Appendix A

Solved University Question Papers of In Sem (March 2019) and End Sem (May 2019)

Machine Learning

March 2019

Q. 1(a) Define Machine learning and state two examples or applications of Machine Learning in our day to day lives. **(Section 1.2)** **(5 Marks)**

Q. 1(b) What do you mean by supervised and Unsupervised learning? Explain one example of each. **(Section 1.2)** **(5 Marks)**

OR

Q. 2(a) What is principal component Analysis (PCA), when it is used? **(Section 2.8)** **(5 Marks)**

Q. 2(b) What do you mean by dictionary learning? What are its applications? **(Section 2.12)** **(5 Marks)**

Ans. :

Applications of Dictionary Learning (a.k.a Sparse coding)

1. Signal processing/statistics
   - De-noising, edge detection, super resolution
   - Block compression for images/video
2. Machine Learning
   - Sparsity as a regularizer to prevent overfitting
   - Learned sparse representation play a key role in deep learning

Q. 3(a) Justify the statement: Raw data has a significant impact on feature engineering process. **(5 Marks)**

Ans. :

- Feature engineering is the process of converting raw data into features that can be used in building a model that can address the underlying problem and result into better accuracy on unseen data.
- Feature engineering is an important part in building any intelligent system.
- There are a wide variety of methods used in building intelligent systems, like machine learning, deep learning and others , however each problem is domain specific.
- In every domain , features play a significant role in system performance. These features are computed from raw data related to the problem and the domain. This is the reason often a data scientist spends 70% of their time in data preparation before the model building process.
- The process of feature engineering is time consuming and needs both domain knowledge and mathematical computations background.
- Features can be of two major types
  1. **Raw features :** Which are obtained directly from the domain specific data set with no extra tansformation.
  2. **Derived features :** These features are obtained by applying feature engineering from the existing data set.
- In the real world data can be of structured and unstructured format types.

Q. 3(b) Explain different mechanisms for managing missing features in a dataset? **(Section 2.5)** **(5 Marks)**

OR

Q. 4(a) With reference to feature engineering, explain data scaling and normalization tasks? **(Sections 2.1 and 2.6)** **(5 Marks)**

Q. 4(b) What are the criteria or methodology for creation of Training and Testing data sets in machine learning methods? **(Section 2.3)** **(5 Marks)**

Q. 5(a) What do you mean by a linear regression? Which applications are best modeled by linear regression? **(Section 3.2)** **(5 Marks)**

Ans. :

Applications modeled by Linear Regression

1. Gaining insights into consumer behaviour, understanding how profitability is influenced by business and other factors.
2. It can be used for trend evaluation and estimating forecasts.
3. It can also be used to check product sale is influenced by marketing effectiveness , pricing and promotions.
4. It is helpful in assessing risk in financial and insurance domain.

Q. 5(b) Write a short note on Types of regression. **(Sections 3.5, 3.5.1 and 3.5.2)** **(5 Marks)**

OR

Q. 6(a) Write short note on Linearly and non-linearly separable data. **(5 Marks)**

Ans. :

It is an important concept in Neural Network. If it is possible to separate points in N-dimensional space using only $N - 1$ dimensions then we say the points are linearly separable else they are non linearly separable points.

1. **One Dimension**

   Let us consider one dimension scale. As shown in Fig. 1-Q. 6(a), there are two possibilities as follows :

   – One you may choose two different numbers, in this case you can always find another number in between them, hence the numbers you chose are linearly separable.

   – The second option, you may choose the same number twice, hence it wont be possible to choose a number in between them as both are the same numbers, in such a case we may call the numbers are non linearly separable.

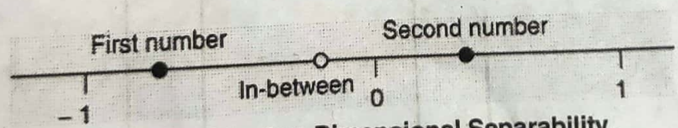

Fig. 1-Q. 6(a) : One Dimensional Separability

2. **Two dimension**

   – We can extend the above concept to two dimension to explore other possibilities. Consider the Fig. 2-Q. 6(a).

   – In the Fig. 2-Q. 6(a) we would like to separate (1,1) from the other two points.

   – Here we can see that there are infinite such possibilities which can separate the point (1,1) from other two points.

   – So the two classes, class A : (1, 1) and the class B: (0, 0), (0, 1) and (1, 0) are linearly separable.

