**FIRST NAME:** ___________________________

**LAST NAME:** ___________________________

**STUDENT ID:** ___________________________

# STA 314H1S, Practice Final (Solution)

**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)**: DDCDCBAACDACDAADCACD

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:

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

**Solution:** We want to find the values of $\beta_0$ and $\beta_1$ that maximize the likelihood function. The log-likelihood is:

$$\begin{aligned}\ell(\beta_0, \beta_1) &= \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)\right) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.\end{aligned}$$

The MLEs are the values that maximize this expression. Taking the partial derivatives with respect to $\beta_0$ and $\beta_1$ and setting them to zero, we get:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0.\end{aligned}$$

Equation above can be rearranged as:

$$n\beta_0 = \sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i \implies \beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Substituting this expression for $\beta_0$ into the original equation, we get:

$$\sum_{i=1}^n y_i x_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0,$$

which can be rearranged and solved for $\beta_1$:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The final solutions for the MLEs are:

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

(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 $\bar{y}$.

**Solution:**

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i, \\ \hat{y}_{\bar{x}} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\ &= \hat{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} \\ &= \hat{y} \end{aligned}$$

(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?

**Solution:** When using a simple linear regression model to predict the number of car accidents, the assumption that the conditional distribution of $y$ follows a normal distribution will be violated. This is because the number of car accidents must be a non-negative integer, whereas the normal distribution is defined for all real values.

To address this issue, we can use the logarithm transformation on the response variable: With the log transformation, we map the positive numbers representing car accidents to the entire real line, making the assumption of normally distributed errors more plausible.

(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?

**Solution:** For two simple linear regression, the conclusion could be based on the training MSE, as the two models have the same model complexity. However when comparing a simple linear regression model with a multiple linear regression model, selection based on the training MSE likely leads to overfitting, as a more complex model will have a smaller training error.

(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)

**Solution:** Recall that:

$$\mathrm{MSE} = \mathrm{Var}(\hat{y}) + \mathrm{Bias}(\hat{y})^2 + \text{Bayes Error}.$$

Using the given expressions for the variance of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, we can derive:

$$\mathrm{Var}(\hat{y}) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = \mathrm{Var}(\hat{\beta}_0) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) + 2\bar{x}\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1).$$

Plugging in the expressions:

$$\mathrm{Var}(\hat{y}) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SSX}} \right) + \frac{\sigma^2 \bar{x}^2}{\mathrm{SSX}} - 2\sigma^2 \bar{x} \frac{\bar{x}}{\mathrm{SSX}} = \frac{\sigma^2}{n}.$$

Since the estimators are unbiased:

$$\mathrm{Bias}(\hat{y})^2 = 0.$$

The Bayes error is the irreducible error, given by:

$$\text{Bayes Error} = \sigma^2.$$

Thus, the testing MSE of $\hat{y}$ when $x = \bar{x}$ is:

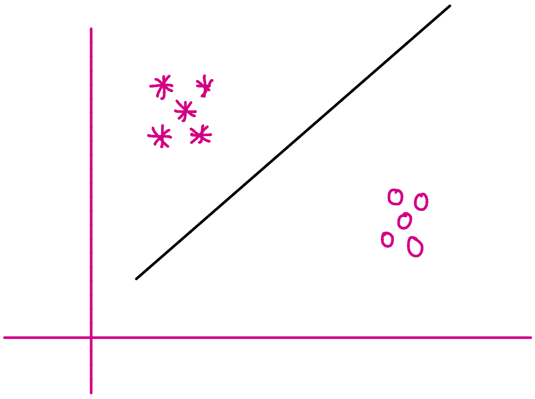
$$\mathrm{MSE} = \sigma^2 \left( \frac{n+1}{n} \right).$$

2. (10 pts) (Classification)

