c) Support vector classifier.

    (d) Regression tree.

12. Suppose our feature $\mathbf{x}^p$ is high-dimensional, following normal distribution with strong correlation within the feature vector. We would like to build a linear regression model based on this feature vector, but we would like to lower the variance of our prediction:

    (a) We should apply the naive Bayes assumption on the feature.

    (b) We should apply the LASSO method to select important features.

    (c) We should apply the PCA method to reduce the dimension of $\mathbf{x}^p$.

    (d) All the methods above will work well.

13. Which of the following is incorrect about the multivariate normal distribution?

    (a) Independence and uncorrelatedness are equivalent for multivariate normal random variables.

    (b) All marginal distributions of a multivariate normal distribution are also normal.

    (c) The covariance matrix of a multivariate normal distribution must be symmetric and positive-definite.

    (d) If x and y are both normal variables, their joint distribution is multivariate normal.

14. Which of the following classification method will always have zero training error rate, if its solution exists.

    (a) Maximal margin classifier

    (b) Support vector classifier

    (c) LDA and QDA classifier.

    (d) None of the methods above.

15. For a prediction $\hat{y}$ yielded from the Ridge regression model, which of the following metric will not change monotonically as λ increases?

    (a) The training RSS

    (b) Variance of the prediction

    (c) Bias of the prediction

    (d) The testing RSS

16. Which of the following classification method will not be defined for a set of linear separable data?

    (a) Logistic regression.

    (b) LDA method.

    (c) Classification tree.

    (d) Maximal margin classifier.

17. Which of the following statement is correct about conjugate prior?

    (a) A prior is conjugate if the prior and the likelihood have the same functional form.

    (b) Given a likelihood, the prior is called a conjugate prior if the posterior has the same functional form with the prior.

    (c) A conjugate prior must be Gaussian.

    (d) None of the statement above is correct.

18. Which of the following statement is true about the SVM classifier?

    (a) It is an example of the unsupervised method.

    (b) The choice of kernel will not affect the bias variance tradeoff of the SVM.

    (c) The number of misclassifications must be less than the size of the budget C.

    (d) All the statements above are correct.

19. Which of the following best describes the Naive Bayes assumption?

    (a) It assumes that all features in the model are mutually independent.

    (b) It assumes that all features in the model are conditionally independent, given the outcome class.

    (c) It assumes all features in the model have the same variance parameter.

    (d) None of the statement is correct.

20. Which of the following is necessarily a valid inner-product for vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$?

    (a) $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y} + c$ where c is some scalar.

    (b) $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{A}\mathbf{y}$ where $\mathbf{A}$ is a symmetric, positive-definite matrix.

    (c) $\langle \mathbf{x}, \mathbf{y} \rangle = (\mathbf{x}^T\mathbf{y})^2$.

    (d) $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{D}\mathbf{y}$, $\mathbf{D}$ is a diagonal matrix with diagonal terms that sum to 1.

1. (15 pts) (Regression) Assume you have collected observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$, where each feature $\mathbf{x}_i \in \mathbb{R}^p$.

   (a) (5 pts) Suppose $p = 5$, and you want to carry out a PC regression using the first two PCs. You carry out the following eigen-decomposition in R for the sample covariance matrix $\hat{\Sigma}$, such that $\hat{\Sigma} = UDU^T$:

```
> ## D matrix:
> D <- round(diag(Eigen_result$values),3)
> D
       [,1]  [,2]  [,3]  [,4] [,5]
[1,] 12.199 0.000 0.000 0.000    0
[2,]  0.000 7.434 0.000 0.000    0
[3,]  0.000 0.000 1.861 0.000    0
[4,]  0.000 0.000 0.000 0.136    0
[5,]  0.000 0.000 0.000 0.000    0
>
> ## U matrix:
> U <- round(Eigen_result$vectors,3)
> U
       [,1]   [,2]   [,3]   [,4]  [,5]
[1,]  0.247 -0.347  0.512 -0.597 0.447
[2,] -0.579  0.534 -0.192 -0.379 0.447
[3,]  0.622  0.071 -0.639  0.003 0.447
[4,] -0.441 -0.657 -0.188  0.372 0.447
[5,]  0.151  0.398  0.507  0.601 0.447
```

   Based on the output above, what are the first two PC directions? Compute the PVE of each PC and sketch a scree-plot.