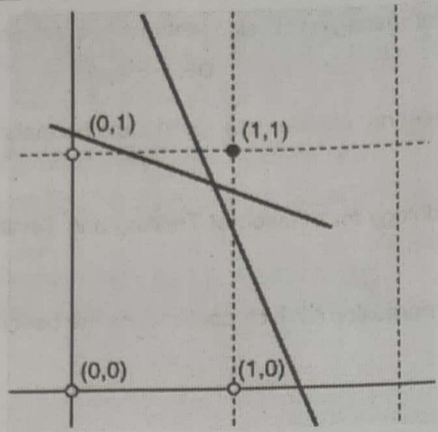

Fig. 2-Q. 6(a)

- Now consider another two dimensional case as shown in Fig. 3-Q. 6(a).
- In the Fig. 3-Q. 6(a) we cannot separate the two classes, one containing dark points and other containing white points.

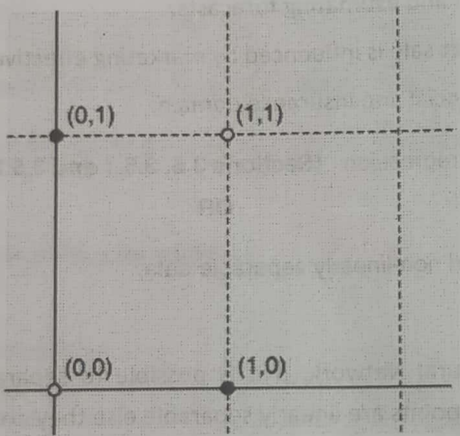

Fig. 3-Q. 6(a)

3. **Three dimension**
   - The concept can be extended to three dimension as shown in the Fig. 4-Q. 6(a)

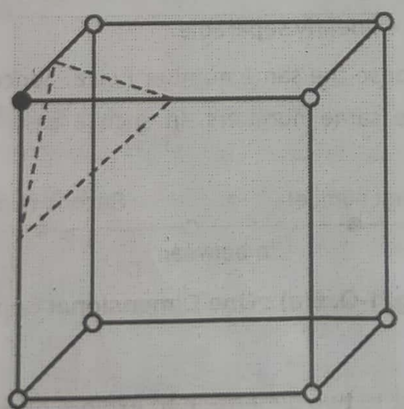

Fig. 4-Q. 6(a) : 3D

- We can see that the two classes, one dark point and rest white points are linearly separable using a plane.
- The concept can be extended to N-dimension and use a N-1 dimensional plane to separate the data points.

Q. 6(b) Write short note on Classification techniques. **(5 Marks)**

Ans. :

Basic Concept : Classification

- Classification constructs the classification model based on training data set and using that model classifies the new data.
- It predicts the value of classifying attribute or class label.

Typical applications

- Classify credit approval based on customer data.
- Target marketing of product.
- Medical diagnosis based on symptoms of patient.
- Treatment effectiveness analysis of patient based on their treatment given.

Various classification techniques

- Linear classifiers: Logistic regression , Naïve bayes classifiers
- Nearest Neighbor
- Support vector machines
- Decision trees
- Boosted trees
- Random forest
- Neural networks

1. **Naïve Bayes classifier**
   - This classification technique is based on Bayes theorem.
   - This method makes assumption that the presence of one feature in a class is unrelated to presence of any other feature.
   - This model is easy to build.
   - It is helpful for large data sets.
2. **Nearest Neighbor**
   - This method takes a set of labelled data points and makes use of these to label other data points.
   - To label a data point it takes into account the closest labelled data point(nearest neighbors).
   - K represents the number of neighbors it checks in the above process of labelling.
3. **Logistic Regression**
   - This method is a statistical method.
   - In this method there are one or more independent variables that are used to determine the outcome.
   - The outcome is a dichotomous variable (two possible outcomes).
   - Goal of the method is to find the best fitting model to describe the relationship between dependent variable and a set of independent variables.

4. Decision trees

- This method builds a classification or regression model in the form of a tree.
- The entire data set is broken down into smaller data sets and the decision tree is built incrementally.
- The final tree obtained has decision nodes and leaf nodes.
- The decision node has two or more branches and the leaf node represents the decision.
- The topmost root node is known as the best predictor variable.

5. Random Forest

- This method is an ensemble learning method for classification, regression or other tasks.
- The method builds a multitude of decision trees during training and output the classification or mean prediction.

6. Neural Network

- A neural network is a network consisting of neurons arranged in layers.
- This network converts the input into some output.
- Each unit takes a input and applies a function, mostly non linear and passes them to the next layer.
- The network can be of two types: feed forward or backward propagation network.

Q. 6(c) Write short note on ROC curve **(Section 3.15)** **(5 Marks)**

May 2019

Q. 1(a) With reference to machine learning, explain the concept of adaptive machines. **(Section 1.1)** **(6 Marks)**

Q. 1(b) Explain the role of machine learning algorithms in following applications. **(6 Marks)**

(a) Spam filtering. **(Section 1.2.1)**

(b) Natural Language processing. **(Section 1.2.2)**

OR

Q. 2(a) Explain Data formats for supervised learning problem with example. **(Section 2.4)** **(6 Marks)**

Q. 2(b) What is categorical data? What is its significance in classification problems? **(Section 2.4)** **(6 Marks)**

Ans. :

Significance of categorical data in Classification Problems

- Any categorical data represents discrete values belonging to a specific finite set of categories or classes.
- These are also called as classes or labels, which are predicted by a machine learning model in classification problem.
- These discrete values can be text or numeric, categorical data have two major classes nominal and ordinal.
- Nominal values have no ordering among the values of those attributes.
- Whereas ordinal has some kind of order among its values.
- So categorical attributes plays its major role in representing classes or labels.

Q. 2(c) Explain the Lasso and ElasticNet types of regression. **(Sections 3.5, 3.5.1 and 3.5.2)** **(8 Marks)**

Q. 3(a) What problems are faced by SVM when used with real datasets? **(Section 4.8.2)** **(3 Marks)**

Q. 3(b) Explain the non-linear SVM with example. (Sections 4.10.2 and 4.10.3) (5 Marks)

Q. 3(c)(i) Write short notes on Bernoulli naive Bayes. (Section 4.4) (3 Marks)

Q. 3(c)(ii) Write short notes on multinomial naive Bayes. (Section 4.5) (3 Marks)

Q. 3(c)(iii) Write short notes on Gaussian naive Bayes. (Section 4.6) (3 Marks)

OR

Q. 4(a) Define Bayes Theorem. Elaborate Naive Bayes Classifier working with example. (Sections 4.1.1, 4.2 and 4.2.1) (8 Marks)

Q. 4(b) What are Linear support vector machines? Explain with example. (Section 4.9) (4 Marks)

Q. 4(c) Explain with example the variant of SVM, the Support vector regression. (Section 4.12) (5 Marks)

Q. 5(a) Explain the structure of binary decision tree for a sequential decision process. (8 Marks)

Ans. :

Structure of a normal decision tree for sequential decision process

- The structure of a binary decision tree is based on sequential decision process.
- The topmost node is called the root and it is the starting point of the decision tree.
- At the root a feature is evaluated and based on the value one of the two branches is selected.
- This process is repeated until a leaf node is reached.
- The leaf node represents the final classification target we have been looking for.
- One of the earliest formulation of decision tree called as Iterative Dichotomizer(ID3) required categorical features.
- This categorical feature requirement in ID3 limited its usage and C4.5 was invented.
- C4.5 managed continuous (binned and discretized) values.
- C4.5 also has the ability to transform a tree into a sequence of conditional expressions

  (if<condition> then <...> else<...>)
- When compared to other classification algorithms, decision trees seemed to simple.
- If the dataset is splittable , the overall process is rather fast in predictions.
- It also works efficiently with unnormalized data sets, as their internal structure is not influenced by values assumed by each feature.
- However decision tress are highly sensitive to unbalanced classes and lead to poor accuracy when a class is dominant, to overcome this one of the resampling methods or class weight parameter may be used.

Binary Decisions

Consider an input dataset, X

$X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\}$ where $\bar{x}_i \in \mathbb{R}^m$

In the representation every vector is made of m features, and each of them can be a node based on a tuple (feature, threshold)

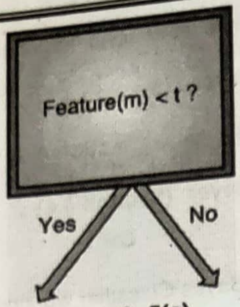

Fig. 1-Q. 5(a)

Single Splitting Node

- The feature that best separates your data should be picked as these condition would guarantee the convergence of the tree which will have minimum uncertainty.
- Let us consider following is the selection tuple $\sigma = [i, t_k]$

  i : Represents the index of the feature at a certain node, and

  t_k : Represents the threshold that determines left and right branch selection
- Overall goal is to reduce the residual impurity in the least number of splits, so that a very short path exists between the sample data and classification result
- Total Impurity measure can be measured by considering two branches as follows :

$$I(D, \sigma) = \frac{N_{left}}{N_D} I(D_{left}) + \frac{N_{right}}{N_D} I(D_{right})$$

