*Total No. of Questions: 6* *Total No. of Printed Pages:3*

**Enrollment No.....**

# Faculty of Engineering

## End Sem (Odd) Examination Dec-2022

## CB3EL01 Machine Learning

Programme: B.Tech. Branch/Specialisation: CSBS

**Duration: 3 Hrs.** **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

Q.1 i. ML is a field of AI consisting of learning algorithms that? **1**

(a) Improve their performance
(b) At executing some task
(c) Over time with experience
(d) All of these

ii. In the mathematical equation of Linear Regression $Y = \beta1 + \beta2X + \epsilon$, ($\beta1$, $\beta2$) respectively refers to _____. **1**

(a) (X-intercept, Slope) (b) (Slope, X-Intercept)
(c) (Y-Intercept, Slope) (d) (slope, Y-Intercept)

iii. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters- **1**

(a) Complete dataset (b) Test dataset
(c) Validation dataset (d) None of these

iv. A _____ is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. **1**

(a) Dataset (b) Confusion matrix
(c) Classifier (d) None of these

v. Which of the following are true about bagging? **1**

(a) In bagging, we choose random subsamples of the input points with replacement
(b) If we use decision trees that have one sample point per leaf, bagging never gives lower training error than one ordinary decision tree
(c) Both options correct
(d) None of these

P.T.O.

vi. Select the correct statements related to Naïve Bayes (NB) classifier. **1**
(a) NB has low variance and high bias
(b) NB has low variance and low bias
(c) NB has high variance and low bias
(d) NB has high variance and high bias

vii. Select the incorrect statements related to "support vector machine" with "linear and nonlinear classification". **1**
(a) SVM can be used for linear classification
(b) SVM can efficiently perform non-linear classification
(c) SVM can not perform non-linear classification
(d) SVM uses kernel trick for non-linear classification

viii. How does the state of the process is described in HMM? **1**
(a) Literal
(b) Single random variable
(c) Single discrete random variable
(d) None of these

ix. Select the correct statements related to kNN classifier. **1**
(a) kNN has high variance and low bias
(b) kNN has high variance and high bias
(c) kNN has low variance and low bias
(d) kNN has low variance and high bias

x. Cluster is a collection of- **1**
(a) Similar types (b) Non-similar types
(c) Linear types (d) Non-linear types

Q.2 i. How does the learning rate affect the performance of a gradient descent algorithm? **4**

ii. Calculate the regression coefficient and obtain the lines of regression for the following data: **6**

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Y | 9 | 8 | 10 | 12 | 11 | 13 | 14 |

OR iii. Can we have overfitting and underfitting at the same time? Explain. How do you overcome overfitting and underfitting problems? **6**

Q.3 i. Discuss methods of evaluation for a machine learning algorithm. **4**

ii. In k-fold cross-validation, what happens when it is too big or too small? How should one decide what value of K to assume for the dataset? **6**



OR iii. Why feature engineering is important in model building and lists out some of the techniques used for feature engineering? **6**

Q.4 i. Explain naive bayes classifier with an example. **4**

ii. Discuss the major drawbacks of K-nearest neighbour learning algorithm and how it can be corrected. **6**

OR iii. Design decision tree for the given dataset showing each step properly: (Take **PROFIT** as the target feature(attribute)) **6**

| **AGE** | **COMPETITION** | **TYPE** | **PROFIT** |
|---|---|---|---|
| Old | Yes | S/W | Down |
| Old | No | S/W | Down |
| Old | No | H/W | Down |
| Mid | Yes | S/W | Down |
| Mid | Yes | H/W | Down |
| Mid | No | H/W | Up |
| Mid | No | S/W | Up |
| New | Yes | S/W | Up |
| New | No | H/W | Up |
| New | No | S/W | Up |

Q.5 i. Discuss association rule mining in detail. **4**

ii. Which problem can be solved with hidden Markov model? What are the limitations of HMM? **6**

OR iii. What are the applications of sequence classification? **6**

Q.6 Write short notes on any two:

i. Ward's algorithm **5**

ii. K-nearest neighbour clustering algorithm **5**

iii. Expectation-maximization algorithm **5**

******

| | | | |
|---|---|---|---|
| | | one decide what value of K to assume for the dataset? | |
| OR | iii | Why Feature Engineering is important in model building and lists out some of the techniques used for Feature Engineering? | 3+3 |
| | | | |
| Q.4 | i. | Explain Naive Bayes Classifier with an Example. | 4 |
| | ii. | Discuss the major drawbacks of K-nearest Neighbour learning Algorithm and how it can be corrected? | 3+3 |
| OR | iii | Design decision tree for the given dataset showing each step properly: (Take **PROFIT** as the target feature(attribute)) | 6 |

| AGE | COMPETITION | TYPE | PROFIT |
|---|---|---|---|
| Old | Yes | S/W | Down |
| Old | No | S/W | Down |
| Old | No | H/W | Down |
| Mid | Yes | S/W | Down |
| Mid | Yes | H/W | Down |
| Mid | No | H/W | Up |
| Mid | No | S/W | Up |
| New | Yes | S/W | Up |
| New | No | H/W | Up |
| New | No | S/W | Up |

| | | | |
|---|---|---|---|
| | | | |
| Q.5 | i. | Discuss association rule mining in detail. | 4 |
| | ii. | Which problem can be solved with hidden Markov model? What are the limitations of HMM? | 3+3 |
| OR | iii | What are the applications of sequence classification? Explain conditional random fields. | 3+3 |
| | | | |
| Q.6 | | Write short notes on any two: | |
| | i. | Ward's Algorithm | 5 |
| | ii. | K-nearest Neighbour Algorithm | 5 |
| | iii | Expectation-Maximization Algorithm | 5 |

******

Sol$^{n}_{2}$ 2.(ii)

| X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|
| 1 | 9 | 1 | 81 | 9 |
| 2 | 8 | 4 | 64 | 16 |
| 3 | 10 | 9 | 100 | 30 |
| 4 | 12 | 16 | 144 | 48 |
| 5 | 11 | 25 | 121 | 55 |
| 6 | 13 | 36 | 169 | 78 |
| 7 | 14 | 49 | 196 | 98 |
| $\Sigma$ 28 | 77 | 140 | 875 | 334 |

(2 Marks)

[Regression Coefficient at X on Y $b_{xy} = 0.929$

Regression Equation at X on Y $X = 0.929Y - 6.219$

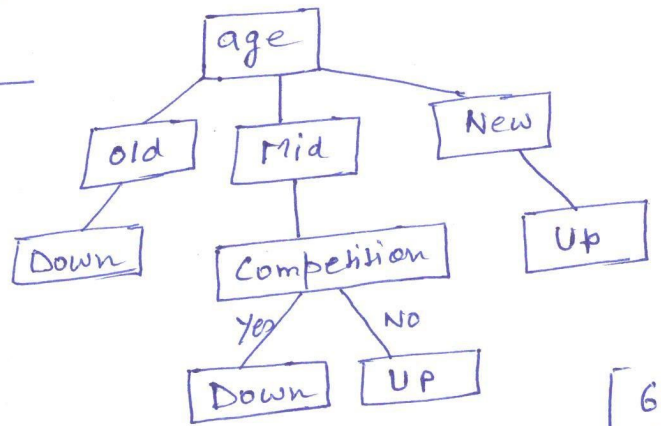
OR

Regression Coefficient at Y on X $b_{yx} = 0.929$

Regression Equation at Y on X $Y = 0.929X + 7.284$

[6 Marks]

Sol$^{n}_{2}$ 4.(iii)




[6 Marks]

# Scheme of Marking

| | |
|---|---|
| *Total No. of Questions: 6* | *Total No. of Printed Pages:2* |
| | **Enrollment No.....** |
| Faculty of Engineering<br>End Sem (Odd) Examination Dec-2022<br>Machine Learning-CB3EL01(T) | |
| Programme: B.Tech. | Branch/Specialisation:CSBS |

**Duration: 3 Hrs.** **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

| | | | |
|---|---|---|---|
| Q.1 | i) | ML is a field of AI consisting of learning algorithms that?<br><br>(a) Improve their performance (b) At executing some task<br>(c) Over time with experience **(d) All of the above** | 1 |
| | **ii)** | In the mathematical Equation of Linear Regression $Y = \beta1 + \beta2X + \epsilon$, ($\beta1$, $\beta2$) refers to _____<br><br>(a) (X-intercept, Slope) (b) (Slope, X-Intercept)<br>**(c) (Y-Intercept, Slope)** (d) (slope, Y-Intercept) | 1 |
| | **iii** | The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters<br><br>(a) Complete Dataset (b) Test Dataset<br>**(c) Validation Dataset** (d) None of these | 1 |
| | **iv** | A _____ is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.<br><br>(a) Dataset **(b) Confusion Matrix**<br>(c) Classifier (d) None of these | 1 |
| | **v)** | Which of the following are true about bagging?<br><br>(a) In bagging, we choose random subsamples of the input points with replacement<br>(b) If we use decision trees that have one sample point per leaf, bagging never gives lower training error than one ordinary decision tree<br>**(c) Both options correct** | 1 |
| | | (d) None of the above | |
| | **vi** | Select the correct statements related to Naïve Bayes (NB) Classifier<br><br>**(a) NB has low variance and high bias**<br>(b) NB has low variance and low bias<br>(c) NB has high variance and low bias<br>(d) NB has high variance and high bias | 1 |
| | **vii** | Select the incorrect statements related to "Support vector machine" with "Linear and Nonlinear classification"<br><br>(a) SVM can be used for linear classification<br>(b) SVM can efficiently perform non-linear classification<br>**(c) SVM can not perform non-linear classification**<br>(d) SVM uses kernel trick for non-linear classification | 1 |
| | **vii** | How does the state of the process is described in HMM?<br><br>a) Literal<br>b) Single random variable<br>**c) Single discrete random variable**<br>d) None of the mentioned | 1 |
| | **ix** | Select the correct statements related to kNN classifier.<br><br>**(a) kNN has high variance and low bias**<br>(b) kNN has high variance and high bias<br>(c) kNN has low variance and low bias<br>(d) kNN has low variance and high bias | 1 |
| | x) | Cluster is a collection of:<br><br>**(a) Similar Types** (b) Non-similar Types<br>(c) Linear Types (d) Non-linear Types | 1 |
| | | | |
| Q.2 | i. | How does the learning rate affect the performance of a Gradient Descent algorithm? | 4 |
| | ii. | Calculate the regression coefficient and obtain the lines of regression for the following data:<br><br>X: 1 2 3 4 5 6 7<br>Y: 9 8 10 12 11 13 14 | 6 |
| OR | iii | Can we have overfitting and underfitting at the same time? Explain. How do you overcome overfitting and underfitting problems? | 3+3 |
| | | | |
| | | | |
| Q.3 | i. | Discuss methods of evaluation for a machine learning algorithm. | 4 |
| | ii. | In k-fold cross-validation, what happens when it is too big or too small? How should | 3+3 |