(b) (5 pts) Continue from the last question, assume you have fitted a PC regression model using the first two PCs in R:

```
> mod <- lm(y~Z)
> mod

Call:
lm(formula = y ~ Z)

Coefficients:
(Intercept)           Z1           Z2
      0.567        3.408        1.836
```

where Z_j denotes the jth PC score. Suppose you have two new data:

$$\mathbf{x}_1 = [1, 0, -1, 2, 0], \quad \mathbf{x}_2 = [0, 0, -1, 1, 0],$$

compute the first two PC scores of each data, as well as its predicted $\hat{y}$.

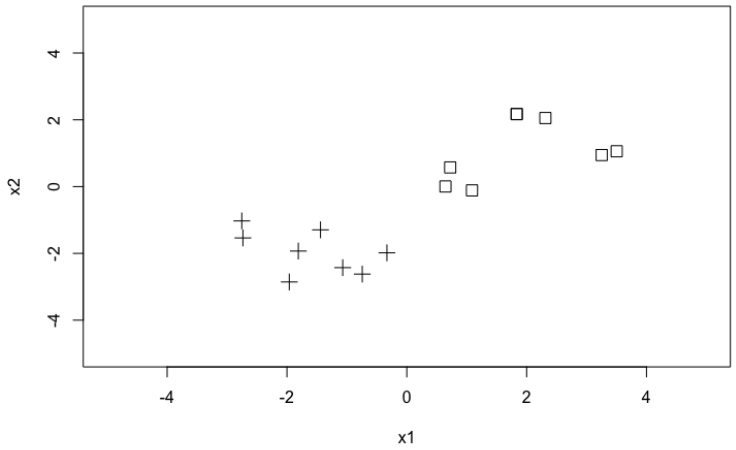
(c) (5 pts) Now suppose $n = 5$ and $p = 1$, and you want to build a predictive model for $x_0 = 0.05$. The training data is shown below:

```
> training_data
       Y      X
   0.004 -0.560
  -0.023 -0.230
   0.341  1.559
  -0.047  0.071
  -0.006  0.129
```

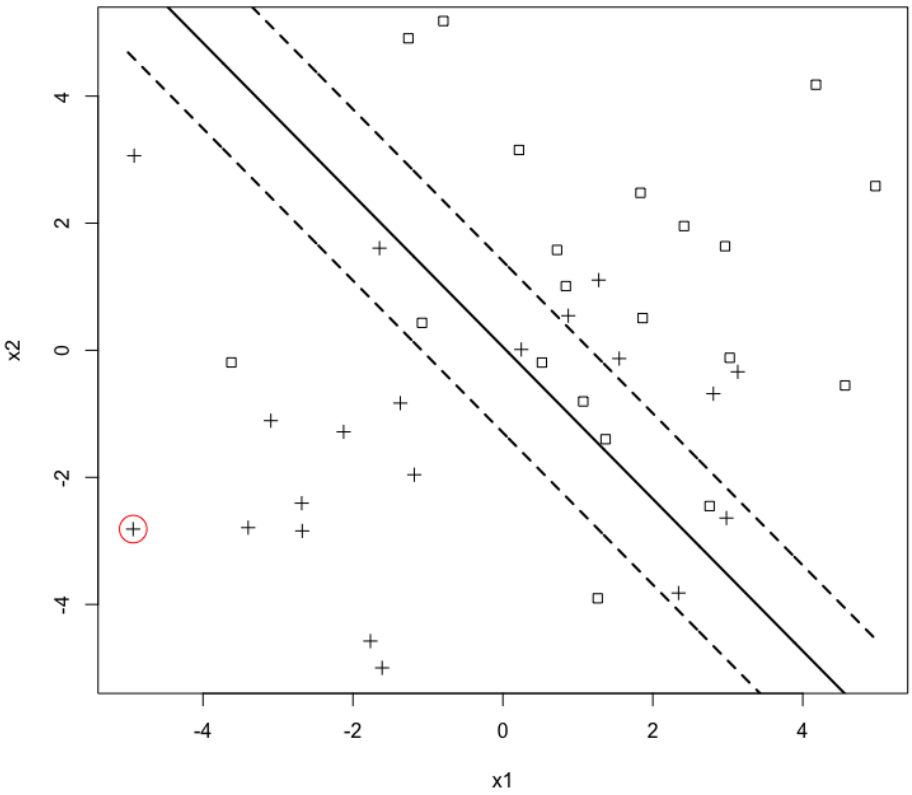
Based on the training data, construct a KNN regression prediction $\hat{y}(x_0)$ (based on the sample mean in the neighborhood) for $K = 1, 2$ and 5. Which prediction has the smallest variance? Which model has the smallest training residual sum of squares (RSS)?

