Question Paper

Exam Date & Time: 29-Nov-2023 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Manipal School of Information Sciences (MSIS), Manipal
First Semester Master of Engineering - ME (Artificial Intelligence and Machine Learning) Degree Examination - November / December 2023

Applied Machine Learning [AML 5102]

Marks: 100 **Duration: 180 mins.**

Wednesday, November 29, 2023

Answer all the questions.

1) **Q1: [CO 2, BT 2] 10 marks 5 questions, 2 marks each** (10)

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?

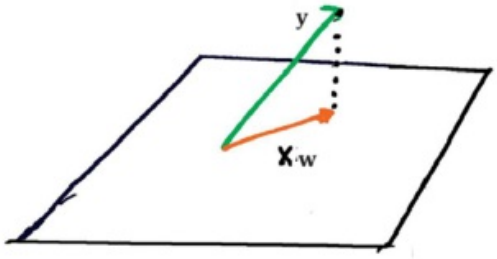


Figure 1: Linear Regression representation

2) **Q2: [CO 2, BT 3] 10 marks 4 questions and 2.5 marks each.** (10)

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

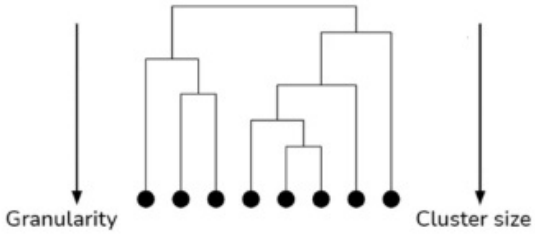


Figure 2: Linear Regression representation

(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

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

3) **Q3: [CO 1, BT 2] 10 marks 5 questions and 2 marks each.** (10)

1. Which of the following imputation is most appropriate for a categorical feature?
   (a) Mean Imputation
   (b) Grouped Mean Imputation
   (c) Median Imputation
   (d) Mode Imputation
2. Logistic Regression is used for
   (a) Binary classification
   (b) Multiclass classification
   (c) both
3. Log normal distributions are
   (a) Left skewed
   (b) heavily left skewed
   (c) heavily right skewed
   (d) symmetric
4. Which of the following is equivalent to Within Cluster Sum of Squares (WCSS) value for a given cluster?
   (a) Cluster centroid
   (b) Cluster variance
   (c) Cluster median
   (d) Cluster standard deviation
   (e) Cluster mean absolute deviation
5. A min max scaler is given by

$$\frac{(x - x_{min})}{(x_{max} - x_{min})}$$

What will be the range of this new feature?

4) **Q4: [CO 2, BT 2] 10 marks 4 questions** (10)

1. (1 mark) Which of these is least sensitive to outliers?
   (a) Mean
   (b) Median
   (c) Standard deviation
2. (5 marks) A toy dataset D = (-1, 3) (-1, 2), (1,4) (2,5) is provided. Assume k = 2 and perform KMeans clustering for 1 iteration using Expectation Maximization algorithm. Choose (-1,3) and (2, 5) as the initial random centroids.
3. (2 marks) A dataset has a column called "Country", is categorical feature and takes values India, Pakistan, Srilanka Which of the following encoding is best?
   (a) Label encoding
   (b) One Hot encoding
   (c) Ordinal encoding
   (d) Binary encoding
   (e) Factor encoding
4. (2 marks) For your chosen answer in the previous question, perform the selected encoding on the feature and show all possible values

5) **Q5: [CO 2, BT 2] 10 marks 5 questions** (10)

Answer the questions below based on the diagram the separating hyperplane and support vectors as shown

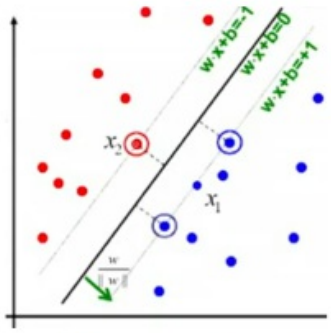
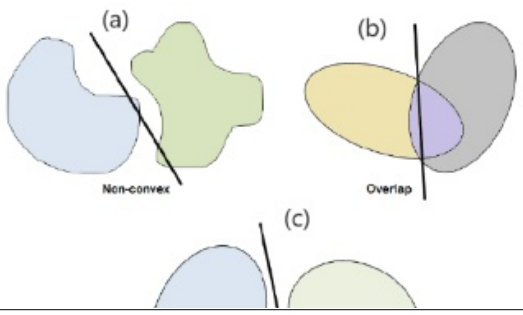


Figure 3: Support Vectors and separating hyperplane

1. (1.5 marks) How many extra support vectors are present in the diagram than necessary?
2. (1.5 marks) What is the distance between two support vectors on the opposing side of the separating hyperplane? Express your answer in terms of w.
3. (1 mark) Does the diagram show a hard margin SVM or soft margin SVM?
4. (2 marks) The diagram shows dataset with twenty data points. 10 are positive data points and 10 are negative. How many inequality constraints exist?
5. (4 marks) How many of the constraints are strict inequalities and how many are strict equalities for the SVM objective function? Give reason for your answer

6) **Q6: [CO 2, BT 2] 10 marks 4 questions** (10)

1. (2 marks) In which of the following cases does the Perceptron learning algorithm does not converge?



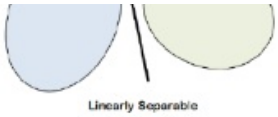


Figure 4: Datasets with different separability

2. (2 marks) In logistic regression, $w^T x + b$ is equal to

   (a) regression value

   (b) Sigmoid function

   (c) Log of odds

   (d) Odds of softmax

3. (4 marks) Figure out the mathematical equation of hinge loss function from the diagram below.

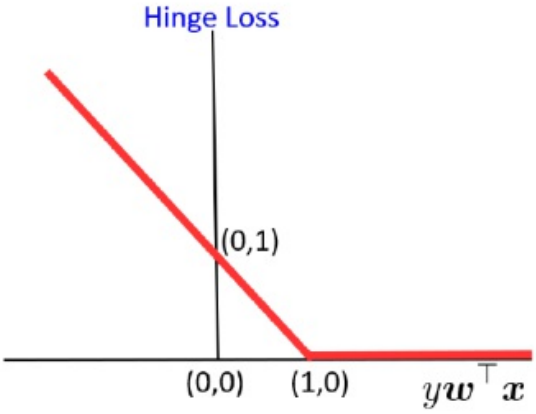


Figure 5: Hinge Loss

4. (2 marks) Which of the following scenario is best suited for using SVM in dual form?

   (a) A housing data set with 10K rows and 5 features to predict the price

   (b) A genetics data for classification using 100 rows and 10K gene expression features

(c) A wine dataset classification with 500 rows and 15 features

7) **Q7: [CO 2, BT 3] 10 marks 3 questions** (10)

1. (3 marks) The figure below shows four axis aligned numbered rectangular spaces for features X1 and X2. Draw a decision tree for the split shown.

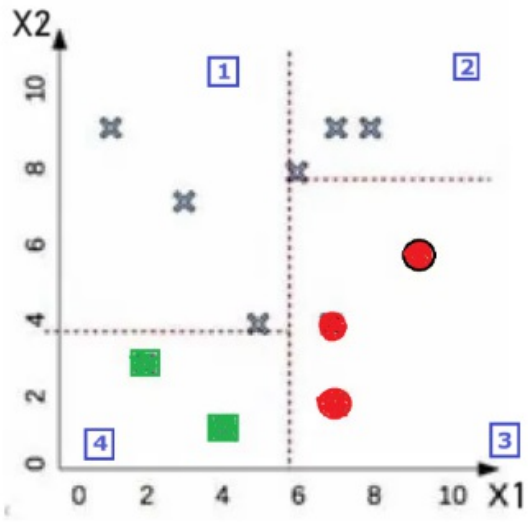


Figure 6: Axis aligned rectangular spaces

2. (6 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 5 features - primary key, age, income, student and credit rate. "default" is the target variable. 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 7: Dataset with categorical features with loan default as target variable

3. (1 mark) Why is k chosen as an odd number in kNN? (1 sentence answer)

8) **Q8: [CO 2, BT 3] 10 marks 5 questions.** (10)

1. (1.5 marks) True or False. A bagging ensemble is always a Random Forest. Give accurate reason for your choice in one sentence.

2. (1.5 marks) True or False. SMOTE under samples the majority class and over samples the minority class in the dataset.

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

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

5. (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

9) **Q9: [CO 3, BT 4] 10 marks 4 questions** (10)

1. (2 marks) 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

2. (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:

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

3. (3 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 8: Binary classification of chocolate muffins and Chihuahua dogs

4. (3 marks) kNN is trained for hyperparameters k = 3, 5, 7, distance = "manhattan", "euclidean" and weight = "uniform", "distance" How many times does the model get trained in total when GridSearch with KFold CV=3 is performed over the hyperparameters?

10) **Q10: [CO 1, BT 3] 10 marks. 5 questions** (10)

1. (3 marks) Two bivariate Gaussian distributions were fit for a dataset with two classes 1 and 2 for the target variable. The contour plots of the distributions and the decision boundary are as shown in the diagram below.

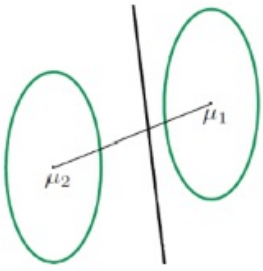


Figure 9: Two bivariate Gaussian distributions with decision boundary

Which of the conclusions are correct in addition to both distributions having same covariance matrices? Justify your reason along with your choice

(a) Same variance for 2 features & both classes are present in equal proportion in the dataset

(b) Same variance for 2 features but the proportion of records from class corresponding to μ_1 is more in the dataset

(c) Same variance for 2 features and the proportion of records from class corresponding to μ_2 is

(c) Same variance for 2 features and the proportion of records from class corresponding to μ_2 is more in the dataset

(d) Different variance for 2 features and the proportion of records from class corresponding to μ_2 is more in the dataset

(e) Different variance for 2 features and the proportion of records from class corresponding to μ_1 is more in the dataset

2. (3 marks)

   (a) (1 mark) Classify the three algorihms, Nearest Centroid, kNN and Kmeans clustering into categories Supervised and unsupervised

   (b) (2 marks) Fill the table below by categorizing the 3 algorithms into appropriate categories. Categories are specified as horizontal rows

Algorithm	Nearest Centroid	kNN	Gaussian Model
Eager vs Lazy	?	?	?
Batch vs Instance	?	?	?
Parametric vs Non-parametric	?	?	?
Discriminative vs Generative	?	?	?

3. (2 marks) What is heteroskedasticity? Answer in 1-2 sentence (max) and draw a diagram to illustrate the concept.

4. (1 mark) Linear Regression is

   (a) Under determmined system of equations

   (b) Undetermined system of equations

   (c) Over determined system of equations

   (d) Exaclty determined system of equations

5. (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

-----End-----