Sessional-2 AML5102 Applied Machine Learning

Q1: [CO 2, BT 2] 10 marks 5 questions and 2 marks each.

1. Which of the following linkage methods in agglomerative clustering results in long chain? State reason in one sentence along with your choice

   (a) Ward linkage

   (b) Complete linkage

   (c) Centroid linkage

   (d) Single linkage

   (e) Average linkage

   Ans: (d) Simple Linkage. Clusters are connectivity by min distance between cluster members results in long chain

2. Which hierarchical clustering mechanism is top down? Why is it called top down? Write in 2 sentences (max)

   Ans: Divisive Clustering

3. Match the type of clustering (left) to an actual clustering algorithm (right). For e.g. if a. on the left corresponding to centroid based clustering matches with iv. on the right viz GMM Clustering, then write as a - iv and so on.

Type of clustering	Clustering algorithm
a. Centroid based clustering	i. DBSCAN
b. Distribution based clustering	ii. KMeans
c. Density based clustering	iii. Divisive clustering
d. Connectivity based clustering	iv. GMM clustering

   Ans: a ii, b iv, c i, d iii

4. Match the agglomerative linkage type (left) between to the formula (right). For e.g. if a. on the left corresponding to complete linkage matches with iv. on the right then write your answer as a - iv and so on.

   Notation convention:

   i. c_i and c_j are clusters and x_i and x_j are points belonging to clusters c_i and c_j respectively.

   ii. $|c_i|, |c_j|$ are number of entries in clusters c_i and c_j respectively.

Linkage type	Linkage formula
a. Complete linkage	i. $\mathcal{D}(c_i, c_j) = \min_{x_i \in c_i,\, x_j \in c_j} \|x_i - x_j\|_2$
b. Average linkage	ii. $\mathcal{D}(c_i, c_j) = \max_{x_i \in c_i,\, x_j \in c_j} \|x_i - x_j\|_2$
c. Centroid linkage	iii. $\mathcal{D}(c_i, c_j) = \frac{1}{\|c_i\|\|c_j\|} \sum_{x_i \in c_i} \sum_{x_j \in c_j} \|x_i - x_j\|_2$
d. Simple linkage	iv. $\mathcal{D}(c_i, c_j) = \left\| \left(\frac{1}{\|c_i\|} \sum_{x_i \in c_i} x_i\right) - \left(\frac{1}{\|c_j\|} \sum_{x_j \in c_j} x_j\right) \right\|_2$

   Ans: a ii, b iii, c iv, d i

5. Which of these regularization causes sparsity in coefficient vector w in linear regression $Xw = \hat{y}$? Select the **most correct** answer and give reason in 2 sentences (max)

   (a) L1 regularization
   (b) L2 regularization
   (c) All Lp regularization where $0 < p \leq 1$
   (d) Both L1 and L2 regularization
   (e) Elasticnet regularization
   (f) All of the above
   (g) None of the above

   Most correct answer: (c). Even though L1 regularization causes sparsity by selecting features and setting weights to be zero for non selected features, looking at the figure below, one can see the distance curve goes from convex to more and more concave as p reduces from zero to one and lesser. Hence those Lp norms with p less than or equal to one all result in sparse matrices.

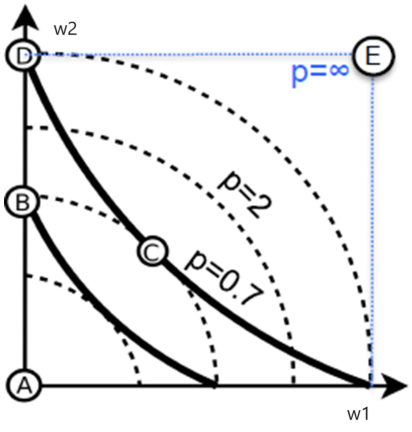

Figure 1: LpNorms

Q2: [CO 2, BT 3] 10 marks 4 questions.

1. (2 marks) The figure below shows four axis aligned numbered rectangular spaces for features X1 and X2. Draw a decision tree for the split shown.

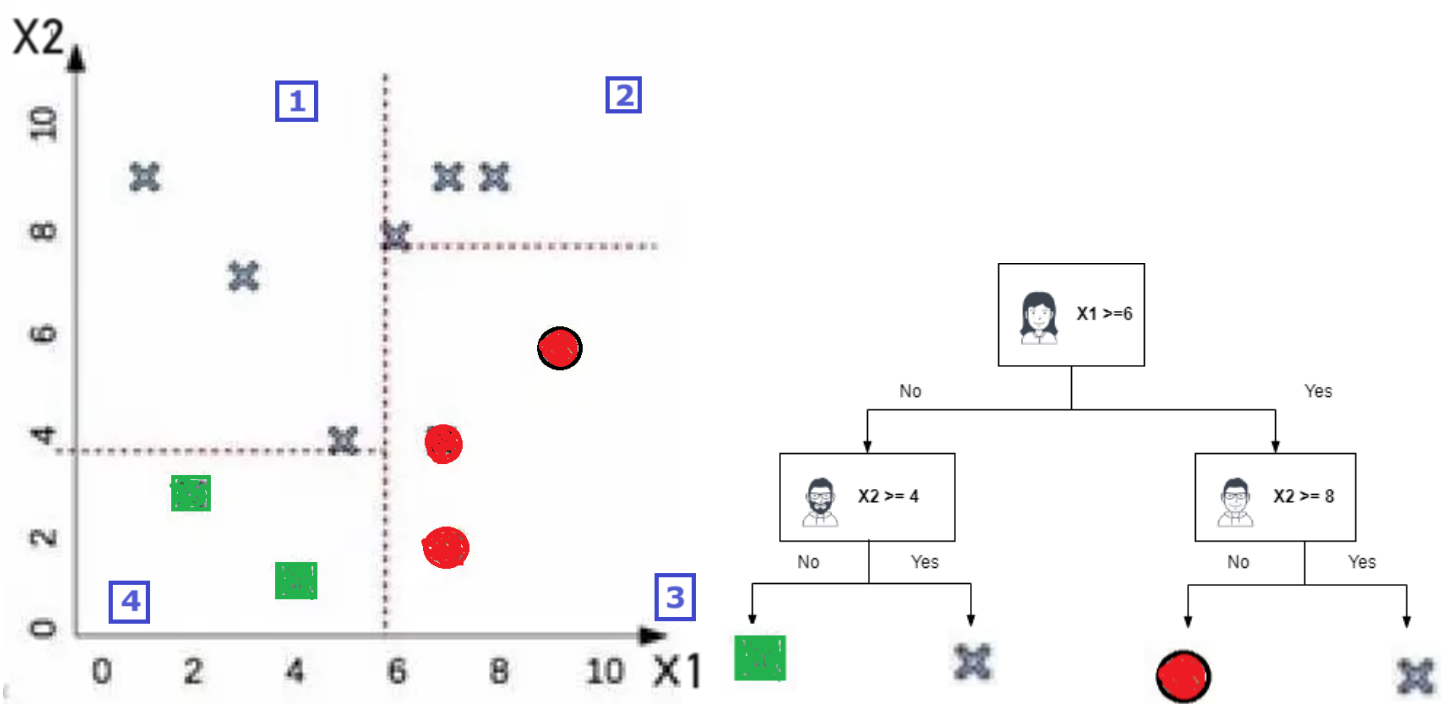

Figure 2: Axis aligned rectangular space and solution

2. (5 marks) The following dataset is given. Use Gini Impurity to select the feature for first split for the Decision Tree classifier. The dataset consists of 6 features - primary key, age, income, student and credit rate. Which feature did you select? Clearly show the calculation steps.

primary key	age	income	student	credit_rate	default
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
2	middle_age	high	no	fair	yes
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
6	middle_age	low	yes	excellent	yes
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
11	middle_age	medium	no	excellent	yes
12	middle_age	high	yes	fair	yes
13	senior	medium	no	excellent	no

Figure 3: Dataset with categorical features with loan default as target variable

3. (1 mark) The nodes struck off with red colored lines indicates post pruning of decision tree. Which of the four decision trees (a, b, c, d) in the diagram has higher accuracy when used for prediction with training dataset after training ? Give one sentence reason.

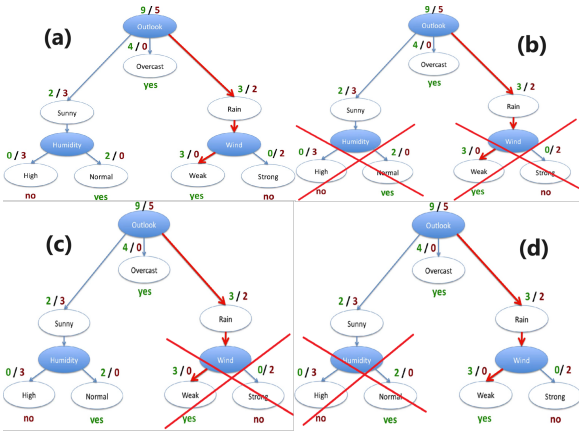

Figure 4: Decision trees pruned in different ways

Ans: (a)

4. (2 marks) Match the decision tree algorithms in the left to their corresponding formulae on the right. For e.g. if a. on the left corresponding to Gini Impurity matches with iv. on the right then write as a - iv and so on.

Notation convention:

i. j indicates a feature chosen at certain depth in the decision tree for doing split at the node.

ii. $|T|$ is the total number of nodes in decision tree. $|T_j|$ is the number of child nodes when the split is done on a feature j

Decision Tree algorithm	Loss function $\mathcal{L}$
a. Gini Impurity	i. $\max_j \sum_{j \in features} G(S)_j$
b. Gini Impurity with Cost complexity pruning	ii. $\frac{N_j}{N}\left(Gini_j - \mathbb{E}[Gini_{j-children}]\right)$
c. Gini Impurity adjusted for high cardinality	iii. $\alpha\|T\| + \sum_{j \in features} G(S)_j$
d. Node Importance	iv. $\alpha\|T\| + \sum_{j \in features} G(S)_j\|T_j\|$
	v. $\arg\min_j \sum_{j \in features} G(S)_j$
	vi. $\frac{\alpha\|T\| + \sum_{j \in features} G(S)_j}{\|T\|}$

Ans: a v, b iii, c iv, d ii

Q3: [CO 2, BT 3] 10 marks 5 questions.

1. (1 mark) True or False. A bagging ensemble is always a Random Forest. Give accurate reason for your choice in one sentence.

   ANs: False. Random Forest is a ensemble. But not all ensembles are random forests.

2. (1 mark) True or False. SMOTE under samples the majority class and over samples the minority class in the dataset.

   Ans: False. SMOTE only oversamples the minority class

3. (2 marks) Which of the following is correct about Random Forest when compared to a decision tree?
   A. Random Forest increases bias

   B. Random Forest decreases bias

   C. Random Forest does not impact bias

   D. Random Forest increases variance

   E. Random Forest decreases variance

   F. Random Forest does not impact variance

   (a) A and D are correct
   (b) A and E are correct
   (c) A and F are correct
   (d) B and D are correct
   (e) B and E are correct
   (f) B and F are correct
   (g) C and D are correct
   (h) C and E are correct
   (i) C and F are correct

   Ans: h

4. (2 marks) Explain the workings of SMOTE-Tomek Links algorithm with a diagram and 2-3 sentences max

5. (1 mark) Mutual information is easiest to calculate for which of the following

   (a) Feature is numerical and target is categorical
   (b) Both feature and target are numerical
   (c) Feature is categorical and target is numerical
   (d) Both feature and target are categorical

   Ans: d

6. (1 mark) Which of the following feature selection methods has feature selection process as part of the machine learning training process itself?

   (a) Filter methods
   (b) Embedded methods
   (c) Wrapper methods
   (d) All of the above

   Ans: b

7. (2 marks) Which of the following are **FALSE**?

   A. Correlated features can be independent

   B. Dependence between feature implies correlation

   C. Uncorrelated features are independent

   D. Independence of feature does not imply correlation

   E. Correlation of features implies dependence

   (a) A, B and C
   (b) B, C and E
   (c) C, D and E
   (d) A, C and D
   (e) A, B and D

   Ans: a

Q4: [CO 2, BT 4] 10 marks 5 questions 2 marks each

1. (2 marks) You developed a machine learning algorithm to detect a very rare disease. Presence of disease is 1, absence of disease is 0. You are given the choice of following metrics and you have to choose two of the most relevant - Precision, Recall, F-2, Accuracy, False Positive Rate.

   Which two metrics will you use and why? Give reasons for your choice without exceeding 1-2 short sentences

   (a) Precision and False Positive Rate (FPR)

   (b) Accuracy and Precision

   (c) F-2 and Accuracy

   (d) F-2 and Recall

   (e) Recall and False Positive Rate (FPR)

   (f) Precision and Recall

   Ans: d

2. (2 marks) Which are the two summary metrics used in ML classification and which of them will you choose for the above ML algorithm of detecting rare disease where presence of disease is 1, absence of disease is 0. State the reason for your choice in 2 sentences max.

   Ans: AOC and PRC. PRC will be used in the above scenario when the dataset is imbalanced

3. (2 marks) Why is F-1 score designed as harmonic mean of precision and recall instead of arithmetic mean of precision and recall ? Answer in 2 sentence max. Formula for F-1 is:

$$2 * \frac{precision \times recall}{precision + recall}$$

   Ans: because it penalizes the extreme value of either of precision and recall and will reduce the score, it follows that a higher F12 can be obtained only when none of them are sacrificed to be low to achieve the other value to be high

4. (2 marks) You developed a machine learning algorithm to classify chocolate muffins versus chihuahua dogs. Your ML algorithm predicts in terms of probability. Chocolate muffins are positive class and chihuahuas are negative class. Currently with the default threshold, your ML model predicts a lot of muffins as chihuahuas. What approach will you take to decrease muffins getting classified as chihuahuas? State your approach in 2 sentences (max) accompanied with necessary distribution diagram.

Figure 5: Binary classification of chocolate muffins and Chihuahua dogs

Ans: With reference to the distributions corresponding to 0 and 1 class, we need to reduce FN. FN can be reduced by reducing the threshold

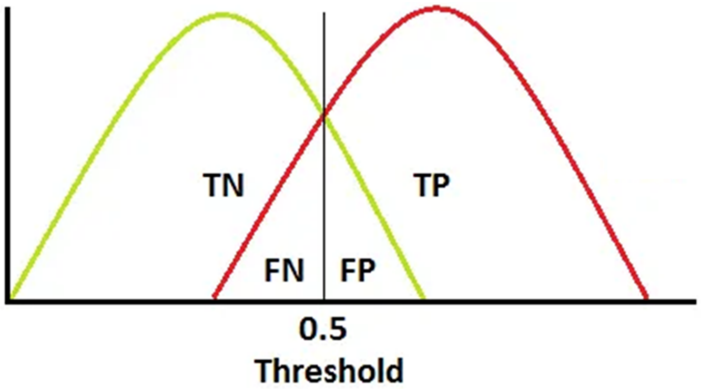

Figure 6: distributions for binary classification

5. The entropy venn diagram below has information for X and Y represented with two circles on left and right respectively. Use the entropy venn diagram to identify the regions corresponding to joint entropy H(X,Y), mutual information I(X,Y), conditional entropies $H(X|Y)$ and $H(Y|X)$. Copy the venn diagram to your answer four times and use each of the copied version to identify the 4 quantities by shading the areas - one shading in each diagram.

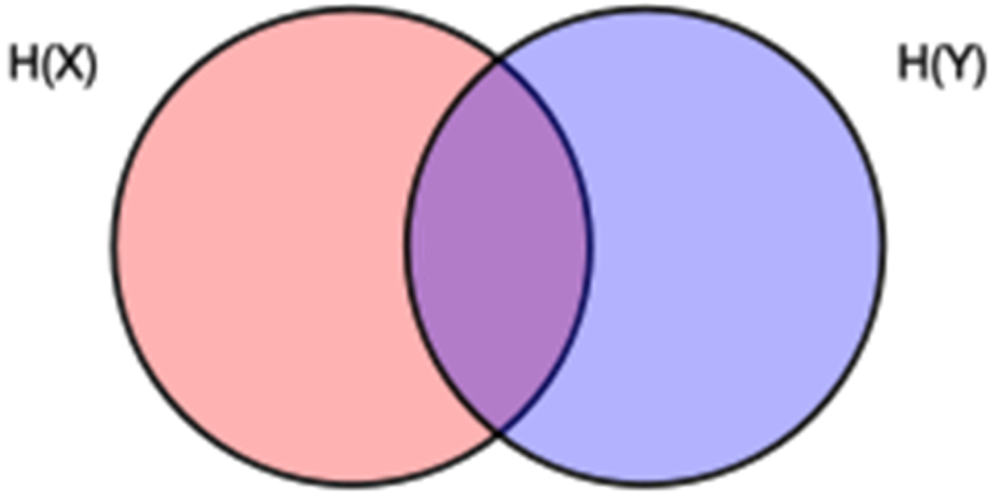

Figure 7: Entropy Venn diagram

Ans:

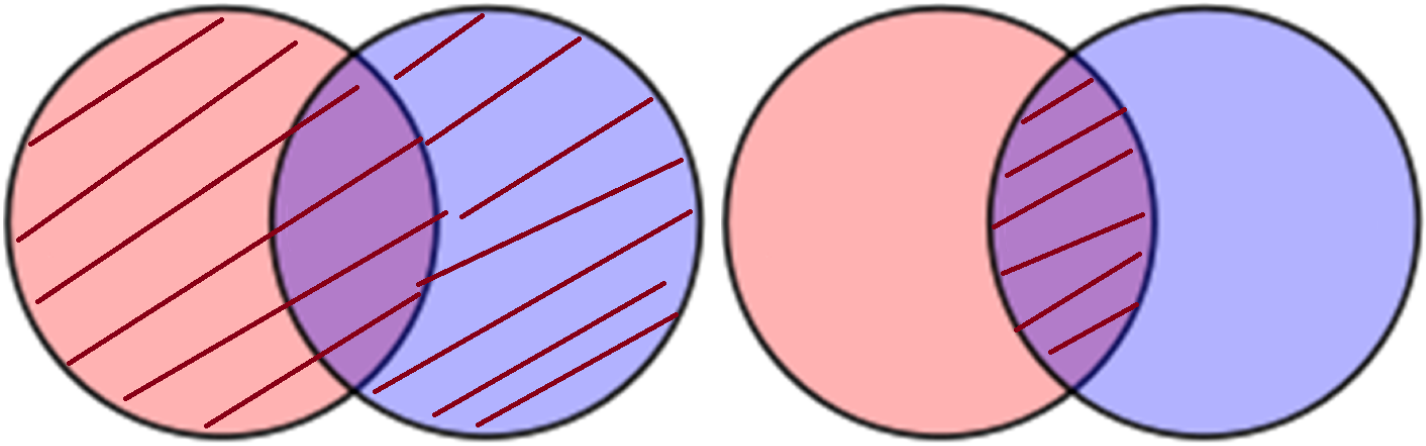

Figure 8: Joint Entropy and Mutual Information

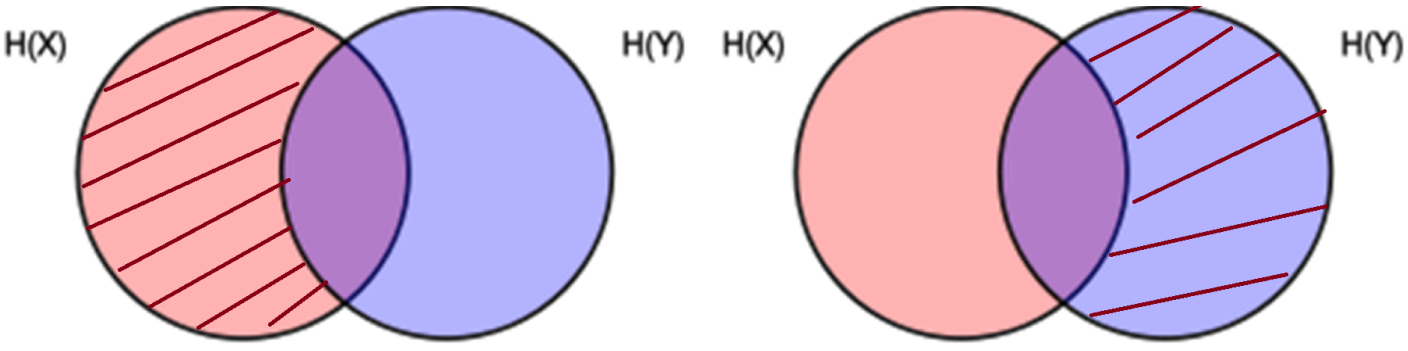

Figure 9: H(X—Y) and H(Y—X)

Q5: [CO 2, BT 3] 10 marks. 6 questions.

1. (2 marks) Which of the following represents the objective function for Linear Regression with L2 Regularization for a data set with m records and n features? θ represents the coefficient vector. β represents the hyperparameter for regularization.

   (a) $\mathcal{J}(x;\theta) = \frac{1}{m}\sum_{i=0}^{m-1}(\theta^T x^{(i)} - y^{(i)})^2 + \beta\sum_{i=1}^{n-1}\theta_i^2$

   (b) $\mathcal{J}(x;\theta) = \frac{1}{2m}\sum_{i=0}^{m-1}(\theta^T x^{(i)} - y^{(i)})^2 + \beta\sum_{j=0}^{n-1}\theta_j^2$

   (c) $\mathcal{J}(x;\theta) = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 + \beta\sum_{i=0}^{n}\theta_i^2$

   (d) $\mathcal{J}(x;\theta) = \frac{1}{2m}\sum_{i=0}^{m}(\theta^T x^{(i)} - y^{(i)})^2 + \beta\sum_{i=1}^{n}\theta_i^2$

   Hint: First let us rule out the incorrect answers. Since there are m records and n features, data matrix X has m rows and n+1 columns. This rules out choice d where the first summation has m+1 rows The L2 regularization penalty term should exclude the bias. (hence n terms). This rules out choice a and c

   We are left with only b. Hence b is the answer

2. (2 marks) Draw the coefficient vector θ and data matrix X for the dataset whose dimensions are specified in previous question. Use standard setup needed for linear regression

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ .. \\ \theta_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1^{(1)} & ... & x_n^{(1)} \\ 1 & x_1^{(2)} & ... & x_n^{(2)} \\ ... & & & \\ 1 & x_1^{(m)} & ... & x_n^{(m)} \end{bmatrix}$$

3. (1 mark) Which of the following is NOT a mechanism for detecting multi collinearity

   (a) VIF

   (b) Eigen decomposition

   (c) L2 regularization

   (d) Pairwise correlation heatmap

   (e) All of the above can be used for detecting multi collinearity

   Ans: c

4. (2 marks) What is heteroskedasticity? Answer in 1-2 sentence (max) and draw a diagram to illustrate the concept.

5. (1 mark) Linear Regression is

(a) Under deternmined system of equations

(b) Undetermined system of equations

(c) Over determined system of equations

(d) Exactly determined system of equations

Ans: (c)

6. (2 marks) Linear regression was applied to a dataset with m records and n features. The coefficients obtained after training is given by θ (excluding the intercept). The intercept is z. Which of the following correctly represents the Linear Regression objective function for above scenario?

(a) $\mathcal{J}_\theta(x) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y} - \theta^T x^{(i)} - z)^2$

(b) $\mathcal{J}_\theta(x) = \frac{1}{m}\sum_{i=0}^{m}(y - \theta^T x^{(i)} + z)^2$

(c) $\mathcal{J}_\theta(x) = \frac{1}{2m}\sum_{i=1}^{m}\left(y - (\theta^T x^{(i)} - z)\right)^2$

(d) $\mathcal{J}_\theta(x) = \frac{1}{m}\sum_{i=1}^{m}(\theta^T x^{(i)} - y)^2$

Ans: Since intercept is separate $\theta^T x$ does not account for it. Using

$$J = \frac{1}{something}\sum_{allrecords}(y - \hat{y})^2$$

and $\hat{y} = \theta^T x + z$ Substitute and Ans is a