Sessional-2 AML5102 Applied Machine Learning

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

A dataset with m records and n features is given. m is very large compared to n. A Linear Regression model is fitted with the equation $Xw = \hat{y}$ where X, w and $\hat{y}$ have standard meanings

1. (2 marks) What is dimension of X so that the equation $Xw = \hat{y}$ is mathematically valid?
2. (2 marks) What is the dimension of w so that the equation $Xw = \hat{y}$ is mathematically valid?
3. (2 mark) Data set has features x_1, x_2, ... x_n, Write the linear combination of all feature vectors in X with the weight coefficients in w
4. (2 marks) Linear combination of all feature vectors in X with the weight coefficients is located in ambient dimension of _____ but really located on a hyperplane of _____ dimension within that ambient space
5. (2 marks) What is the equation for the vector corresponding to dotted line in the following diagram?

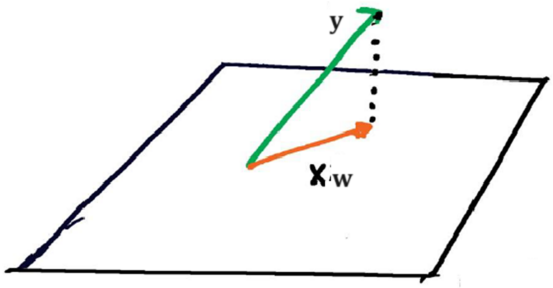

Figure 1: Linear Regression representation

Q2: [CO 2, BT 3] 10 marks 5 questions and 2 marks each.

1. What happens to the cluster size and granularity in the direction of arrow in hierarchical clustering

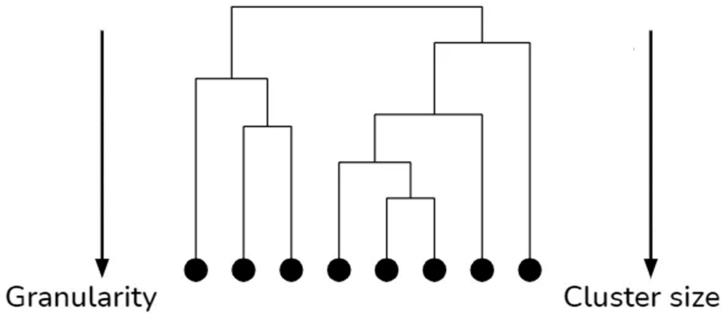

Figure 2: Linear Regression representation

   (a) Cluster size decreases and Granularity decreases
   (b) Cluster size decreases and Granularity increases
   (c) Cluster size increases and Granularity decreases
   (d) Cluster size increases and Granularity increases

2. Which of the statements is TRUE?

   (a) Divisive clustering is computationally efficient and has better accuracy than agglomerative clustering
   (b) Divisive clustering is computationally less efficient than agglomerative clustering but provides higher accuracy
   (c) Divisive clustering is neither computationally efficient nor provides better accuracy than agglomerative clustering
   (d) Divisive clustering is computationally efficient but agglomerative clustering has better accuracy

3. If the choice of agglomerative clustering is to choose merger of clusters based on minimization of their variance after merger, then the choice of linkage is

   (a) Ward linkage
   (b) Simple linkage
   (c) Average linkage
   (d) Complete linkage
   (e) Centroid linkage

4. 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

5. 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. $\mathcal{D}(c_i, c_j) = \left\| \left(\frac{1}{\|c_i\|} \sum_{x_i \in c_i} x_i\right) - \left(\frac{1}{\|c_j\|} \sum_{x_j \in c_j} x_j\right) \right\|_2$

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.

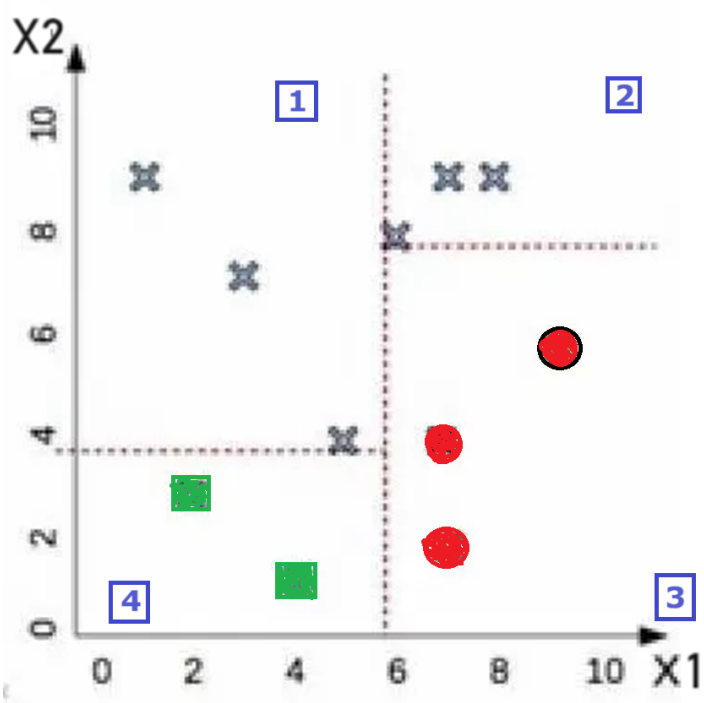

Figure 3: Axis aligned rectangular spaces

2. (4 marks) The following dataset is given. "Play Tennis" is the target variable. Others are features. What is the information gain if temperature was used for the first feature split? Clearly show the calculation steps.

Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure 4: Data set with categorical features

3. (2 marks) Why can't L1 or L2 regularization be used with Decision Trees to prevent overfitting? State your reason with 2-3 sentence max

4. (2 marks) Arrange these statements in the descending order of information of contained in them. Give your reasons preferably with mathematical equations and steps

   (a) New year is celebrated on January 1st

   (b) India will win ICC cricket world cup 2023

   (c) You will get Nobel Prize 2024 in Physics

Q3: [CO 2, BT 3] 10 marks 6 questions.

1. (1 mark) True or False. A boosting ensemble may take longer time to execute training than bagging ensemble. Give your reason for your choice in one sentence.

2. (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

3. (4 marks) What is bootstrapping? (1 sentence)
   What is role of bootstrapping in Random Forest? (2 sentence)
   What happens if bootstrapping is not performed in Random Forest? (2 sentence)

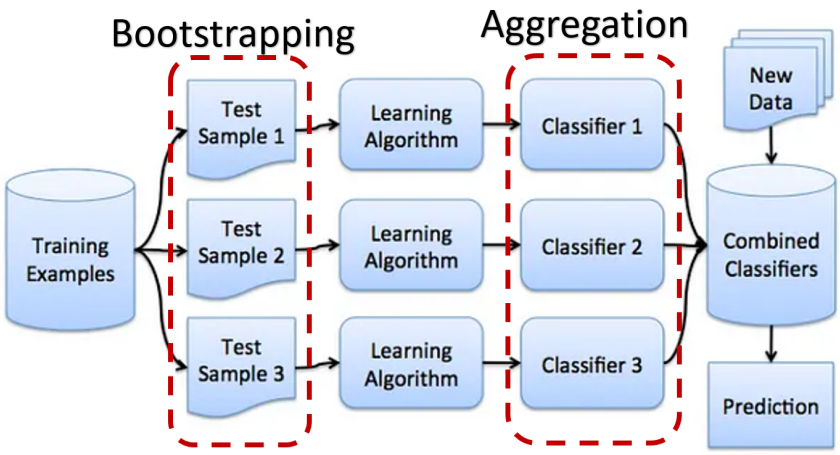

Figure 5: Bagging

4. (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

5. (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

6. (1 mark) Which of the following is not an advantage of Random Forest?

   (a) Feature correlation does not matter

   (b) Data need not be scaled

   (c) need not decide on parameters for the model

   (d) best feature among dependent features automatically gets highest feature importance

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

2. (2 marks) When and for what purpose will you use PR curve? Of the given curves, which is the best PR curve and why?

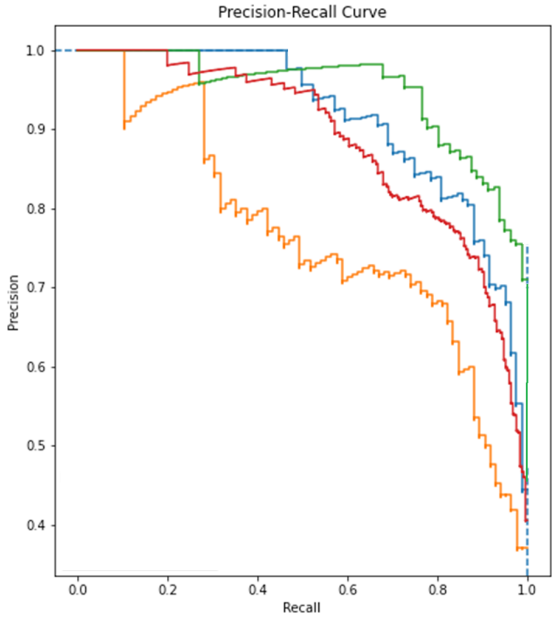

Figure 6: Bagging

3. (2 marks) Why is F-2 score a better metric than F-1 for evaluating model performance when the positive class is minority?

$$(1+\beta^2) * \frac{precision \times recall}{\beta^2 * precision + recall}$$

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 7: Binary classification of chocolate muffins and Chihuahua dogs

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.

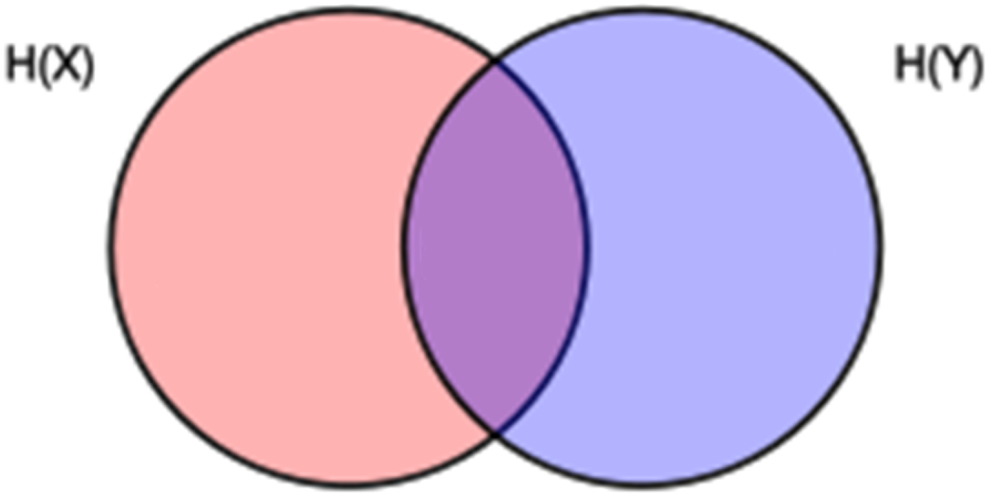

Figure 8: Entropy Venn diagram

Q5: [CO 2, BT 4] 10 marks. 5 questions.

1. (2 marks) Confusion matrix below with four cells of TP, TN, FP and FN. State each of the cells in terms of joint probabilities

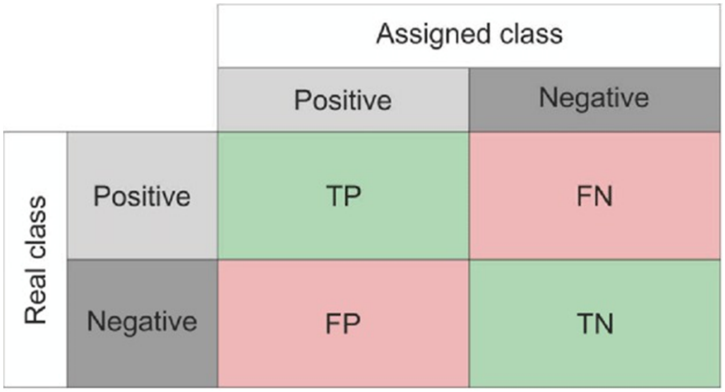

Figure 9: Entropy Venn diagram

2. (2 marks) Why is the sum of FP and FN always a constant for a given model? Answer in 2 sentences and a distribution diagram

3. (2 marks) Two features of a data set are correlated with correlation coefficient of 0.85. Which of the following are FALSE about Eigen decomposition of dataset. You can select multiple answers

   (a) Eigen decomposition cannot be performed on dataset with features with correlation greater than 0.80

   (b) Eigen decomposition can be performed but the eigen values are zero

   (c) Eigen decomposition cannot be performed because inverse cannot be calculated

   (d) Eigen decomposition can be performed on covariance matrix but not on the dataset

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

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