2. (10 pts) (Classification)

   (a) (2 pts) Discuss the difference between the maximal margin classifier and the support vector classifier (SVC). On the figure below, draw the separating hyperplane as well as its margins, using the maximal margin classifier.



(b) (3 pts) Suppose you fit the following SVC using your training data, how many support vectors do you have? How many observations are missclassified by the SVC? What will happen to the fitted SVC if you change that circled training data point to $\mathbf{x} = [x_1, x_2] = [-10, -10]$?



(c) (4 pts) Discuss what will happen to the figure in (b) as the budget C increases. Also, what will happen to the bias and variance of your SVC model? What will happen to the number of support vectors?

(d) (1 pts) Based on the figure in (b), classify the following two test data points:

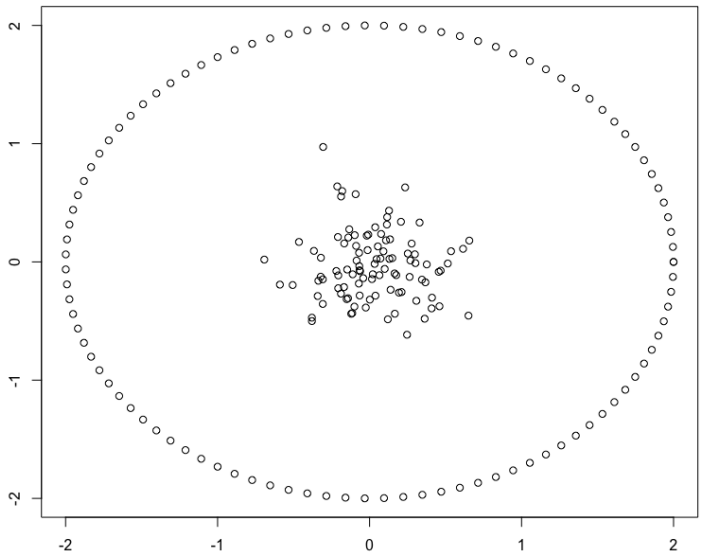
$$\mathbf{x}_1^* = [x_1, x_2] = [0, -1], \quad \mathbf{x}_2^* = [x_1, x_2] = [1, 4].$$

3. (10 pts) (Clustering)

   (a) (5 pts) Describe the procedure of the K-means algorithm, and describe what is the problem of local optimal in this algorithm, and how this problem can be mitigated to improve the performance of the algorithm.

(b) (3 pts) Define a dendrogram in the context of hierarchical clustering, and how it can be used for clustering.

(c) (2 pts) Suppose you have a dataset shown in the plot below, and you wish to carry out the Hierarchical clustering on this dataset to obtain two clusters. Which type of linkage would you prefer to use, complete or single linkage? Briefly explain why.



4. (20 pts) (Tree)

   (a) (5 pts) Assume you have the following training set, construct a classification tree to predict the type of the animal, such that each terminal leaf has zero misclassification rate. How many terminal leaves do you have? Do you expect the tree to overfit or underfit?

```
> animals_data
  fur.color body.weight   type
1    yellow           5    cat
2    yellow           4    cat
3     black          15    dog
4     black           6    cat
5     white          16    dog
6     white           2 rabbit
```

(b) (2 pts) Assume you have the following two observations, use the tree from (a) to classify them.

	fur.color	body.weight	type
1	yellow	5	?
2	white	10	?