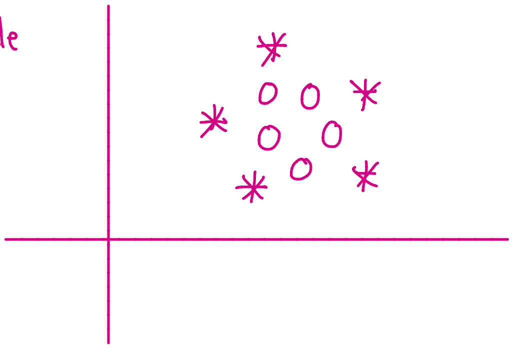
(a) (2 pts) Assume the feature $\mathbf{x} \in \mathbb{R}^2$, and you have $n = 10$ observations, with 5 observations in each class. Draw a scatterplot where the observations are linearly separable, and a scatterplot where they are not.

**Solution**:

Linearly separable :



Not linearly separable



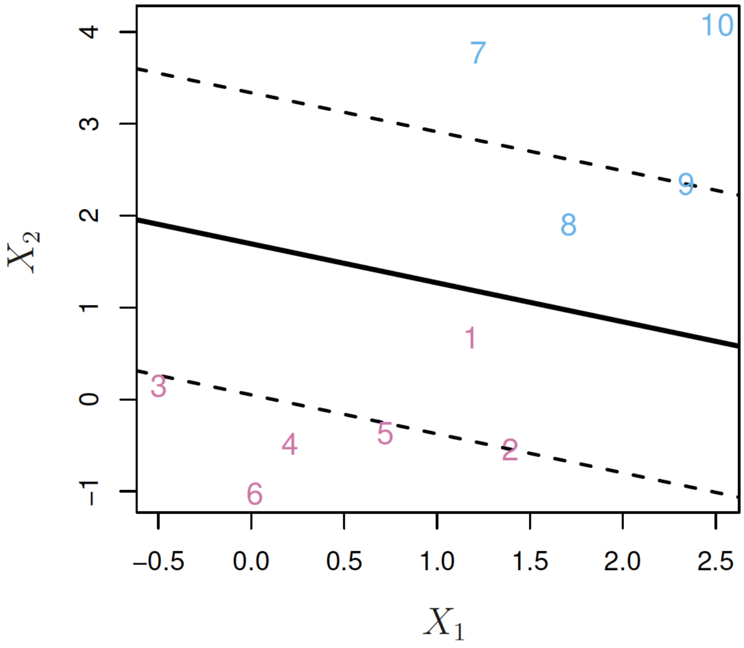
(b) (4 pts) Define the notion of the support vectors in the context of the support vector classifier. What properties make them different from the other observations? Also, how does the notion of support vectors make SVC more robust to certain kind of measurement error than Logistic regression?

**Solution**: Support vectors (SVs) are those observations that either lie on the margin, or on the wrong side of the margin.
SVs are the only observations that determine the fitted classifier. Whereas for other non-SV observations, their position can be changed without affecting the fitted classifier, provided that their new position is still non-SV.
When serious measurement error exists for those non-SV observations, the logistic regression will be severely affected, but not the SVC.

(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?

**Solution**: Mercer's theorem states that a kernel function can be expressed as an inner product in a high-dimensional space. This implies by using the kernel trick, the input space is implicitly mapped into a new, possibly higher-dimensional feature space.
The computation efficiency will be significantly improved when the mapped space is high or infinite dimensional, in which the direct method will not be computationally feasible.

(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 $\mathbf{x} = [1, 1]$?



**Solution**: This cannot be a MMC as observations 1 and 8 are on the wrong sides of the margin, whereas in MMC all the observations must be either on the margin or on the correct side of the margin.
The observation should be classified as red.

3. (10 pts) (Clustering)

(a) (5 pts) Prove the following identity

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

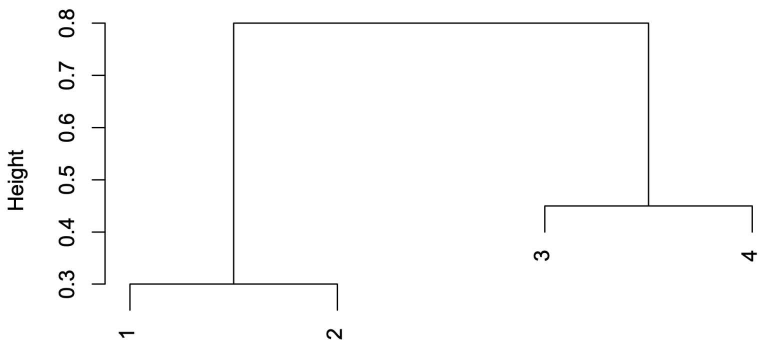
where $\bar{\mathbf{x}}_k$ denotes the sample mean (centroid) for cluster $k$. Explain why that implies the K-means algorithm will never increase the within cluster variation after each iteration.
(Note: you can assume for simplicity that the feature is univariate.)

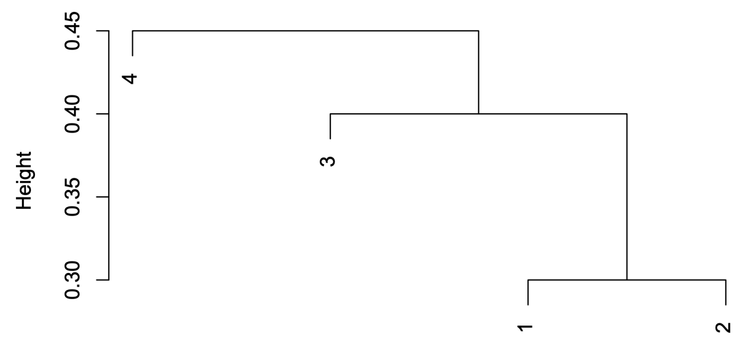
**Solution**:

$$\begin{aligned}
&\frac{1}{|C_k|} \sum_{i,j \in C_k} (x_i - x_j)^2 \\
&= \frac{1}{|C_k|} \sum_{i,j \in C_k} (x_i - x_j + \bar{x}_k - \bar{x}_k)^2 \\
&= \frac{1}{|C_k|} \sum_{i,j \in C_k} \big((x_i - \bar{x}_k)^2 + (x_j - \bar{x}_k)^2 + 2(x_j - \bar{x}_k)(x_i - \bar{x}_k)\big) \\
&= \frac{1}{|C_k|} \sum_{i \in C_k} \sum_{j \in C_k} \big((x_i - \bar{x}_k)^2 + (x_j - \bar{x}_k)^2 + 2(x_j - \bar{x}_k)(x_i - \bar{x}_k)\big) \\
&= \frac{1}{|C_k|} \sum_{i \in C_k} \left[ |C_k|(x_i - \bar{x}_k)^2 + \sum_{j \in C_k} \big((x_j - \bar{x}_k)^2 + 2(x_j - \bar{x}_k)(x_i - \bar{x}_k)\big) \right] \\
&= \frac{1}{|C_k|} \sum_{i \in C_k} \left[ |C_k|(x_i - \bar{x}_k)^2 + \sum_{j \in C_k} \big((x_j - \bar{x}_k)^2\big) + \sum_{j \in C_k} (2(x_j - \bar{x}_k)(x_i - \bar{x}_k)) \right] \\
&= \frac{1}{|C_k|} \sum_{i \in C_k} \left[ |C_k|(x_i - \bar{x}_k)^2 + \sum_{j \in C_k} \big((x_j - \bar{x}_k)^2\big) + (x_i - \bar{x}_k) \sum_{j \in C_k} (x_j - \bar{x}_k) \right] \\
&= \frac{1}{|C_k|} \sum_{i \in C_k} \left[ |C_k|(x_i - \bar{x}_k)^2 + \sum_{j \in C_k} \big((x_j - \bar{x}_k)^2\big) \right] \\
&= \frac{1}{|C_k|} \left[ \left[ \sum_{i \in C_k} |C_k|(x_i - \bar{x}_k)^2 \right] + \left[ \sum_{j \in C_k} |C_k|(x_j - \bar{x}_k)^2 \right] \right] \\
&= 2 \sum_{i \in C_k} |C_k|(x_i - \bar{x}_k)^2,
\end{aligned} \tag{1}$$

the seventh equality holds since $\sum_{i=1}^{n}(x_i - \bar{x}_k) = 0$.

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


Cluster Dendrogram

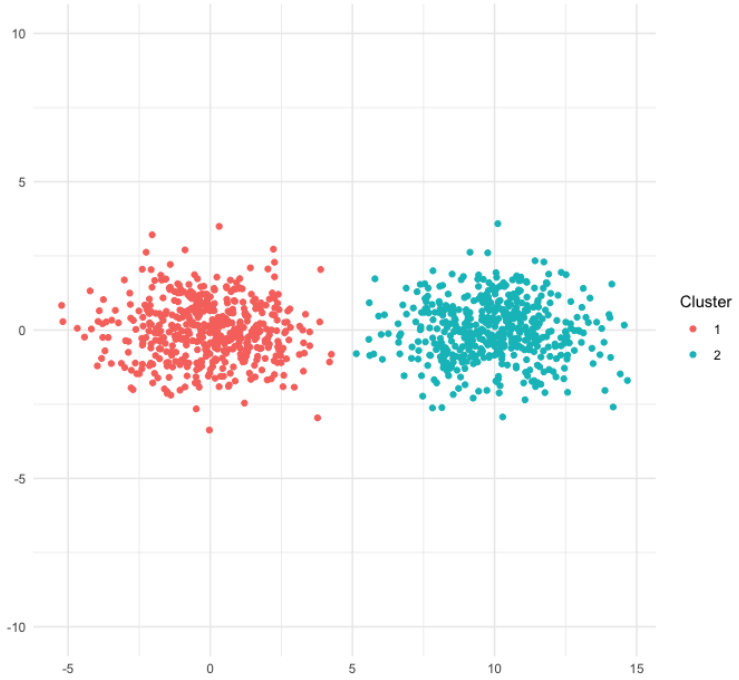

Cluster Dendrogram

**Solution:** When you cut at 0.35, the first dendrogram will have three clusters: $C_1 = \{1, 2\}, C_2 = \{3\}, C_3 = \{4\}$. The second dendrogram will have three clusters: $C_1 = \{4\}, C_2 = \{3\}, C_3 = \{1, 2\}$.
When you cut at 0.7, the first dendrogram will have two clusters: $C_1 = \{1, 2\}, C_2 = \{3, 4\}$. The second dendrogram will have one cluster: $C_1 = \{1, 2, 3, 4\}$.

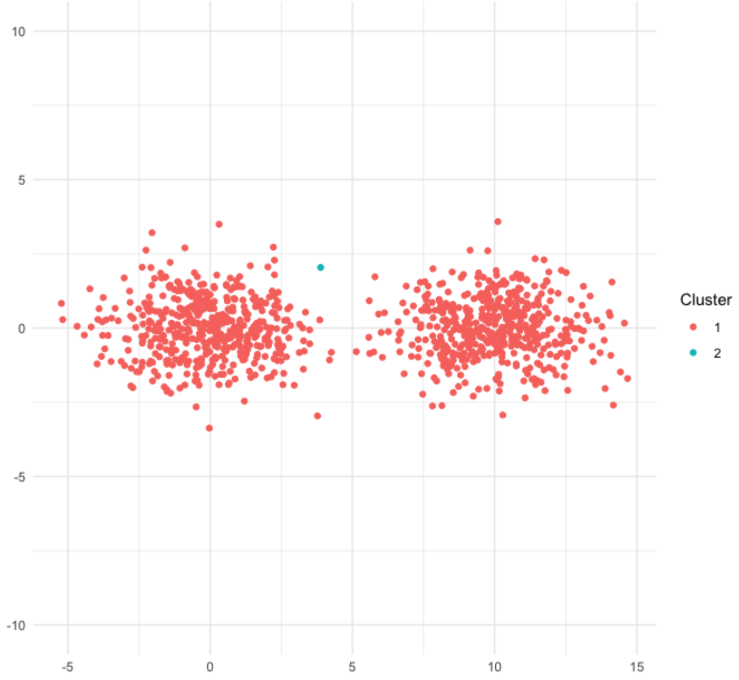
(c) (1 pts) Give an example with two clusters that is hard to be identified using single linkage, but easier using complete or average linkage.

**Solution:**

Complete Linkage



Single 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?

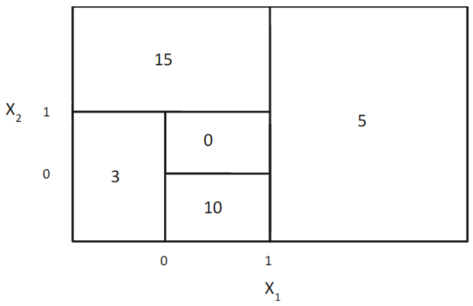
**Solution:** Node impurity measures the dissimilarity of outcomes within a node.

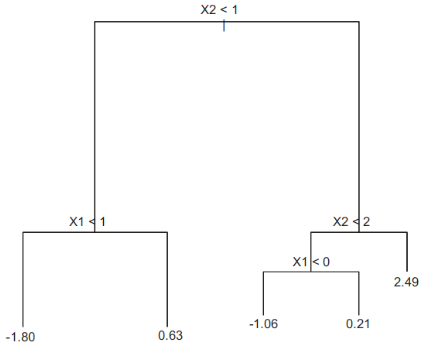
- In **regression trees**, it can be measured by the **MSE**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

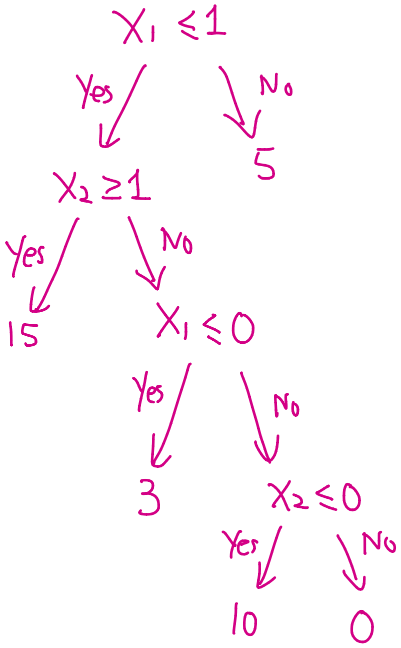
- In **classification trees**, it can be quantified using various methods:
  - **Gini index**: $\sum_{i=1}^{k} p_i(1 - p_i)$
  - **Entropy**: $-\sum_{i=1}^{k} p_i \log(p_i)$
  - **Classification Error**: $1 - \max(p_1, p_2, \ldots, p_k)$

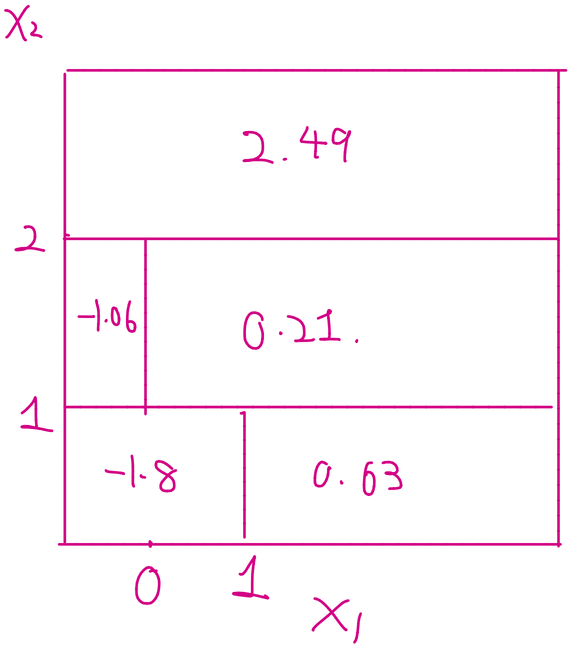
(b) (6 pts) Refer to the following figure, draw the corresponding tree of the feature space partition in the figure, and draw the corresponding partition of the tree in the figure: (the numbers inside the boxes indicate the mean)



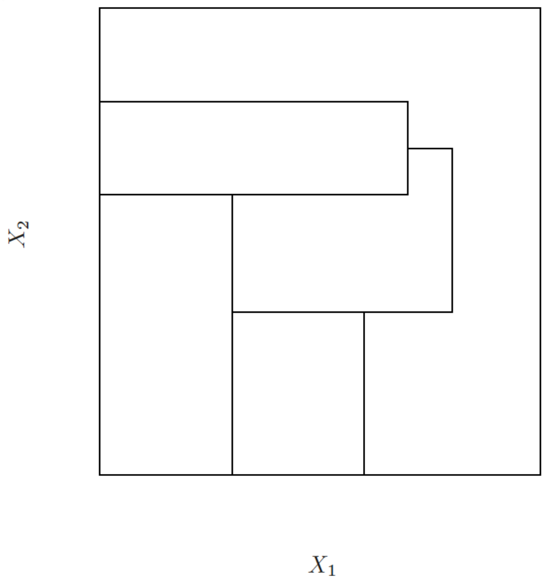


**Solution:**





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



**Solution:** No, if binary recursive split is used, there must exist a split that cuts the entire feature space into two half-planes.

(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.

**Solution:** Your approach using a classification tree would likely perform better in the following scenarios:

i. **Non-Linear Relationships**: Classification trees can model non-linear relationships between the features $x_1, x_2$ and the target $y$, whereas logistic regression assumes a linear boundary in the log odds.

ii. **Interactions between Features**: Classification trees can capture complex interactions between $x_1$ and $x_2$ by creating splits based on both features, whereas logistic regression might struggle with complex interactions.

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

**Solution:** As the number of terminal nodes in a regression tree increases:

- **Training MSE**: It typically decreases, as the tree can fit the training data more closely.
- **Testing MSE**: It usually decreases initially as the model becomes better at capturing the underlying patterns in the data. After a certain point, it starts to increase, showing that the model is overfitting the training data and not generalizing well to new, unseen data.

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 $\hat{y}^{(b)}$, but has a improved variance.

**Solution:**

$$\begin{aligned}
\text{Bias}(\hat{y}_{\text{agg}}) &= \frac{1}{B}\sum_{b=1}^{B} E(\hat{y}^{(b)}) - E(y) \\
&= \frac{B}{B}E(\hat{y}) - E(y) \\
&= \text{Bias}(\hat{y}) \\
\text{Var}(\hat{y}_{\text{agg}}) &= \text{Var}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{y}^{(b)}\right) \\
&= \frac{1}{B^2}\text{Var}\left(\sum_{b=1}^{B} \hat{y}^{(b)}\right) \\
&= \frac{1}{B^2}\sum_{b=1}^{B} \text{Var}(\hat{y}^{(b)}) \\
&= \frac{1}{B}\text{Var}(\hat{y})
\end{aligned}$$

(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.

**Solution:** Bootstrapping Procedure:

i. **For** $b = 1$ to $B$:

   A. Sample $n$ observations from the original dataset $\{y_i, x_i\}_{i=1}^n$ **with replacement** to create the $b$-th bootstrap dataset $\{y_i^{(b)}, x_i^{(b)}\}_{i=1}^n$.

Because the bootstrap samples will have a lot of overlapped observations, the predictions $\hat{y}^{(b)}$ are no longer independent when obtained from bootstrap samples. Hence: $\text{Var}\left(\sum_{b=1}^B \hat{y}^{(b)}\right) \neq \sum_{b=1}^B \text{Var}\left(\hat{y}^{(b)}\right)$ in the proof above.

(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 $\hat{\theta}$?

**Solution:**

i. **Compute the MLEs:** From each of the $B$ bootstrap samples, compute the MLE $\hat{\theta}^{(b)}$.

ii. **Calculate the sample mean of the bootstrap MLEs:**

$$\bar{\theta} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}.$$

iii. **Calculate the sample standard standard deviation of the bootstrap MLEs:**

$$\sqrt{\mathrm{Var}(\hat{\theta})} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\theta}^{(b)} - \bar{\theta} \right)^2}.$$

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 $P(\theta|\mathbf{y})$.

**Solution:**

**Likelihood Function:**

$$L(\mathbf{y}|\theta) = \prod_{i=1}^n (1-\theta)^{y_i-1}\theta = \theta^n(1-\theta)^{\sum_{i=1}^n (y_i-1)}$$

**Prior Distribution:**

$$P(\theta) = \frac{\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}{B(\alpha_0,\beta_0)}$$

**Posterior Distribution:**

$$\begin{aligned} P(\theta|\mathbf{y}) &\propto L(\mathbf{y}|\theta)\cdot P(\theta) \\ &\propto \theta^{n+\alpha_0-1}(1-\theta)^{\beta_0+\sum_{i=1}^n (y_i-1)-1} \end{aligned}$$

Since the posterior has the same functional form to the Beta distribution, we conclude that the posterior is a beta distribution with $\alpha = n + \alpha_0, \beta = \beta_0 + \sum_{i=1}^n (y_i - 1)$.

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

**Solution:** To find the MAF, note that the mode of a Beta distribution can be found at:

$$\frac{d}{d\theta}\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) = (\alpha-1)\theta^{\alpha-2}(1-\theta)^{\beta-1} - (\beta-1)\theta^{\alpha-1}(1-\theta)^{\beta-2} = 0$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{\alpha-1}{\alpha+\beta-2}$$

Plug in the value of $\alpha$ and $\beta$ derived for the posterior:

$$\hat{\theta}_{MAP} = \frac{n+\alpha_0-1}{n+\alpha_0+\beta_0+\sum_{i=1}^{n}(y_i-1)-2}$$

The MLE can be found by directly maximizing the log-likelihood:

$$\begin{aligned}\ln P(\mathbf{y}|\theta) &= n\ln\theta + \left(\sum_{i=1}^{n}(y_i-1)\right)\ln(1-\theta)\\ \frac{d}{d\theta}\ln P(\mathbf{y}|\theta) &= \frac{n}{\theta} - \frac{\sum_{i=1}^{n}(y_i-1)}{1-\theta} = 0\\ \hat{\theta}_{MLE} &= \frac{n}{n+\sum_{i=1}^{n}(y_i-1)}\end{aligned}$$

The MAP estimator incorporates prior information, while the MLE relies solely on the observed data. When a uniform prior is used (i.e. $\alpha_0 = \beta_0 = 1$), the two estimators are the same.

(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?

**Solution:** First, you could compute the posterior predictive distribution for $y_{new}$ using your result from (b).

$$P(y_{new}|\mathbf{y}) = \int P(y_{new}|\theta)P(\theta|\mathbf{y})\,d\theta$$

Using $P(y_{new}|\mathbf{y})$, you could either compute the mean $E(y_{new}|\mathbf{y})$ or the mode as your point prediction for $y_{new}$.

(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)

**Solution:** First, you will compute the joint posterior distribution for $\theta, \alpha$:

$$P(\theta, \alpha|\mathbf{y}) = \frac{P(\mathbf{y}|\theta)P(\theta|\alpha)P(\alpha)}{\int_\theta \int_\alpha P(\mathbf{y}|\theta)P(\theta|\alpha)P(\alpha) d\alpha d\theta}.$$

For the next step, since $y_{new}$ is not dependant on $\alpha$ given $\theta$, you can integrate out the hyperparameter from the posterior:

$$P(\theta|\mathbf{y}) = \int_\alpha P(\theta, \alpha|\mathbf{y})\, d\alpha.$$

Finally, the posterior distribution can be computed as before:

$$P(y_{new}|\mathbf{y}) = \int P(y_{new}|\theta)P(\theta|\mathbf{y})\, d\theta.$$