Index no. | | | | | | | |

UNIVERSITY OF MORATUWA, SRI LANKA
Faculty of Engineering
Department of Electronic & Telecommunication Engineering
B.Sc. Engineering
Semester 5 Quiz

EN 3150—Pattern Recognition

Time Allowed: 1 hour and 15 minutes Sep. 2023

ADDITIONAL MATERIAL

- Solution of ordinary least squares $\hat{\boldsymbol{w}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
- Scaling formulas
  Standard$(x_i) = \dfrac{x_i - \text{mean}(\boldsymbol{x})}{\text{std}(\boldsymbol{x})}$, Min-max $(x_i) = \dfrac{x_i - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$ and Max-abs $(x_i) = \dfrac{x_i}{\max(|\boldsymbol{x}|)}$.

INSTRUCTIONS TO CANDIDATES

- This quiz contains 15 questions on 5 pages.
- Answer **all** the questions.
- This quiz accounts for 15% of the module assessment. The total maximum mark attainable is 100. The marks assigned for each question & sections there of are indicated in square brackets.
- This is a closed-book examination.
- The symbols used in this paper have their usual meanings.
- Clearly state any assumptions that you may make.
- Neat and orderly presentation is important.
- Write your answers in the space provided.
- If you have any doubts as to the interpretation of the wording of a question, make your own decision, but clearly state it on the script.
- Electronic/Communication devices are not permitted. Only equipment allowed is a calculator approved and labeled by the Faculty of Engineering

Q1. Suppose an individual is scheduled to undergo a medical examination, and in the event of a positive outcome, the individual will be diagnosed with disease "A". The test's sensitivity is 90%, indicating that if the disease is present, there is a likelihood of 0.9 for the medical examination to produce a positive result. It is given that the probability of having the disease is 0.0001. Further, it is given that out of the 500 individuals who faced the aforementioned medical assessment, 50 yielded positive results without having the disease. If we denote an individual with the disease "A" as "$a = 1$" and a positive medical examination result as "$t = 1$".
What are the values of $p(t = 1|a = 1)$, $p(a = 1)$ and $p(t = 1|a = 0)$. Here, $p(x)$ indicates the probability of event x. **[05 marks]**

$p(t = 1|a = 1) = 0.90$, $p(a = 1)$=0.0001, and $p(t = 1|a = 0) = 50/500 = 0.1$.

Q2. Consider the information given in **Q1**. If the medical examination yields a positive result, what is the probability of actually having the disease ? **[10 marks]**

Using Bayes' theorem.

$$\begin{aligned} p(a = 1|t = 1) &= \frac{p(t = 1|a = 1)p(a = 1)}{p(t = 1)}, \\ &= \frac{p(t = 1|a = 1)p(a = 1)}{p(t = 1|a = 1)p(a = 1) + p(t = 1|a = 0)p(a = 0)}, \\ &= \frac{0.90 \times 0.0001}{0.90 \times 0.0001 + 0.1 \times (1 - 0.0001)} = 0.0009 = 0.09\%. \end{aligned}$$

Q3. Given a data set of observations $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)^T$ and two machine learning models namely model 1 and 2. The likelihood values for model 1 are expressed as $p(\boldsymbol{x}_1|\boldsymbol{\theta}_1) = 0.5$, $p(\boldsymbol{x}_2|\boldsymbol{\theta}_1) = 0.7$ and $p(\boldsymbol{x}_3|\boldsymbol{\theta}_1) = 0.4$, while for model 2 they are expressed as $p(\boldsymbol{x}_1|\boldsymbol{\theta}_2) = 0.6$, $p(\boldsymbol{x}_2|\boldsymbol{\theta}_2) = 0.5$ and $p(\boldsymbol{x}_3|\boldsymbol{\theta}_2) = 0.4$. Here, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represents the parameters of the model 1 and 2, respectively. Which model would be chosen based on the utilization of maximum likelihood analysis? What is the primary (main) assumption being used here? **[05 marks]**

In maximum likelihood analysis, the goal is to choose the model that maximizes the likelihood of the observed data.
For model 1

$$l_1 = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}_1) = 0.5 \times 0.7 \times 0.4 = 0.14.$$

For model 2

$$l_2 = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}_2) = 0.6 \times 0.5 \times 0.4 = 0.12.$$

Model 1 would be chosen because it has a higher joint likelihood for the given dataset.
The primary assumption being used here is that the observations are independent and identically distributed (i.i.d). This means that the likelihood of the entire dataset is the product of the individual likelihoods for each observation.

Q4. Linear regression is a unsupervised learning technique. Is this statement true? **[05 marks]**

False.

Q5. State three different variant of gradient decent algorithm. **[05 marks]**

(a) Batch Gradient Descent.

(b) Stochastic Gradient Descent.

(c) Mini-batch Gradient Descent.

Q6. Consider the linear regression model of $y(\boldsymbol{x}_i) = w_0 + w_1 x_{1,i} + w_2 x_{2,i}$ and loss function as $\sum_{i=1}^{N}(y_i - y(\boldsymbol{x_i}))^2$. Suppose that for the j-th iteration $w_0^{(j)}$, $w_1^{(j)}$ and $w_2^{(j)}$ are given by 1, 0.5, 1.5, respectively. Find the value of w_2 for (j+1)-th iteration using stochastic gradient descent (SDG) algorithm

with a learning rate of 0.1. Here, the data sample is used for the SGD is $y_i = 6$, $x_{1,i} = 2$, and $x_{2,i} = 2$, respectively. [**10 marks**]

$L(\boldsymbol{w}) = \sum_{i=1}^{N}(y_i - y(\boldsymbol{x_i}))^2 = \sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1,i} - w_2 x_{2,i})^2$.

$$\frac{\partial L(\boldsymbol{w})}{\partial w_2} = 2\sum_{i=1}^{N}(y_i - y(\boldsymbol{x_i}))(-x_{2,i}) = 2\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1,i} - w_2 x_{2,i}))(-x_{2,i}).$$

For one data sample

$$\frac{\partial L(\boldsymbol{w})}{\partial w_2} = -2(y_i - w_0 - w_1 x_{1,i} - w_2 x_{2,i}))(x_{2,i}) = -2(y_i - y(\boldsymbol{x_i}))x_{2,i}.$$

Gradient descent update

$$\begin{aligned} w_2^{(j+1)} &\leftarrow w_2^{(j)} - \alpha\frac{\partial L(\boldsymbol{w})}{\partial w_2}, \\ &\leftarrow w_2^{(j)} + \alpha \times 2(y_i - y(\boldsymbol{x_i}))x_{2,i}, \\ &\leftarrow 1.5 + (0.1) \times 2(6 - 1 - 0.5 \times 2 - 1.5 \times 2) \times 2, \\ &\leftarrow 1.9. \end{aligned}$$

The value of w_2 for (j+1)-th iteration is 1.9.

Q7. Consider the information given in **Q6**. Find the value of w_0 for (j+1)-th iteration using SDG. [**10 marks**]

$$\frac{\partial L(\boldsymbol{w})}{\partial w_0} = 2\sum_{i=1}^{N}(y_i - y(\boldsymbol{x_i}))(-1) = 2\sum_{i=1}^{N}(y_i - w_0 - w_1 x_{1,i} - w_2 x_{2,i}))(-1).$$

For one data sample

$$\frac{\partial L(\boldsymbol{w})}{\partial w_0} = -2(y_i - w_0 - w_1 x_{1,i} - w_2 x_{2,i})) = -2(y_i - y(\boldsymbol{x}_i)).$$

$$\begin{aligned} w_0^{(j+1)} &\leftarrow w_0^{(j)} + \alpha\ 2\ (y_i - y(\boldsymbol{x_i}))1, \\ &\leftarrow 1.0 + (0.1) \times 2\ (6 - 1 - 0.5 \times 2 - 1.5 \times 2), \\ &\leftarrow 1.2. \end{aligned}$$

The value of w_0 for (j+1)-th iteration is 1.2.

Q8. Consider the following data set ($\boldsymbol{X}$), which consists of five samples of ten distinct features. Out of these scaling methods (a) Standard scaling (b) Min-max scaling (c) Max-abs scaling, which scaling methods is appropriate for preserving the structure of this data set? [**05 marks**]

$$\boldsymbol{X} = \begin{bmatrix} 0 & 0 & 0.5 & 0 & 200 & 3 & 0 & 0.1 & 0 & 100 \\ 4 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.75 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 \\ 0.1 & 0 & 0 & 0 & 2 & 12 & 0 & 1000 & 1000 & 0 \end{bmatrix}$$

Max-abs scaling is suitable. Other two methods do not preserve the sparse nature of the data.

Q9. Is the statement "A simple model will have low bias and low variance, and a complex model will have high bias and high variance" true? [**05 marks**]

False. A simple model typically has high bias and low variance. A complex model often has low bias but high variance.

Q10. A data set of three data samples are given in table Q10. Suppose that linear regression model $f(\boldsymbol{x}) = w_0 + w_1 x_1 x_2$ fits to the given data set. Use all the data samples to find parameters of the linear regression model (w_0 and w_1). [**10 marks**]

Table Q10: Data set for **Q10**.

Sample index (i)	$x_{1,i}$	$x_{2,i}$	y_i
1	$\frac{1}{5}$	5	2
2	3	$\frac{2}{3}$	4
3	3	1	6

$f(\boldsymbol{x}) = y = w_0 + w_1 x_1 x_2 = w_0 + w_1 z$, where $z = x_1 x_2$. Now, this becomes a linear regression model.

$\hat{\boldsymbol{w}} = \left(\boldsymbol{Z}^T \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T \boldsymbol{y}$. Here, $\boldsymbol{Z} = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ 1 & z_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1}x_{2,1} \\ 1 & x_{1,2}x_{2,2} \\ 1 & x_{1,3}x_{2,3} \end{bmatrix}$.

$$\hat{\boldsymbol{w}} = \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

$w_0 = 0$ and $w_1 = 2$.