(c) (3 pts) Describe how to construct a single prediction of the animal's type using the idea of Bagging.

(d) (5 pts) For classification tree, there are two types of impurity metric: the classification error rate and the entropy. Describe the difference between the two types of metric, and why is entropy often preferred over error rate. Also, provide a specific example of two sets that have the same error rate but different entropies.

(e) (5 pts) Describe how to use a 3-fold Cross-Validation (CV) to select the optimal value of the cost-complexity parameter (α) when fitting a regression tree using the recursive binary splitting method and cost-complexity pruning.

5. (10 pts) (Ensemble method) Suppose you have $p = 100$ features $x_1, ..., x_p$ to build a regression model to predict y.

   (a) (5 pts) To start with, you decide to obtain a bagging prediction with B weak learners, where each weak learner is a high-dimensional linear regression using all the features. Describe the procedure to obtain the bagging prediction, and explain why the weak learners will be correlated with each other.

(b) (3 pts) Propose an approach to decorrelate the weak learners from the previous questions. Explain why your approach could reduce the correlation. Do you expect the variance of your final ensemble prediction to increase or decrease?

(c) (2 pts) Now you wish to carry out a boosting algorithm with B weak learners, where each weak learner will be a tree stump, fill in the missing step in the following algorithm, and discuss whether choosing a large B could lead to overfit:

Algorithm

i. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.

ii. For b in 1 to B, repeat the following steps:

   A. Fit a tree $\hat{f}^b$ with 1 split to the training data with $\{r_i\}_{i=1}^n$ as the response.

   B. Update $\hat{f}$ by $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$

   C. **The Missing Step**

iii. Output the boosted model: $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$

6. (15 pts) (Bayesian inference)

   (a) (3 points) Assume you collected the data $\mathbf{y}$ with the likelihood denoted as $L(\mathbf{y}|\theta)$. Prove that the MAP of θ is the same as the MLE of θ when the prior $P(\theta)$ is a diffuse uniform (i.e. $P(\theta) = c$ for all $\theta \in \mathbb{R}$, where c is some constant).

(b) (5 points) Suppose $y_1, ..., y_n$ are $i.i.d$ observations from a Poisson distribution with rate λ. The prior for λ is Gamma(a, b). Derive the posterior distribution of λ.
Hint: If $x \sim$ Gamma(a, b), $f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$.
If $x \sim$ Poisson(λ), $p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$.

(c) (4 points) Assume you have a linear regression model with no intercept parameter, such that

$$y_i|x_{i1}, x_{i2} \sim N(\beta_1 x_{i1} + \beta_2 x_{i2}, 1)$$

You have collected a set of independent data $\{x_{i1}, x_{i2}, y_i\}_{i=1}^n$.
Show that the LASSO regression estimates of β_1, β_2 is equivalent to their MAP, with independent laplace priors on β_1, β_2.
Hint: The multivariate normal density is given by

$$f(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right),$$

and the density of Laplace distribution is given by

$$f(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

(d) (3 points) Consider a hierarchical Bayesian model with two levels:

1. Observations $\mathbf{y}$ are influenced by a parameter β. Their likelihood is given by $L(\mathbf{y}|\beta) = \prod_{i=1}^{n} L(y_i|\beta)$.
2. β follows a prior distribution conditional on λ, represented by $P(\beta|\lambda)$.
3. λ has its own prior distribution, $P(\lambda)$.

Under this setting, suppose you can evaluate any kinds of integral involving $L(\mathbf{y}|\beta)$, $P(\lambda)$, and $P(\beta|\lambda)$, explain how you would compute the following quantities step by step:

i. The posterior distributions $P(\beta|\mathbf{y})$ and $P(\lambda|\mathbf{y})$.

ii. The posterior predictive distribution for a new $i.i.d$ observation y_{new}.

Extra page:

Extra page:

Extra page:

Extra page:

Extra page: