Last (Family) name: _____

First (Given) name: _____

Student ID (10 digits): _____

University of Toronto
Faculty of Arts & Science

DECEMBER 2023 EXAMINATIONS

STA314H1F
Statistical Methods for Machine Learning I

Duration: 2 hours
Aids Allowed: None

Exam Reminders:

- Fill out your name and student number on the top of this page.
- If a scantron and/or exam booklets are required: Ensure you fill in your name and student number on the scantron and/or exam booklet(s)
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- As a student, you help create a fair and inclusive writing environment. 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.

Special Instructions: There are 6 problems in total. Answer Problems 1 - 5 on the exam paper. Answer Problem 6, the multiple choices questions, on the scantron sheet.

Exam Format and Grading Scheme: The exam has 50 points in total. Each sub-question has its points indicated.

Students must hand in all examination materials at the end

Problem 1 (3 points)

Suppose we have n data points $x_1, \ldots, x_n$ generated i.i.d. from a 1-dimensional normal distribution $N(\mu, 1)$ with some unknown mean μ. Consider the sample mean estimator

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

State at least two ways of estimating the variance of $\widehat{\mu}$, based on $x_1, \ldots, x_n$.

Problem 2 (8 points)

Consider a regression problem

$$Y = f(X) + \varepsilon$$

with $X, Y \in \mathbb{R}$, where we use the natural cubic splines to parametrize $f(X)$.

(a) (3 points) For three specified knots $\xi_1 < \xi_2 < \xi_3$, verify that

$$f(X) = \beta_0 + \beta_1 X + \beta_2 \left[\frac{(X-\xi_1)_+^3 - (X-\xi_3)_+^3}{\xi_3 - \xi_1} - \frac{(X-\xi_2)_+^3 - (X-\xi_3)_+^3}{\xi_3 - \xi_2} \right]$$

represents a natural cubic spline. Here $(a)_+ = a$ if $a > 0$ and 0 otherwise, for any $a \in \mathbb{R}$.

(continue your answer of part (a).)

(b) (3 points) For the three specified knots $\xi_1 < \xi_2 < \xi_3$, write down a basis representation of $f(X)$ for cubic splines.

Based on this setting with 3 knots, deduce the number of parameters (including the intercept term) we need to estimate for natural cubic splines with K knots? State your reasoning.

(c) (2 points) If we consider to use natural cubic splines with K knots for predicting Y, how do you expect the test MSE to change as K increases? (You may reason based on the bias-variance decomposition)

Problem 3 (9 points)

Suppose we are interested in classifying if a student will score A in this class by using 2 features: the number of hours spent in this semester (X_1), and the current undergrad GPA (X_2). Let us encode $Y = 1\{\text{the student scores A}\}$.

(a) By using the past training data, we fit a logistic regression and obtain the estimated coefficients as $\widehat{\beta}_0 = -6$, $\widehat{\beta}_1 = 0.05$ and $\widehat{\beta}_2 = 1$.

(a1) (1 point) Estimate the probability that a student who studies 40 hours and has an undergrad GPA of 3.5 gets an A in this class. (Your final answer may contain the Euler's number e)

(a2) (1 point) How many hours would the student in part (a1) at least need to study to have a chance of getting an A in this class greater than 50%?

(b) Suppose we only focus on a group of students with undergrad GPA of 3.0 and would like to examine the effect of X_1 on the probability of getting an A in this class via logistic regression. Suppose we have data points of two students: (x_1, y_1) and (x_2, y_2).

(b1) (2 points) The MLE of the coefficient β_0 and β_1 are defined as

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg \min_{\beta_0, \beta_1 \in \mathbb{R}} \sum_{i=1}^{2} \left[-y_i \left(\beta_0 + \beta_1 x_{i1} \right) + \log \left(1 + e^{\beta_0 + \beta_1 x_{i1}} \right) \right] \tag{1}$$

Suppose we use the gradient descent to solve (1) with step size equal to α. Write down the iterative updates of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

(b2) (3 points) In the gradient descent you derived above, it is oftentimes difficult to specify α. For this reason, the Newton's method is used in practice. For solving the root of a given differentiable function f, that is, $f(x) = 0$, the Newton's method iterates as

$$x^{(t+1)} = x^{(t)} - \frac{f(x^{(t)})}{f'(x^{(t)})}, \qquad \text{for all } t = 1, 2, \ldots.$$

with $f'(\cdot)$ being the derivative of $f(\cdot)$.

Suppose $\beta_0 = -3$ in (1) and we only minimize over β_1. State how to adopt the Newton's method for finding the $\widehat{\beta}_1$ in (1). (You need to explicitly specify the expressions of $f(\cdot)$ and $f'(\cdot)$)

(continue your answer of part (b2).)

(b3) (2 points) Suppose the training data (x_1, y_1) and (x_2, y_2) are *well-separated*, that is, $\mathbb{P}(y_i = 1 \mid X = x_i)$ are arbitrarily close to 1 or 0, for all $i \in \{1, 2\}$.

Explain why using the Newton's method in part (b2) would lead to numerical instability at any iterate $\widehat{\beta}_1^{(t)}$ that is close to the true β_1.

Problem 4 (12 points)

Consider the problem of classifying a random, discrete, class label $Y \in \{0, 1\}$ based on a 1-dimensional feature $X \in \mathbb{R}$. We assume

$$\mathbb{P}(Y = 1) = \eta, \qquad \mathbb{P}(Y = 0) = 1 - \eta, \tag{2}$$

for some $0 < \eta < 1$, and

$$X \mid Y = 0 \sim N(-\mu, 1), \qquad X \mid Y = 1 \sim N(\mu, 1). \tag{3}$$

Let $(y_1, x_1), \ldots, (y_n, x_n)$ be n i.i.d. realizations of (Y, X) following (2) and (3). Answer the following questions.

Recall that the probability density function of $N(\mu, \sigma^2)$ at x is

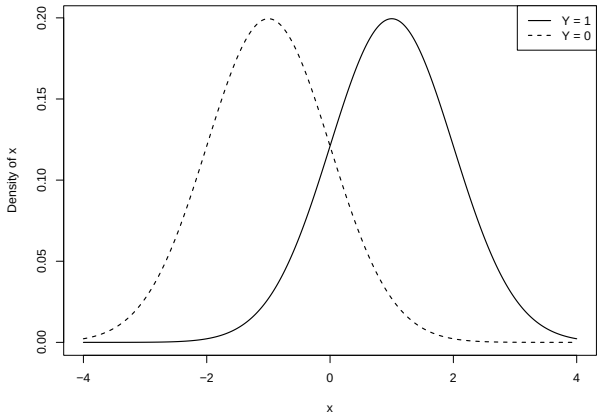
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

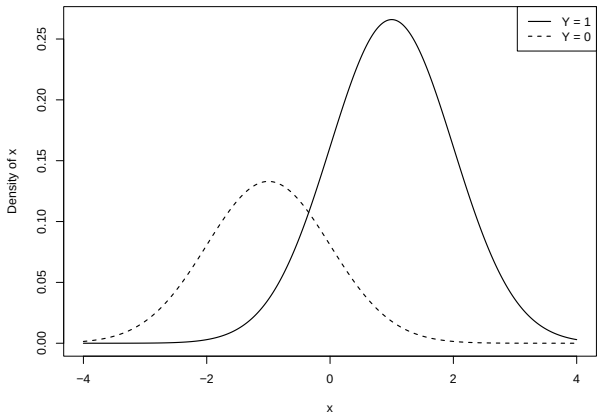
(a) (2 points) Prove that the Bayes classifier f^* is

$$f^*(x) = \begin{cases} 1 & \text{if } 2\mu x \geq \log\left(\frac{1-\eta}{\eta}\right) \\ 0 & \text{otherwise} \end{cases}, \qquad \text{for all } x \in \mathbb{R}.$$

(b) Assuming $\mu > 0$ throughout part (b).

(b1) (1 points) For the Bayes classifier you derived in part (a), draw its decision boundaries for $\eta = 1/2$ (in the first plot) and $\eta = 2/3$ (in the second plot), respectively. (You may use $\log_e(2) \approx 0.7$ and assume $\mu = 1$)





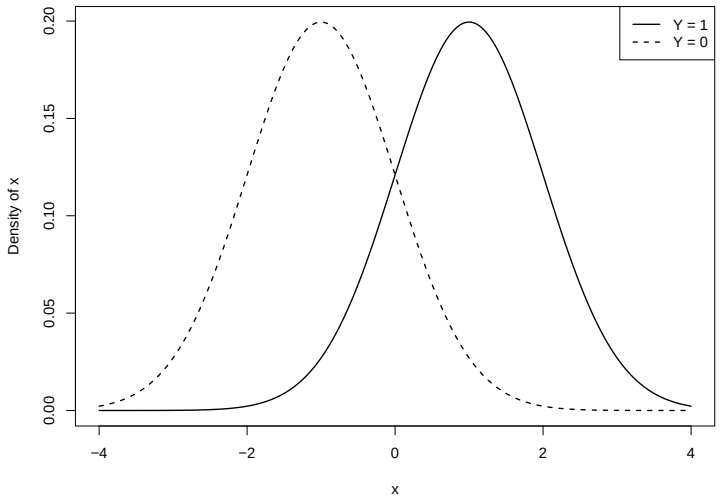
(b2) (4 points) For $\eta = 1/2$, indicate the Bayes error $\mathbb{P}(Y \neq f^*(X))$ on the plot below, and proves that

$$\mathbb{P}(Y \neq f^*(X)) = \Phi(-\mu).$$

Here $\Phi(t)$ is the c.d.f. of $N(0,1)$ at $t \in \mathbb{R}$, that is,

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$

Comment on how the Bayes error changes with μ.



(Continue your answer of part (b2).)

(c) (3 points) Prove that the maximum likelihood estimators (MLE) of η and μ are

$$\widehat{\eta} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \left[y_i x_i - (1-y_i)x_i\right]. \tag{4}$$

(Continue your answer of part (c).)

(d) (2 points) Suppose the training data set contains: $(y_1, x_1) = (1, 1)$, $(y_2, x_2) = (1, 0)$, $(y_3, x_3) = (1, 1/2)$ and $(y_4, x_4) = (0, -1/2)$. Suppose we estimate the Bayes classifier by $\widehat{f}$, the plug-in estimator based on the MLEs of η and μ in (4).

Compute the predicted label of $\widehat{f}$ at the test point $x = -1$. (You might use the fact that $\log_e(3) > 1$)

Problem 5 (8 points)

Answer the questions based on the following two fitted regression trees of using two features X_1 and X_2.

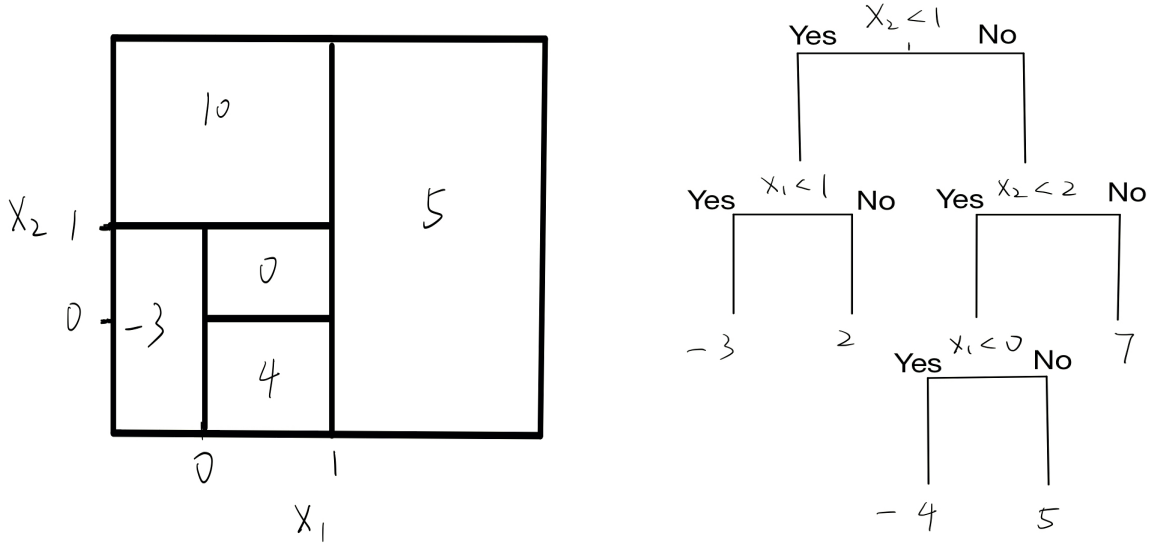


Figure 0.1: Two fitted decision trees

(a) (2 points) Sketch the fitted decision tree corresponding to the partition of the feature space in the left-hand panel of Figure 0.1. The numbers inside the boxes indicate the averaged responses of the training data within each region.

(b) (2 points) Create a diagram similar to the left-hand panel, using the tree illustrated in the right-hand panel of Figure 0.1. You should divide up the feature space into the correct regions, and indicate both the axis value of each split and the mean of each region.

(c) (2 points) For the three new data points $x_1 = (0.3, 0.8)$, $x_2 = (-0.5, 2)$ and $x_3 = (3, 0)$, state their predicted values from *each* of the individual decision tree above.

If the two fitted decision trees are from the bagging procedure, state the final predicted values of x_1, x_2 and x_3.

(d) (2 points) For the fitted tree corresponding to the left panel in Figure 0.1, denote it by T_0. Suppose each region of T_0 contains the same number of training data points.

The tree-pruning prcedure seeks a subtree $T \subset T_0$ such that

$$\sum_{j=1}^{|T|} \sum_{i \in R_j} (y_i - \bar{y}_j)^2 + \alpha |T| \tag{5}$$

is minimized. Here R_j denotes the jth region in the partition induced by T, $|T|$ is the number of leaf nodes of T and $\bar{y}_j$ is the averaged responses of the data points in R_j.

Suppose we use a value of α such that solving (5) gives a subtree $\widehat{T}$ with $|\widehat{T}| = 3$. Draw the subtree $\widehat{T}$ (similar to the right-hand panel in Figure 0.1) and briefly state your reasoning.

Problem 6 (10 points, 1 point for each subquestion)

Be sure to mark your answers on the answer sheet of multiple choice questions. There can be **one to four** correct answers to each question. One point is assigned to a multiple choice question **if and only if** all correct answers to this question are checked and no incorrect answer to this question is checked.

1. Which of the following statements are true

   A The Bayes classifier usually has the smallest training error rate among all possible classifiers.

   B The Bayes classifier usually has the smallest expected test error rate among all possible classifiers.

   C When Y has three categories (e.g., $Y \in \{1, 2, 3\}$), the Bayes error rate can be greater than 0.5.

   D The training error rate of the Bayes classifier is always greater than 0.

2. Which of the following statements are true

   A Logistic regression is a regression approach.

   B The maximum likelihood estimator (MLE) of the coefficients under logistic regression has a closed-form expression.

   C The Bayes classifier in logistic regression has a decision boundary that is linear in the feature space.

   D Logistic regression cannot handle qualitative features.

3. Which of the following statements are true

   A The decision boundary of the Bayes classifier in LDA is linear in X.

   B LDA cannot handle qualitative features.

   C LDA cannot be used when the response Y has three categories (e.g., $Y \in \{1, 2, 3\}$).

   D In LDA, we cannot estimate the conditional probability $\mathbb{P}(Y = 1 \mid X = x)$.

4. Which of the following statements are true

   A QDA is a more flexible method than LDA.

   B Naive Bayes classifier is a more flexible method than QDA.

   C QDA usually has smaller training error rate than LDA.

   D QDA usually has smaller test error rate than LDA.

5. Which of the following statements are true

   A Gradient descent can only be used to solve convex optimization problems

   B Gradient descent can only be used to solve optimization problems with differentiable loss function

   C Stochastic gradient descent uses only one data point per iteration

   D Stochastic gradient descent can cost more iterations to converge

6. Which of the following statements are true

   A For convex optimization problems, using gradient descent always converges regardless of the choice of step size

   B A too small step size leads to overfitting

   C A too small step size takes longer to converge

   D A too large step size renders the algorithm not to converge

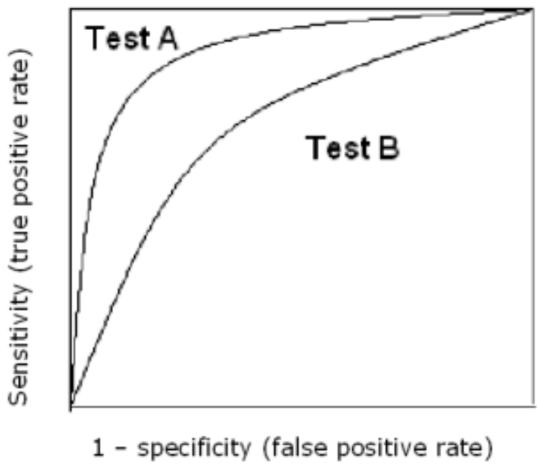
7. In the following plot, we compare two classification methods (called Test A and Test B) based on their ROC curves on the training data. Which of the following statements are true?

   A Test B is better than A in the training data set.

   B Test A is better than B in the training data set.

   C Test A has smaller test error than B.

   D Test B has smaller test error than A.



8. Which of the following statements are true between two classifiers

A The classifier with lower false negative rate (FNR) must have higher false positive rate (FPR)

B The classifier with lower FNR must have smaller misclassification rate

C The classifier with both lower FNR and lower FPR must have smaller misclassification rate

D The classifier with both lower FNR and lower FPR does not necessarily have smaller misclassification rate

9. Which of the following statements are true?

A Simple decision trees are easy to interpret

B Simple decision trees cannot handle categorical features

C Fitted simple decision trees can have large variance

D Simple decision trees can have zero training error

10. Which of the following statements are true

A Bagging and random forests are used to reduce the bias of simple decision-trees.

B Both bagging and random forests are used to reduce the variance of simple decision-trees.

C Bagging and boosting can only be used in the context of decision-trees.

D Both bagging decision trees and random forests use bootstrap samples.

(You may use this page as scratch paper if needed. This is the last page of the exam.)