Q11. For this linear regression model $y(x) = w_0 + w_1 x_1 + w_2 x_2$, it is given that p-values for w_0, w_1, and w_2 are 0.3, 0.02, and 0.001, respectively. What can be say about the significance of these features (x_1,and x_2) to the dependent variable? State the reason for this selection. **[05 marks]**

p-values ≤ 5% is the common significance level. It says that the suggests that the feature is likely to have a significant impact on the target variable.
x_1 is statistically significant with a p-value of 0.02.
x_2 is highly statistically significant with a p-value of 0.001.

Q12. For the lasso regression minimizes the following loss function,

$$J(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2 + \lambda |\boldsymbol{w}|_1.$$

Fig. Q12 shows behavior of different components of the loss function with respect to different λ values. In this figure, minimum of each plot and respective co-ordinates are shown. Based on Fig. Q12, choose a value for λ and state reason for your selection. **[10 marks]**

Selecting the λ value that minimizes the loss is a valid selection (provides the best trade-off between bias and variance). Based on this criteria, λ is 1.96. (see curve $r + \lambda|\boldsymbol{w}|_1$).
There are other choices as well, if your primary goal is to obtain a sparse solution in Lasso regression, select the λ value that results in the lowest l_1-norm ($|\boldsymbol{w}|_1$) of the coefficients. In this case $\lambda = 10$. (see $|\boldsymbol{w}|_1$ curve). However, here you may get higher residual error value (r). If you need a balance between the sparsity and residual error, it is better to consider the both components r and $\lambda|\boldsymbol{w}|_1$.

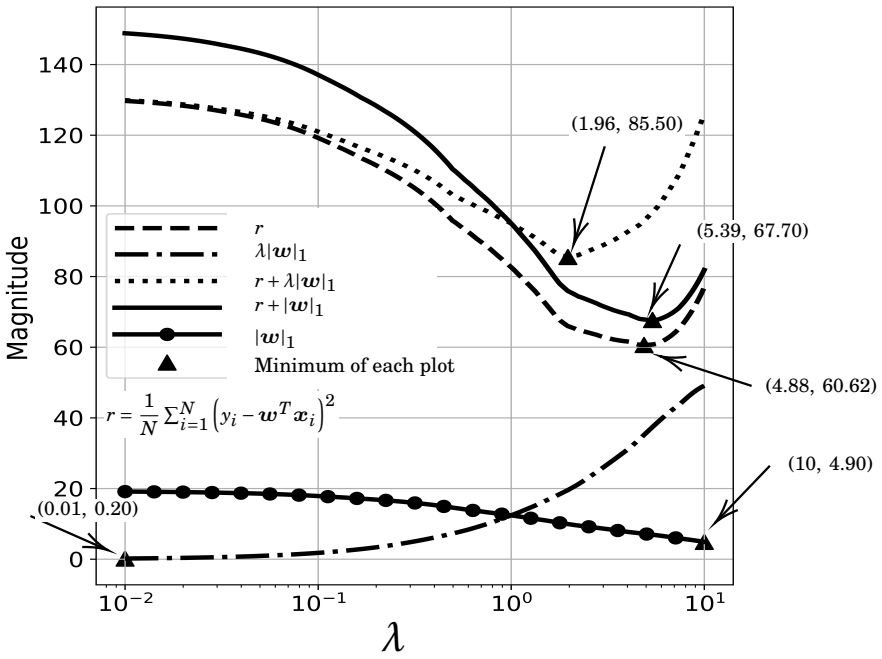

Figure Q12: Impact of λ.

Table Q13: Bias, variance, and irreducible error of two machine learning models.

Model	Bias	Variance	Irreducible error
Model 1	−0.5	0.5	0.1
Model 2	0.6	0.3	0.1

Q13. Table Q13 provides the bias, variance, and irreducible error of two machine learning models for a given data set. Based on this information, which model would you choose? State the reason for this selection. [**05 marks**]

MSE=bias^2+variance+Irreducible error.
For model 1: 0.85 and model 2: 0.76. So based on the MSEs, we have selected model 2, which has the lowest MSE.

Q14. Suppose you have a huge data set and you have time constrains. Here, which variation of the gradient descent algorithm would you select and what is the primary reason behind your selection? [**05 marks**]

If you have a large dataset and time constraints, Stochastic Gradient Descent (SGD) is a good choice. SGD is a variation of the gradient descent algorithm that uses a random subset of the training data to estimate the gradient at each iteration. This makes it computationally more efficient than the standard gradient descent algorithm, which uses the entire dataset to compute the gradient.

Q15. Suppose that both dependent and independent variables have measurement errors. Which algorithm would you choose between the ordinary least-squares and total least squares in this scenario? What is the main reason behind your decision? [**05 marks**]

In this scenario, the total least squares (TLS) method is a better choice. The main reason behind this decision is that ordinary least-squares assumes that only the dependent variable has measurement errors, while TLS accounts for errors in both dependent and independent variables.

The End.