IAF 604 - Machine Learning and Predictive Analytics

Exam 1

Name: Aman Tej Vidapu

Part 1: Multiple choice questions, each 5 points total 50 points (may contain more than one choice, underline all the correct answer(s))

1. The three V's of Big Data are:

    a. Volume, Viability and Vastness

    b. Volume, Viability and Velocity

    c. Volume, velocity and Veracity

    d. [✓ Volume, Variety and velocity]

    e. Volume, Variety and Viability

2. Technologies for big data machine learning include the following:

    a. [✓ Map/reduce]

    b. Webpage design

    c. [✓ Hadoop DFS]

    d. Virtual reality simulation

3. It involves the classification of different types of data and extraction of useful information from massive and complex datasets, which one is it?

    a. [✓ Big data classification]

    b. Logistic regression

    c. Support vector machine

    d. Clustering techniques

4. Which of the following ARE included in the phases of machine learning?

    a. [✓ Training]

    b. [✓ Testing]

    c. [✓ Validation]

    d. Optimization

5. Which big data controller is the contributor to imbalanced data problem?

   a. Observations

   b. Features

   c. Labels

   d. Subspace

6. If a data set has the observed values {1,2,3,4,5}, then which of the following are possible bootstrap samples to be used for a random forest technique?

   a. 5,2,1,4,3

   b. 2,2,1,3,3,4

   c. 5,1,6,3,3

   d. 2,2,5,4,4

7. Bootstrapping is applied as which part of the random forest algorithm

   a. Testing

   b. Training

   c. Validation

   d. Pruning

8. Which big data controller is the contributor to sparsity problem?

   a. Observations

   b. Features

   c. Labels

   d. Subspace

9. A trained SVM classifier for a two-dimensional data is given by $3x_1+4x_2+12 =0$. Which of the following point is NOT in the class of point (1,1)?

   a. (-1, -1)

   b. (1/3, -7/2)

   c. (-1/3, 7/2)

   d. (3, 2/7)

10. Elastic-net regression is defined by

    a. $e = (y-ax)^2$

    b. $e = (y-ax)^2 + \lambda a^2$

    c. $e = (y-ax)^2 + \lambda|a|$

    d. $\underline{e = (y-ax)^2 + \lambda a^2 + \lambda|a|}$

Part 2: Short questions, 5 points each

1. What is data science and which subject fields make major contributions to this hot interdisciplinary field?

Answer:

Data science is focused on gathering data and extracting/structuring useful insights from a given dataset, understanding of the systems that produce the data. system that works together to do tasks, such as collecting data, facts, or statistics about the environment the system is supposed to monitor, using a set of rules and statistics. The benefits from a large amount of data is helpful only if it is processed effectively. Currently, the need for storage increased dramatically.

The real world systems may produce a very large amounts of data known as Big data. Data here may be complex, unstructured, and make analysis. To handle this, a new field known as bid data science is introduced which has got huge research advancements on this problem discussed above. Example: big data classification, to classifys different types of data and the extraction of useful information from the massive and complex data sets. Handling massive dataset from real-world problems is challenging task which is being solved by this big data science which makes it as a robust in this field.

2. What are the main challenges to big data machine learning? What computer technologies can help to handle these challenges?

Answer:

The main Big data machine learning challenges are based on current techniques and technologies. The challenges caused with the techniques may be categorized as classification, scalability, and analysis. The challenges caused with the technologies may be categorized as computation, communication, and storage.

computer technologies can help to handle these challenges:

1. MapReduce programming model - modern programming frameworks

2. Hadoop - distributed file system

3. What are the major steps in MapReduce coding?

Answer:

Major steps in MapReduce coding:

1. mapper()

2. mapreducer()

3. reducer()

The mapper() function is used to parametrize a key domain and its value range. The mapper() function can be used to visualize the operation on a data table. For n-observations, if the data has p columns, then the goal of the function is to divide the data vertically into its key domain and value range.

The second one, mapreduce() is the mapping operation from mapper to reducer functions. It has a formal parameter list that duplicates the input data of the mapper() function and the reducer() function and runs them according to parametrization, sorting, and parallelization procedures.

while the third one, reducer() is to parallelizing the results. On the data returned by the mapper() function, the reducer() function may be seen as a horizontal (or row) action. The function picks the key's values in their value ranges and conducts the operations if the key domain has k key labels (i.e., a variety of keys) key1,key2,...,keyk.

4. What are the main supervised learning algorithms?

Answer:

1. Support-vector machines

2. Linear regression

3. Logistic regression

4. Naïve Bayes

5. Decision trees

6. K-nearest neighbor algorithm

7. Neural networks - Multilayer perceptron

8. Random Forest

Part 3: Linear Regression Problem, 30 points

a) Derive a straight line to fit the following points: (1,4), (2, 1), (3, 2), (4, 1), (5, 3). Give the equation of the line and explain the steps how to get that. (20)

Answer:
Linear Regression equation: y=Ax
A = yx'(xx')$^{-1}$
Considered text book code to solve the line equation:

3) a) Given data,

x	y
1	4
2	1
3	2
4	1
5	3

Straight line to fit given data as per linear regression | standard regression.

$$Y = Ax$$

~~Here~~ 'A' is given by $yx^T(xx^T)^{-1}$

Linear regression equation for

$$A = yx^T(xx^T)^{-1}.$$

$$x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \qquad y = \begin{bmatrix} 4 \\ 1 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

$$x^T = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

Matrix multiplication can be done accordingly as per dimensions.

$$\Rightarrow \quad x^T x = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$= 55$$

$$\Rightarrow (x x^T)^{-1} = 1/55 = 0.018$$

$$\Rightarrow \quad x^T (x x^T)^{-1}$$

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \end{bmatrix} \cdot 0.018$$

$$= \begin{bmatrix} 0.018 & 0.036 & 0.054 & 0.072 & 0.09 \end{bmatrix}$$

$$\Rightarrow \quad y x^T . (x x^T)^{-1}$$

$$= \begin{bmatrix} 0.018 & 0.036 & 0.054 & 0.072 & 0.09 \end{bmatrix} . \begin{bmatrix} 4 \\ 1 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

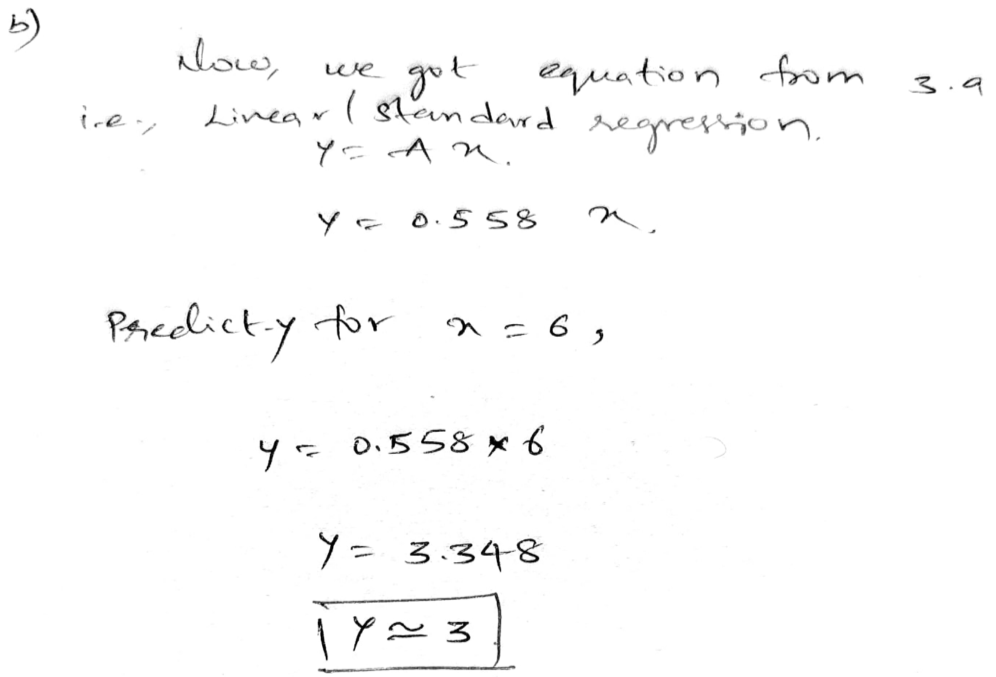
$$A = 0.558$$

b) