PLEASE HAND IN

PLEASE HAND IN

UNIVERSITY OF TORONTO
Faculty of Arts and Science

DECEMBER 2021 EXAMINATIONS

CSC 2515HF

Duration: Please submit on MarkUs by Monday December 6 at 16:59.

Aids Allowed: You may consult the course slides and your notes.

Student Number: 1008046124

Last (Family) Name(s): Zhu

First (Given) Name(s): Bingzhang

Please read *carefully* every reminder on this page.

- Fill out your name and student number above—do it now, don't wait!
- This take-home test consists of 12 questions on 21 pages (including this one).
- You may either (1) print these pages, answer each question directly on the examination paper, and then scan it and upload it as a PDF file, or (2) type your answers using your favourite word processor using the same order of questions as here. In the latter case, make sure you use the right question numbers and part (e.g., 4(a), 7(c), etc.). If you don't, you may not get the mark. Points will be deducted if we have a hard time reading your solutions.
- You should help us having a fair take-home test. You may consult the slides and your notes. We in fact encourage you to do so. But do not discuss the questions with anyone else (including on Piazza). And do not search for answer to these questions on the Internet.
- Do not share this take-home test with anyone else, even after the semester ends, as we may reuse some of these questions in the future.
- There will not be an auto-fail in this take-home test. But try hard to do a good job. This will be a good opportunity to practice your ML skills once more, and get feedback.
- **Late Submission:** There is no grace period or a gradual late penalty, unless there is an emergency.
- **Collaboration:** The take-home test must be done individually, and you **should not** collaborate with others (this is different from your homework assignments).

MARKING GUIDE

Nº 1:	_____/	12
Nº 2:	_____/	16
Nº 3:	_____/	5
Nº 4:	_____/	6
Nº 5:	_____/	9
Nº 6:	_____/	5
Nº 7:	_____/	8
Nº 8:	_____/	10
Nº 9:	_____/	8
Nº 10:	_____/	4
Nº 11:	_____/	15
Nº 12:	_____/	2
TOTAL:	_____/	100

Good Luck!

OVER. . .

Question 1. True or False Questions [12 MARKS]

For each of these questions, a correct answer is +1 point, an empty answer is 0 point, and a wrong answer is −1 point.

	Statement	True	False
(1)	Deciding whether a mushroom is edible or not based on its image is an example of a regression problem		X
(2)	A K-Nearest Neighbourhood method with larger K may generalize worse than a one with a smaller K	X	
(3)	Decision trees can achieve zero classification error on any training data (assuming each training data point is unique)	X	
(4)	A lower entropy implies lower uncertainty	X	
(5)	Linear regression has to be linear in both parameters and features		X
(6)	Covariance matrix can have negative values	X	
(7)	Squared error loss in regression is only suitable if the target values are from a Gaussian distribution		X
(8)	The ℓ_1 regularization cannot shrink parameters to zero, hence it can be used for the purpose of feature selection		X
(9)	PCA can be used as a feature selection method	X	
(10)	Bagging decreases bias		X
(11)	AdaBoost cannot overfit		X
(12)	Projection to the highest variance direction is the same as projection to the direction that minimizes the total squared norm of each point to its projection	X	

Question 2. Short Answer Questions [16 MARKS]

Answer these questions concisely. In most cases, one or two sentences are enough.

Part (a) Fill in the blanks. [3 MARKS]

- Information Gain $IG(Y, X) = H(\ldots Y \ldots) - H(\ldots Y \ldots | X)$, where $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ is the entropy.
- If X and Y are independent, then $H(X, Y) = \ldots H(X) + H(Y) \ldots$
- If X and Y are independent, then $H(X|Y) = \ldots H(X) \ldots$.

Part (b) Fill in the blanks, using the following terms Posterior, Likelihood, Prior, Evidence. [3 MARKS]

$$\text{Posterior} = \frac{(\text{Prior}) \times (\text{Likelihood})}{(\text{Evidence})}$$

Part (c) Explain the relationship between posterior probability and prior probability given the likelihood. [1 MARK]

Assuming Evidence stay unchanged, the prior probability will increase as the posterior probability increase.

Part (d) Why do we use validation set? Explain briefly. [1 MARK]

The validation set is used to tune hyper-parameters and avoid over-fitting.

Part (e) What happens if we use the test set for tuning hyper-parameters? Explain briefly. [1 MARK]

The model will have high accuracy on the test data. but poor accuracy on the unseen data.

Part (f) Suppose that your classifier achieves poor accuracy on both the training and test sets. Which would be a better choice to try to improve the performance: Bagging or Boosting? Briefly justify your answer. [2 MARKS]

Boosting. As the classifier performs bad on both dataset, it has a high bias. The boosting method works on the predicts missclassified sample to reduce bias, while Bagging works well for high variance and low bias estimator.

Part (g) What is the effect of Boosting and Bagging on the bias/variance tradeoff? [4 MARKS]

Boosting could help reduce bias but increase variance; Bagging could help reduce variance but the bias will not change much.

Part (h) What is the main modelling assumption in the Naïve Bayes classifier? [1 MARK]

Naïve Bayes assumes that feature x_i are conditionally independent given the class t, so that,

$$p(x_i, x_j \mid t) = p(x_i \mid t)\, p(x_j \mid t)$$

Question 3. Hyper-Parameter Identification [5 MARKS]

Consider the following ML methods. Write down one or more hyper-parameters used in each method.

Part (a) KNN (1 answer) [1 MARK]

k

Part (b) Deep Neural Networks (2 answers) [2 MARKS]

Activation function. learning rate

Part (c) Support Vector Machines (1 answer) [1 MARK]

Kernal method

Part (d) PCA (1 answer) [1 MARK]

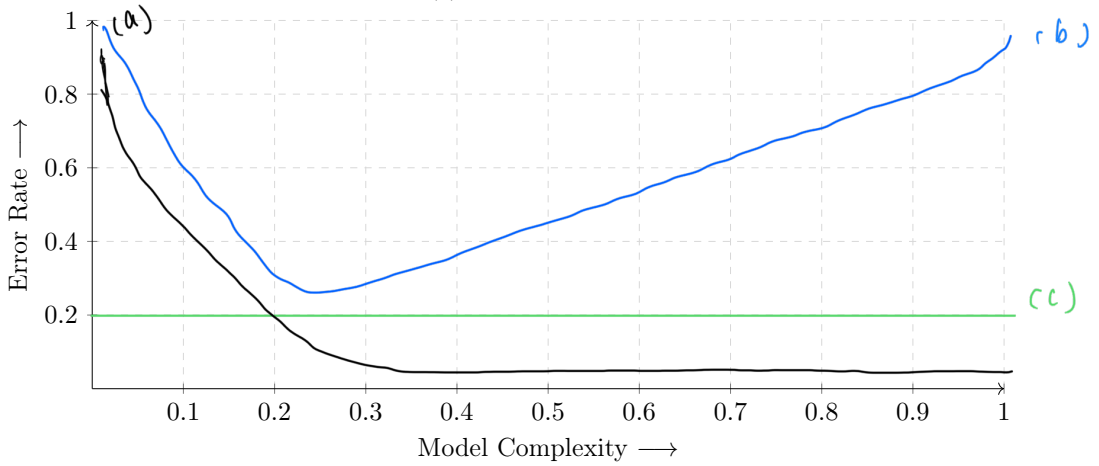
k = Number of the low dimensional subspace.

Question 4. Training/Testing Error Curves [6 MARKS]

This question asks you to show your general understanding of underfitting and overfitting as they relate to model complexity and training set size. Given the provided axes, where both vertical and horizontal are scaled between 0 and 1, draw your graphs with increasing error upwards and increasing complexity/training set size rightwards. Make sure that you clearly mark each curve.

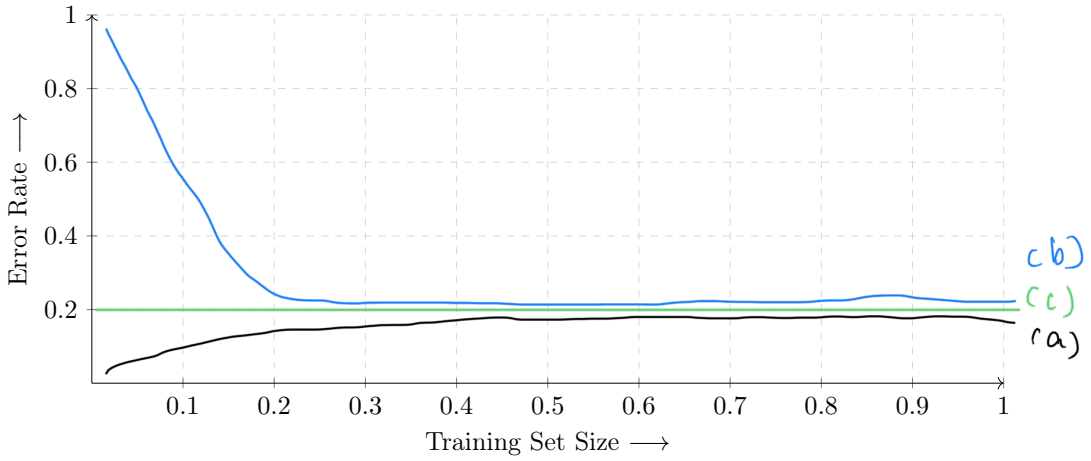
Part (a) [3 MARKS]

For a fixed training set size, (a) sketch a graph of the typical behaviour of training error rate versus model complexity in a learning system. (b) Add to this graph a curve showing the typical behaviour of test error rate (for an infinite test set drawn independently from the same input distribution as the training set) versus model complexity, on the same axes. (c) Mark a horizontal line showing the Bayes error.



Part (b) [3 MARKS]

For a fixed model complexity, (a) sketch a graph of the typical behaviour of training error rate versus training set size in a learning system. (b) Add to this graph a curve showing the typical behaviour of test error rate (again on an iid finite test set) versus training set size, on the same axes. (c) Mark a horizontal line showing the Bayes error.



Question 5. Decision Trees [9 MARKS]

The two dimensional input space shown in Figure 1 is partitioned into five (5) regions R1, R2,..., R5. We also show training data points consisting of circles and crosses.

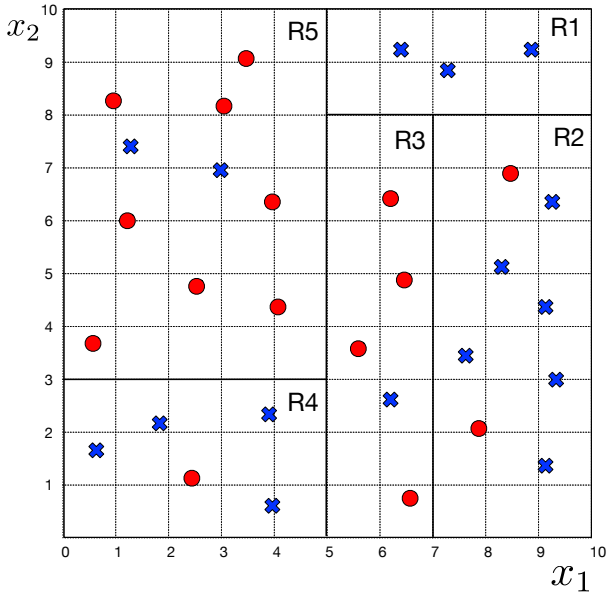
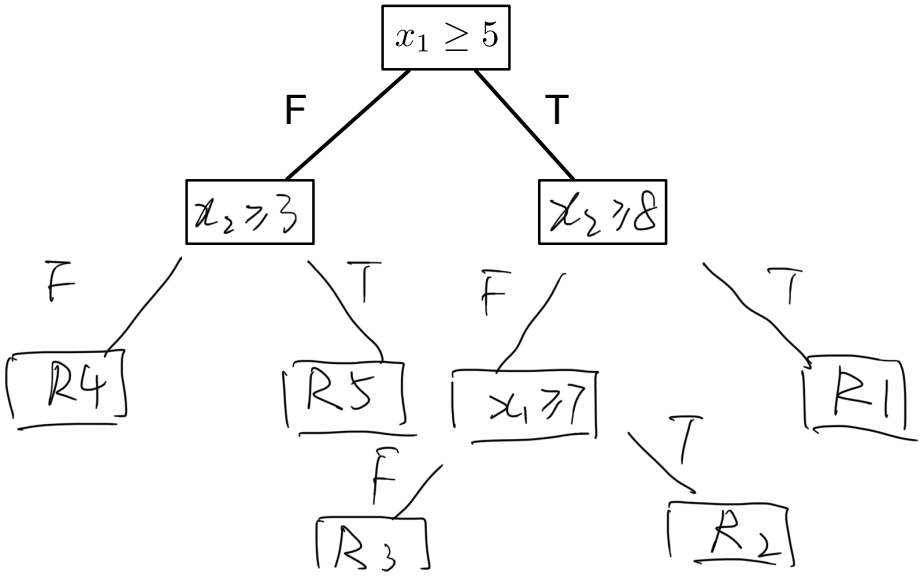


Figure 1: Training points consists of two classes of crosses (X) and circles (O), and five regions partitioning the input space.

Part (a) Design a decision tree that partitions the space into these regions. You should complete the following figure, which only has the root of the tree. The leaves of the tree should be one of the regions R1, R2, R3, R4, or R5. [4 MARKS]



R1	X
R2	X
R3	O
R4	X
R5	O

Table 1: Fill in with cross (X) or circle (O).

Part (b) What the predictions of the leaves (corresponding to regions R1, R2, ..., R5) should be in order to minimize the classification error over the training set? Fill Table 1 with crosses (X) and circles (O). [3 MARKS]

Part (c) Suppose that we constructed a tree with the partitions as in Figure 2. Given the training data points in each region, do you expect it to generalize better or worse compared to the tree in the previous part? Briefly justify your answer. [2 MARKS]

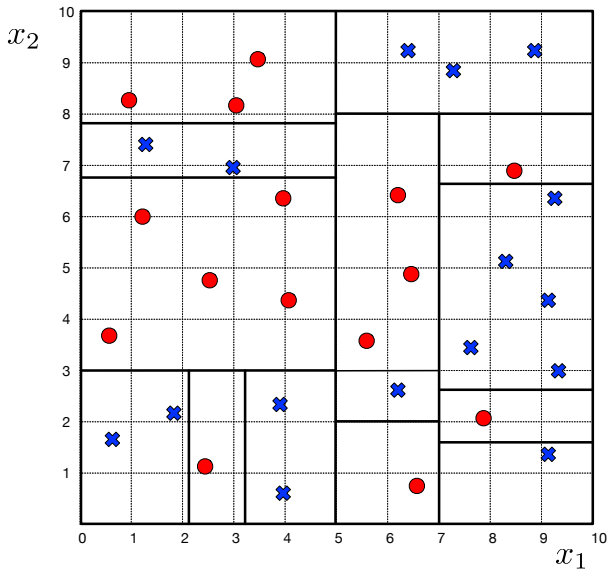


Figure 2: Training points consists of two classes of crosses (X) and circles (O), and fourteen (14) regions partitioning the input space.

It will generate worse, because it learn the training data so detail that the model is too complex to have a high generalization performance. This means it could be over-fitting and cannot have a good accuracy on unseen dataset.

 CONT'D...

Question 6. Multi-Class Classification for Regression Problems [5 MARKS]

Suppose that you have a regression problem with data $(\mathbf{x}, t)$ with the target values t being between $[0, 1]$. We usually solve this type of problems using one of a regression method by minimizing the squared error loss. But assume that you only need to predict t with the resolution of 0.1, e.g., it does not matter whether you predict 0.72 or 0.79; predicting 0.7 would be enough. Briefly describe how this problem can be formulated as a multi-class classification problem.

1° class → t

class	t
0	[0, 0.1)
1	[0.1, 0.2)
2	[0.2, 0.3)
3	[0.3, 0.4)
4	[0.4, 0.5)
5	[0.5, 0.6)
6	[0.6, 0.7)
7	[0.7, 0.8)
8	[0.8, 0.9)
9	[0.9, 1]

old t is a scalar

2° use one-hot encoding for each class. such as:

class 0 ⇒ $[1, 0, 0, \cdots, 0]^T$ (10 dim)

class 1 ⇒ $[0, 1, 0, \cdots, 0]^T$ (10 dim)

new t is a vector

3° Now the input $\mathbf{x}$ has d dimensions

the target t has 10 dimensions

then, the output y should also has 10 dimentions.

$$z_{d\times 1} = W_{d\times d}\, x_{d\times 1} + b_{d\times 1}$$

y have the same form with z ⟵ $y_{d\times 1} = \sigma(z)_{d\times 1}$

$$= \text{softmax}(z_1, z_2, \cdots, z_{10})$$

We use cross-entropy as the loss function

$$L_{CE} = -\sum_{k=1}^{10} t_k \log y_k$$

$$= -t^T (\log y)$$

where log is applied elementwise.

 OVER...

4° use SGD to update weight

5° class 0 represent 0, class 1 represent 0.1 and so on

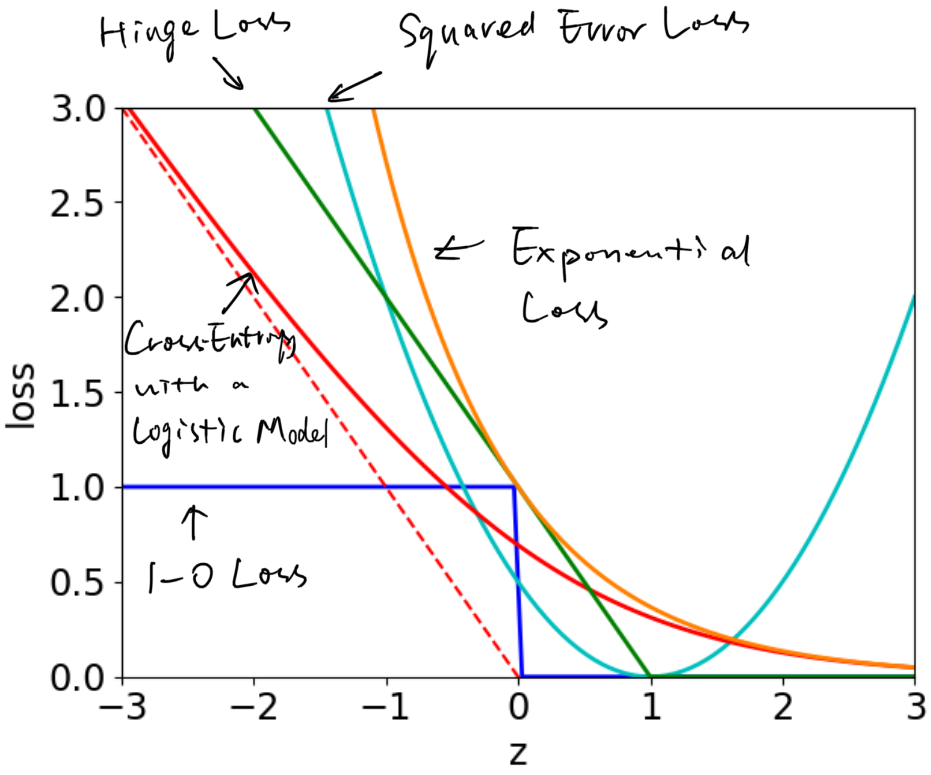
Question 7. Loss Functions for Classification [8 MARKS]

Consider a model whose output is z, e.g., the output of a linear classifier is $z = w^\top x + b$. We have been introduced to a variety of classification loss functions $\mathcal{L}(z, t)$, where t is the target (either $\{0, 1\}$ or $\{-1, +1\}$ for binary classification). This question asks you about these loss functions.

Part (a) Identifying Loss Functions [5 MARKS]

Identify the following loss functions on the figure (or write down their formulate, if you cannot write down on the figure). In this figure, you should assume that the target is $t = 1$. Make sure you identify them clearly without any ambiguity.

- $0 - 1$ Loss
- Exponential Loss
- Hinge Loss
- Cross-Entropy with a Logistic Model
- Squared Error Loss



Part (b) Why don't we minimize the $0-1$ loss function with a linear model? [1 MARK]

Because almost any point on it has 0 gradient, so we cannot minimize it with a linear model.

Part (c) What ML method uses the Hinge loss? [1 MARK]

SVM

Part (d) What ML method can be interpreted as using the exponential loss? [1 MARK]

Adaboost

Question 8. Optimization of Deep Linear Neural Networks [10 MARKS]

Consider the simplest deep linear neural network that is described by the following equations:

$$\therefore \quad \begin{cases} z = w_1 x, \\ y = w_2 z, \end{cases} \quad \Rightarrow \quad y = w_1 w_2 x$$

with $x, w_1, w_2, y \in \mathbb{R}$. This is indeed a very simple DNN that receives a one dimensional input, has one hidden layer with one unit, and one output. This simplicity is to ensure that the calculations are all easy. Consider the squared error loss function $l(y,t) = \frac{1}{2}(y-t)^2$.

Part (a) Show that one can replace this 2-layer NN with a 1-layer NN (show the relation of the input x to the output y). [2 MARKS]

$\therefore y = kx$, where $k = w_1 w_2$

Part (b) Compute the gradient of the loss of the 2-layer NN with respect to w_1 and w_2. [4 MARKS]

$$\frac{\partial l}{\partial w_2} = \frac{\partial l}{\partial y}\frac{\partial y}{\partial w_2} = (y-t)z$$

$$\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w_1} = (y-t)w_2 x$$

Part (c) Is the loss function of this 2-layer NN convex with respect to w_1 and w_2 or not? Prove your claim. [4 MARKS]

The loss function is not convex.

$$l(w_1, w_2) = \frac{1}{2}(w_1 w_2 x - t)^2$$

The Hessian matrix of the loss function is

$$H = \begin{bmatrix} \frac{\partial^2 l}{\partial w_1^2} & \frac{\partial^2 l}{\partial w_1 w_2} \\ \frac{\partial^2 l}{\partial w_2 w_1} & \frac{\partial^2 l}{\partial w_2^2} \end{bmatrix} = \begin{bmatrix} w_2^2 x^2 & 2w_1 w_2 x^2 - tx \\ 2w_1 w_2 x^2 - tx & w_1^2 x^2 \end{bmatrix}$$

Since when $w_1 = w_2 = 0$ we have $\frac{\partial l}{\partial w_1} = \frac{\partial l}{\partial w_2} = 0$

So $(w_1, w_2) = (0,0)$ is stagnation point

At that point:

$|H| = -3w_1^2 w_2^2 x^4 + 4 w_1 w_2 t x^3 - t^2 x^2$

$|H| = -t^2 x^2 \leq 0$ when $t \neq 0$, $x \neq 0$, we have $|H| < 0$, which mean the stagnation point $(0,0)$ is not the global minimum.

So, the loss function is not convex.

CONT'D...

Question 9. Bayesian Classifier [8 MARKS]

Consider a binary classification problem with input x being a scalar. The data generation process works as follows:

- First, a target t is sampled from $\{0, 1\}$ with equal probability.
- If $t = 0$, x is sampled from a uniform distribution over the interval $[0, 1]$.
- If $t = 1$, x is sampled from a uniform distribution over the interval $[0, 2]$.

Part (a) Write down the formula for $P(x|t = 0)$, $P(x|t = 1)$, $P(t = 0)$, and $P(t = 1)$. [4 MARKS]

$$P(x|t=0) = \int_0^x \frac{1}{1-0}\,dx = x$$

$$P(x|t=1) = \int_0^x \frac{1}{2-0}\,dx = \frac{1}{2}x$$

Since target t is sampled from $\{0, 1\}$ with equal probability, we have $P(t=0) = P(t=1)$

$\because P(t=0) + P(t=1) = 1 \quad \therefore P(t=0) = P(t=1) = \frac{1}{2}$

Part (b) Compute the posterior probability $P(t = 0|x)$ as a function of x. [4 MARKS]

$$P(t=0|x) = \frac{P(x|t=0)\,P(t=0)}{P(x)}$$

1° when $x \in [0,1]$,

$P(x|t=0) = x \quad P(t=0) = \frac{1}{2}$

$P(x) = P(x|t=0)P(t=0) + P(x|t=1)P(t=1)$

$= x \times \frac{1}{2} + \frac{1}{2}x \times \frac{1}{2}$

$= \frac{3}{4}x$

$\therefore P(t=0|x) = \frac{2}{3}$

2° when $x \in (1,2]$,

$P(x|t=0) = 0$

Therefore, we have

$$P(t=0|x) = \begin{cases} \frac{2}{3}, & x \in [0,1] \\ 0, & x \in (1,2] \end{cases}$$

$C_{1,2} = \log \frac{|\Sigma_2^{-1}|}{|\Sigma_1^{-1}|} + 2 \log \frac{P(t_1)}{P(t_2)}$

Question 10. Gaussian Discriminant Analysis [4 MARKS]

Consider a GDA model with two classes with covariance matrices Σ_1 and Σ_2.

Part (a) Write down the formula describing the decision boundary (you should be able to find this from the slides). [2 MARKS]

It is where $\log p(t_1 | x) = \log p(t_2 | x)$

$$(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + C_{1,2}$$

$$x^T \Sigma_1^{-1} x - 2\mu_1^T \Sigma_1^{-1} x = x^T \Sigma_2^{-1} x - 2\mu_2^T \Sigma_2^{-1} x + C_{1,2}$$

Part (b) If the covariance matrices are shared between two classes (i.e., $\Sigma_1 = \Sigma_2 = \Sigma$), mathematically show that the decision boundary is linear. [2 MARKS]

If $\Sigma_1 = \Sigma_2 = \Sigma$

we have,

$$x^T \Sigma^{-1} x - 2\mu_1^T \Sigma^{-1} x = x^T \Sigma^{-1} x - 2\mu_2^T \Sigma^{-1} x + C_{1,2}$$

$$2(\mu_2^T \Sigma^{-1} - \mu_1^T \Sigma^{-1}) x - C_{1,2} = 0$$

Therefore, the decision boundary is linear.

Question 11. Estimation Problems in Casino [15 MARKS]

You are at a casino and you decide to play a slot machine. The machine works as follows: On each round of the game, you pull an arm. The machine tells you whether you have won or lost. If you lose, it costs you \$1. If you win, you get a known value \$r, e.g., $r = 5$. You play with the machine until you win, and then it restarts, and then you play another round. In other words, each round consists of playing until a win.

The number of times that you have to pull an arm before winning (and finishing a round) is a random variable K that can take values of $0, 1, 2, \ldots$. For each round, you record this value. For example, if you play three rounds and get LLLLW (round 1), LLW (round 2), and W (round 3), you have $K_1 = 4$, $K_2 = 2$, $K_3 = 0$.

Let us model this process. The probability of winning at each arm pull is θ and the probability of losing is $1 - \theta$, for some unknown $\theta \in [0, 1]$ that depends on the machine. You can assume that the probability of winning at each arm pull is independent from each other and does not change as you play the game. With this assumption, the probability of k losses before winning has the following distribution:

$$P(K = k|\theta) = (1 - \theta)^k \theta, \qquad k = 0, 1, 2, \ldots .$$

As a perfectly rational person, deciding whether or not to play this game should depend on the expected money that you gain.[1] You can calculate the expected gain as follows (you do not need to understand this derivation completely to answer this question): If at one round you suffer k losses before winning, you have lost $k \times \$1$ and you gained $\$r$. This happens with probability $P(K = k|\theta)$. Therefore, your expected gain is

$$\begin{aligned}\text{expected gained money} &= \sum_{k\geq 0}(-k + r)P(K = k|\theta) = r\sum_{k\geq 0} P(K = k|\theta) - \sum_{k\geq 0} kP(K = k|\theta) \\ &= r - \frac{1 - \theta}{\theta}.\end{aligned}$$

So, if $r > \frac{1-\theta}{\theta}$, playing the game is worth it as the expected gain is positive; otherwise, it is not, and you better get out of the casino as soon as possible! Likewise, you can say that if $\theta > \frac{1}{1+r}$, the game is worth playing.

Since you do not know θ, it is crucial to estimate it in order to make an informed decision. This question asks you to develop some estimators for θ.

You play for N rounds and collect a dataset of $\mathcal{D}_N = \{K_1, K_2, \ldots, K_N\}$ describing the result of each rounds. For example, $\{4, 2, 0\}$ are the values for the example above. We assume that this data has already been collected, so you do not have to worry about your gain or loss so far. You can also assume that each round is independent from the previous ones (this is not an extra assumption, as it is implied by the independence of each arm pull).

[1] In reality, we are not rational agents, but that is another story.

Part (a) Likelihood [2 MARKS]

Write the likelihood function $P(\mathcal{D}_N|\theta)$ given a dataset $\mathcal{D}_N = \{K_1, K_2, \ldots, K_N\}$. It should be in the following form:

$$\theta^{\cdots\cdots} \times (1-\theta)^{\cdots\cdots},$$

each round is independent from the previous one

where $\cdots$ should be completed by you.

$\because P(K=k|\theta) = (1-\theta)^k \theta$

$\therefore P(\mathcal{D}_N|\theta) = P(k_1 k_2 \cdots k_N|\theta) = P(k_1|\theta)P(k_2|\theta)\cdots P(k_N|\theta)$

$= \theta^N (1-\theta)^{\sum_{i=1}^N k_i}$

Part (b) Log-likelikelihood [1 MARK]

Write the log-likelihood function $\ell(\theta) = \log P(\mathcal{D}_N|\theta)$ given a dataset $\mathcal{D}_N = \{K_1, K_2, \ldots, K_N\}$.

$$l(\theta) = \log P(\mathcal{D}_N|\theta) = \log\left(\theta^N (1-\theta)^{\sum_{i=1}^N k_i}\right)$$

$$= N\log\theta + \sum_{i=1}^N k_i \log(1-\theta)$$

Part (c) MLE [2 MARKS]

Find the Maximum Likelihood Estimator. You need to show your derivations in order to get any mark.

$$\frac{dl}{d\theta} = \frac{d}{d\theta}\left(N\log\theta + \sum_{i=1}^N k_i \log(1-\theta)\right)$$

$$= \frac{N}{\theta} - \frac{\sum_{i=1}^N k_i}{1-\theta}$$

Set this to zero, then we get:

$$\hat{\theta}_{ML} = \frac{N}{N + \sum_{i=1}^N k_i}$$

Part (d) Encoding Prior Belief [2 MARKS]

You are skeptical that a casino would setup a game such that you win money. You believe that they set their slot machine such that its θ is small enough to make you lose money in average. You can formulate this belief as a prior distribution on θ. Your skepticism can be expressed by stating that the prior probability that $\theta < \frac{1}{1+r}$ should be high.

As you are already familiar with the Beta distribution, you decide to use it as your prior. Recall that

$$\text{Beta}(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1},$$

where Γ is the gamma function. How do you reasonably encode your prior belief with a Beta distribution? You need to specify conditions on a and b. Note that there is no single correct answer. Specify a relation between a and b and briefly justify your answer.

$a = 1$, $b = r + 999$ $\Rightarrow$ $b = 999a + r$

Since the prior distribution follows the Beta distribution, we have $E(\theta) = \frac{a}{a+b}$, so, when $a = 1$, $b = r + 999$, $E(\theta) = \frac{1}{r+1000}$, which means $P(\theta < \frac{1}{1+r})$ is very high.

Part (e) MAP [2 MARKS]

Assume that you have selected a Beta distribution with a particular choice of a and b to encode your prior belief. Find the Maximum A-Posteriori (MAP) estimate. Your answer should be a function of $\mathcal{D} = \{K_1, K_2, \ldots, K_N\}$ and a and b. You should not use the particular a and b that you found in the previous question, but write it for any a and b.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D_N)$$

$$= \arg\max_{\theta} P(\theta) P(D_N \mid \theta)$$

$$= \arg\max_{\theta} \left[\theta^{a-1}(1-\theta)^{b-1}\right]\left[\theta^{N}(1-\theta)^{\sum_{i=1}^{N} k_i}\right]$$

$$= \arg\max_{\theta} \theta^{N+a-1}(1-\theta)^{\sum_{i=1}^{N} k_i + b - 1}$$

$$= \arg\max_{\theta} \underbrace{(N+a-1)\log\theta + \left(\sum_{i=1}^{N} k_i + b - 1\right)\log(1-\theta)}_{= L(\theta)}$$

Let $\frac{\partial L(\theta)}{\partial \theta} = 0$

$$\Rightarrow \frac{N+a-1}{\theta} - \frac{\sum_{i=1}^{N} k_i + b - 1}{1-\theta} = 0$$

We get $\hat{\theta}_{MAP} = \dfrac{N + a - 1}{N + \sum_{i=1}^{N} k_i + a + b - 2}$

Part (f) Bayesian Posterior [2 MARKS]

Calculate the Posterior probability of θ given the prior $\text{Beta}(\theta; a, b)$ and the data $\mathcal{D}_N = \{K_1, K_2, \ldots, K_N\}$. (Hint: Notice that the Beta distribution is a conjugate prior for this likelihood, so your posterior is in the form of a Beta distribution too.)

$$P(\theta|\mathcal{D}) \propto p(\theta)\, p(\mathcal{D}_N|\theta)$$
$$\propto [\theta^{a-1}(1-\theta)^{b-1}][\theta^N (1-\theta)^{\sum_{i=1}^N K_i}]$$
$$= \theta^{a-1+N}(1-\theta)^{b-1+\sum_{i=1}^N K_i}$$

Part (g) Bayesian Estimation of $\mathbb{E}[\theta]$ [1 MARK]

What is the expected value of the parameter θ according to the posterior distribution?

$$E(\theta|\mathcal{D}_N) = \frac{N + a}{N + \sum_{i=1}^N K_i + a + b}$$

Part (h) Comparison of MLE, MAP, and Bayesian Estimation [3 MARKS]

Briefly explain what the advantages and disadvantages of each of MLE, MAP, and Bayesian Estimation are (you should be able to answer this even if you have not calculated the estimators correctly.)

	advantages	disadvantages
MLE	1) it is easy to implement, because we can do gradient descent.	not suitable for small dataset
MAP	easy to calculate like MLE	only find one parameter instead of having a probability distribution
Bayesian Estimation	1) could deal with data sparsity better 2) having a probability distribution	1) it assumes that features are conditionally independent, (impractical in real world senario) 2) choosing prior is not easy

CONT'D...

Question 12. Course Evaluation [2 MARKS]

Please fill the course evaluation! I will read all your comments. It helps me improve the course in the future. Also feel free to send me an email if you want to provide more detailed comments.

If you did, please specify it here. You get the point if you have filled it.

Yes, I did it.

Use the space on this "blank" page for scratch work, or for any solution that did not fit elsewhere.
Clearly label each such solution with the appropriate question and part number.

Use the space on this "blank" page for scratch work, or for any solution that did not fit elsewhere.
Clearly label each such solution with the appropriate question and part number.