Where, **D** : Whole data set, **I** : Impurity measure and

D_{left} and D_{right} : Resulting subsets

Impurity measures

Following are some of the impurity measures that may be used

1. Gini impurity index

$$I_{Gini}(j) = \sum_i p(y = i \mid Node = j)\,(1 - p(y = i \mid Node = j))$$

2. Cross-entropy impurity index

$$I_{Cross-Entropy}(j) = -\sum_i p(y = i \mid Node = j) \log(p(y = i \mid Node = j))$$

3. Misclassification impurity index

$$I_{Missclassification}(j) = 1 - \max p(y = i \mid Node = j)$$

Q. 5(b) With reference to Clustering, explain the issue of "Optimization of clusters" (5 Marks)

Ans. :

The potential problem are the challenges faced in choosing the following parameter with respect to clustering techniques

1. **Identification of distance measure**

For numerical attributes, different cases of minkowski distance measure may be used , but incase of categorical attributes it is difficult to identify a measure.

2. **Number of clusters**

- Determining the number of clusters is a difficult task if the number of class labels are not known before hand.
- Choosing the right number of clusters is important for producing correct results.

3. **Lack of class labels**

For real time data sets , distribution of data has to be done to understand the class labels

4. **Structure of database**

- Real time datasets may have data from which it is difficult to identify the number of clusters.
- There may be issues with missing values in the tuples.

5. **Types of attributes**

- The datasets may not always contain numerical and categorical attributes.
- It may contain other types like nominal, ordinal, binary.
- Here conversion of data types will be needed.

6. **Choosing the initial clusters**

- Partition based methods needs initial k clusters to be chosen randomly.
- For which a careful examination of data is required.

Q. 5(c) Explain Evaluation methods for clustering algorithms. **(4 Marks)**

Ans. :

- Evaluating or validating a clustering algorithm is an important process
- Different aspects that may be considered for the validation of clustering algorithm includes
  - o Clustering tendency in the data
  - o Correct number of clusters
  - o Quality of clusters
  - o Comparing two sets of clusters to find which is better
- Clustering evaluation methods can be broadly classified as follows
- Internal Validation methods and external validation methods

Internal Validation methods

- Using internal validation methods the quality of the clustering methods can be evaluated without using external information
- Two types of internal validation metrics can be used

1. **Cohesion :** This metric evaluates how closely the elements are within a cluster
2. **Separation :** This metric evaluates the level of separation between the clusters. These metrics are also known as internal indices as they do not use any external information

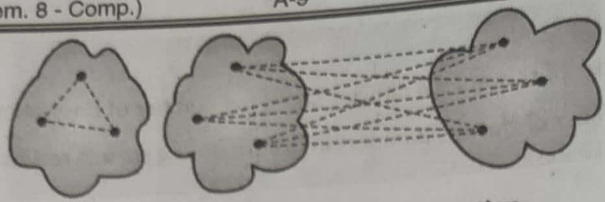

Fig. 1-Q. 5(c) : Cohesion and separation

Internal validation methods can be classified as follows

(A) Partitional

1. **Cohesion and separation :** Cohesion and separation do no perform well for algorithms based on density
2. **Proximity matrix :** Construction of a proximity or similarity matrix is computationally expensive. This method cannot be used without sampling for large data sets

(B) Hierarchical

External validation

- These validation methods can be associated with supervised learning problems
- This method requires additional information like external class labels for the training examples
- As we know unsupervised algorithms are used when such information is not available external validation methods are not used with most of clustering problems

OR

Q. 6(a) With reference to Meta Classifiers, explain the concepts of Weak and eager learner. **(Section 5.4.1)** **(8 Marks)**

Q. 6(b) Write short notes on : **(9 Marks)**

(a) Ada boost. **(Section 5.1.4(B))**

(b) Gradient Tree Boosting. **(Section 5.1.4(C))**

(c) Voting Classifier. **(Section 5.1.4(D))**

Q. 7(a) With reference to Hierarchical Clustering, explain the issue of connectivity constraints. **(Section 6.3.5)** **(8 Marks)**

Q. 7(b) What are building blocks of deep networks, elaborate. **(Section 6.5.2(D))** **(8 Marks)**

OR

Q. 8(a) With reference to Deep Learning, Explain the concept of Deep Architectures? **(Section 6.5.2)** **(8 Marks)**

Q. 8(b) Justify with elaboration the following statement :

The k-means algorithm is based on the strong initial condition to decide the Number of clusters through the assignment of 'k' initial centroids or means. **(Section 5.2.2(C))** **(8 Marks)**