**FIRST NAME:** ____________________________

**LAST NAME:** ____________________________

**STUDENT ID:** ____________________________

# Summer 2023: STA 314H1S, Practice Final

**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 classification algorithm?
   (a) Linear regression
   (b) PCA
   (c) K-means algorithm
   (d) None of the above

2. Which of the following statement is true about a positive definite matrix $A$?
   (a) $|A| > 0$
   (b) $\operatorname{tr}(A) > 0$
   (c) $A$ must be invertible
   (d) All the statements above

3. Which of the following is a reason to use the elastic net over LASSO for penalized regression?
   (a) LASSO does not have closed form solution, whereas the elastic net has.
   (b) The computation of elastic net is easier than LASSO when the number of observations is large.
   (c) LASSO is less stable than elastic net when features are highly correlated.
   (d) None of the above.

4. Which of the following statement is true for a multivariate random vector $\mathbf{x} = (x_1, ..., x_p)^T \sim N(\mathbf{0}, \Sigma)$?
   (a) $A\mathbf{x} \sim N(\mathbf{0}, A\Sigma A^T)$ for any matrix with $p$ columns
   (b) The marginal distribution of each $x_i$ must be normal
   (c) All the elements $\{x_i\}_{i=1}^p$ are independent if $\Sigma$ is diagonal
   (d) All the statements above

5. Which of the following metric assesses the performance of a classifier using a single number, while considering both the sensitivity and specificity of the classifier at different classification threshold:

   (a) RSS

   (b) ROC

   (c) AUC

   (d) Overall error rate

6. Which of the following algorithm relies on the kernel trick to accommodate the non-linear decision boundary?

   (a) KNN classification

   (b) Support vector machine

   (c) Maximal margin classifier

   (d) QDA

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

   (a) Recursive binary split

   (b) Cross-validation

   (c) PCA

   (d) All the methods above

8. Which of the following model could be trained using a simple gradient descend (GD) method?

   (a) Logistic regression

   (b) K-means algorithm

   (c) LASSO regression

   (d) Every ML model can be trained using simple GD

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

   (a) AUC

   (b) RSS

   (c) Entropy

   (d) Tree depth

10. Which of the following is an approach to simplify the computation of the Bayesian Hierarchical model?

    (a) Naive Bayes

    (b) Bayes classifier

    (c) Forward selection

    (d) Empirical Bayes

11. Which of the following linkage computes the dissimilarity between two clusters as the maximal pairwise dissimilarity?

    (a) Complete linkage

    (b) Average linkage

    (c) Simple linkage

    (d) None of the above

12. Given the sample covariance matrix $\hat{\Sigma}$ of $\mathbf{x}$, which of the following computes the second PC score in a correct way?

    (a) Compute the eigen-decomposition of $\hat{\Sigma}$, and the second eigenvalue will be that PC score.

    (b) Compute the eigen-decomposition of $\hat{\Sigma}$, and the second eigenvector $\mathbf{v}_2$ will be that PC score.

    (c) Compute the eigen-decomposition of $\hat{\Sigma}$ and record the second eigenvector $\mathbf{v}_2$. The PC score can be computed as $\mathbf{v}_2^T\mathbf{x}$.

    (d) None of the above

13. Which of the following best describes the performance of a regression model, as the number of feature $p$ increases?

    (a) As $p$ increases, the training RSS will monotonically decrease.

    (b) As $p$ increases, the squared bias of the model will monotonically decrease.

    (c) The testing MSE will never be smaller than the bayes error, regardless of the choice of $p$.

    (d) All of the above

14. Which of the following is a classification method that uses the conditional distribution of the feature $\mathbf{x}$?

    (a) LDA

    (b) K-means algorithm

    (c) Logistic regression

    (d) None of the methods above

15. Which of the following ensemble method decorrelates the weak learner by restricting the choice of features in the construction of each weak learner?

    (a) Random Forest

    (b) Bagging

    (c) Boosting

    (d) Cost complexity pruning

16. Which of the following classification method is only defined for a set of linearly separable data?

    (a) Logistic regression

    (b) Classification tree

    (c) LDA method

    (d) Maximal margin classifier

17. In order to obtain the MAP of a Bayesian model, one must:

    (a) compute the marginal likelihood $p(\mathbf{y})$

    (b) use a prior that is conjugate to the likelihood function

    (c) know the functional form of the prior and the likelihood

    (d) None of the statement above is correct

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

    (a) It replaces the inner product in the computation of SVC with a given choice of the kernel function.

    (b) The bias variance tradeoff of the SVM is only controlled by the size of the budge $C$.

    (c) The SVM classifier requires the conditional distribution of $\mathbf{x}$ to be multivariate normal.

    (d) All the statements above are correct.

19. Select the sentence that has the most accurate description:

    (a) Clustering is a set of supervised learning approaches, that aim to study the distribution $P(\mathbf{x})$.

    (b) Classification is a set of supervised learning approaches, that aim to study the distribution $P(y)$.

    (c) Both PCA and clustering are methods that do not use require the training data to be labelled.

    (d) K-NN is a non-parametric clustering method, and K-means is a non-parametric classification method.

20. Suppose you want to build a linear regression model to predict $y$ using the feature vector $\mathbf{x} \in \mathbb{R}^p$, where $p > n$:

    (a) If $\mathbf{x}$ is a multivariate normal random vector with strong correlation, you could consider obtain the first $k < n$ PCs and carry out a PC regression.

    (b) If you have a high-performance computer where the computation speed is not an issue at all, you could consider to find a smaller model using the best subset selection.

    (c) If you want to obtain the prediction in a Bayesian method, you could assign a Gaussian prior to the regression coefficient $\boldsymbol{\beta}$ and then compute the posterior predictive distribution.

    (d) All the statements are correct.

1. (17 pts) (Regression) Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\{y_i, x_i\}_{i=1}^n$ is a set of independent observations.

(a) (5 pts) Show that the MLEs are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}.$$

(b) (3 pts) Write out the formula of the predicted value $\hat{y}_i$ for $x = x_i$. Show that the predicted value of $y$ when $x = \bar{x}$ is $\bar{y}$.

(c) (2 pts) Suppose you want use this simple linear regression model to predict the number of car accidents based on the temperature today. Explain why the simple linear regression above may not work well due to some violated assumption. What transformation could you think of to make the assumption less violated?

(d) (2 pts) Suppose you want to compare the predictive performance of two simple linear regression models with different features, could you conclude based on the training MSE? What if you want to compare a simple linear regression with a multiple linear regression?

(e) (5 pts) Recall that for the simple linear regression model:

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SSX}} \right), \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\mathrm{SSX}}, \quad \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\mathrm{SSX}},$$

where $\mathrm{SSX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ and you can assume without proof that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators. Derive the testing MSE of $\hat{y}$ when $x = \bar{x}$. What are the values of the variance, squared bias and bayes error?
(Note: the feature can be treated as fixed constant)

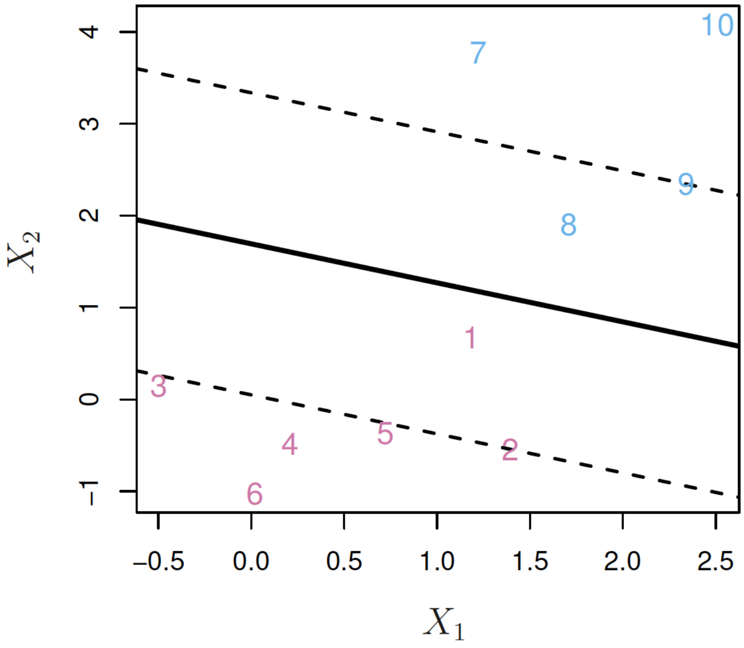
2. (10 pts) (Classification)

   (a) (2 pts) Assume the feature $\mathbf{x} \in \mathbb{R}^2$, and you have $n = 10$ observations, with 5 observations in each class. Draw a scatterplot where the observations are linearly separable, and a scatterplot where they are not.

(b) (4 pts) Define the notion of the support vectors in the context of the support vector classifier. What properties make them different from the other observations? Also, how does the notion of support vectors make SVC more robust to certain kind of measurement error than Logistic regression?

(c) (2 pts) Discuss the implication of the Mercer's theorem in the context of the support vector machine. How that helps to justify the computational advantage of the kernel trick compared to the direct feature mapping?

(d) (2 pts) Suppose you see the following fitted classifier, do you think it is a maximal margin classifier or a support vector classifier? How will you use that to classify an observation with $\mathbf{x} = [1, 1]$?



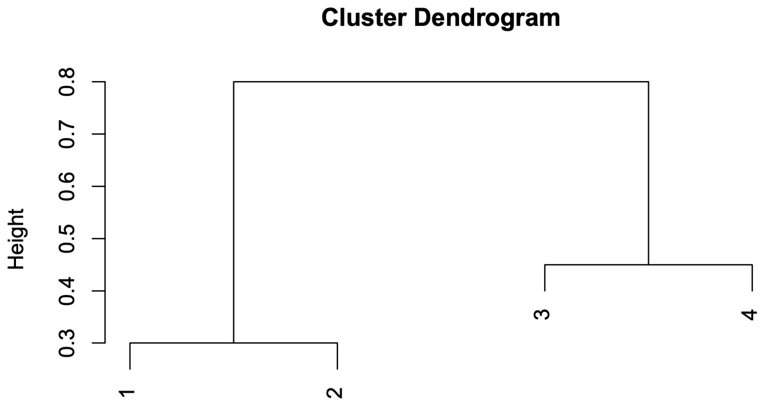
3. (10 pts) (Clustering)

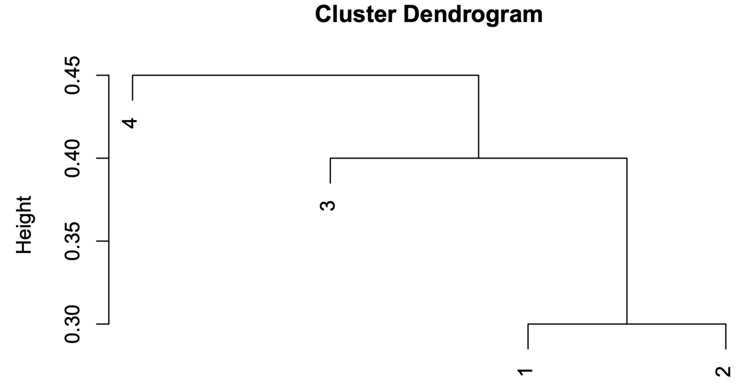
(a) (5 pts) Prove the following identity

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2 \sum_{i \in C_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2,$$

where $\bar{\mathbf{x}}_k$ denotes the sample mean (centroid) for cluster $k$. Explain why that implies the K-means algorithm will never increase the within cluster variation after each iteration.

(b) (4 pts) Suppose you obtain the following two dendrogram for your training data, using two different linkage. What would be your clustered result from each dendrogram if you decide to cut at 0.35? What if you cut at 0.7?



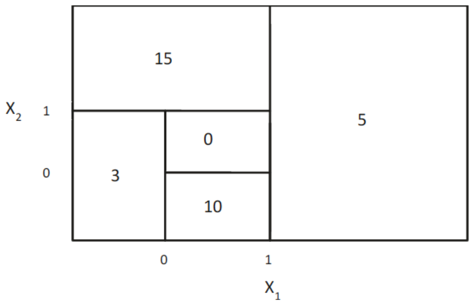


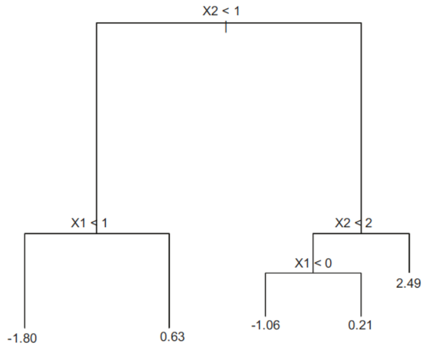
(c) (1 pts) Give an example with two clusters that is hard to be identified using single linkage, but easier using complete or average linkage.

4. (18 pts) (Tree)

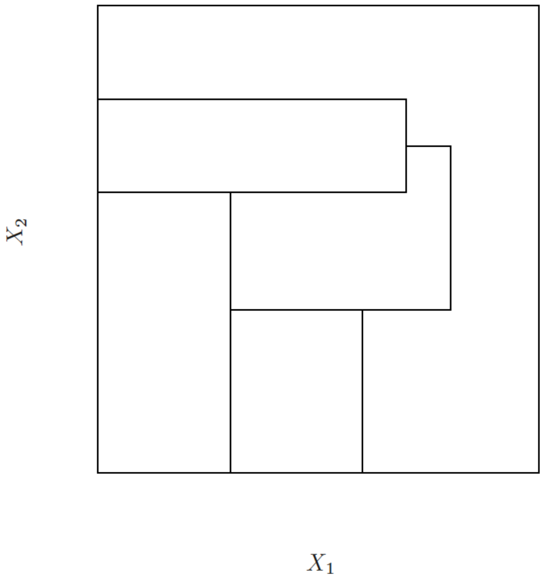
   (a) (4 pts) Define what is node impurity in the context of tree method. How could you measure that in regression tree, and in classification tree?

(b) (6 pts) Refer to the following figure, draw the corresponding tree of the feature space partition in the figure, and draw the corresponding partition of the tree in the figure: (the numbers inside the boxes indicate the mean)





(c) (2 pts) Is the following partition a possible result of the binary recursive split? Why or why not?



(d) (4 pts) Suppose both you and your friend want to build up a classifier using two features $x_1$ and $x_2$. Your friend considers to use a logistic regression model such that

$$\text{logit}P(y = 1|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

But you decide to use the classification tree method. In which scenario, your approach will be likely to perform better? Explain your answer.

(e) (2 pts) Discuss how would the training and testing MSE changes for a regression tree, as the number of terminal node increases.

5. (10 pts) (Ensemble method)

   (a) (4 pts) Suppose you have $B$ independent datasets $\{y_i^{(b)}, x_i^{(b)}\}_{i=1}^n$ for $b \in \{1, 2, ...B\}$, all from the same distribution $P_\theta(Y, X)$, where $\theta$ is some parameter that parametrizes the distribution. In each dataset, you obtain the prediction $\hat{y}^{(b)}$ for a new outcome $y_{new}$.
   Prove that the aggregated prediction $\hat{y}_{agg} = \sum_{b=1}^B \hat{y}^{(b)}/B$ will have the same bias as a single prediction $\hat{y}$, but has a improved variance.

(b) (4 pts) However, you only have access to one training set $\{y_i, x_i\}_{i=1}^n$. Explain how you could obtain $B$ datasets using the bootstrap. Also explain why the above result from part (a) no longer holds, when the $B$ datasets are obtained from bootstrap.

(c) (2 pts) Assume the target now is to estimate the unknown parameter $\theta$. From the $B$ bootstrap samples in (b), you compute the MLEs $\hat{\theta}^{(b)}$ for $b \in \{1, ..., B\}$. How will you estimate the standard error (i.e. deviation) of your MLE $\hat{\theta}$?

6. (15 pts) (Bayesian inference)

(a) (5 pts) Let $\mathbf{y} = \{y_i\}_{i=1}^{n}$ be *iid* sampled from a Geometric distribution with probability of success $\theta$:

$$P(Y_i = y|\theta) = (1-\theta)^{y-1}\theta, \quad y = 1, 2, \ldots$$

Assume that the prior distribution for the unknown parameter $\theta$ follows a Beta distribution with parameters $\alpha = \alpha_0$ and $\beta = \beta_0$:

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < \theta < 1$$

where $B(\alpha, \beta)$ is the Beta function. Derive the posterior $P(\theta|\mathbf{y})$.

(b) (5 pts) Based on (a), compute both $\hat{\theta}_{MAP}$ and $\hat{\theta}_{MLE}$. What do you notice from their difference?

(c) (2 pts) Suppose you want to obtain a prediction for a new independent observation $y_{new}$ using the posterior predictive distribution, how would you do it?

(d) (3 pts) Now suppose instead of using a fixed value $\alpha = \alpha_0$, you assign the prior $P(\alpha)$ to $\alpha$. How will you compute the posterior predictive distribution in (c) now? Briefly explain the new procedure.
(you don't need to solve the integral explicitly)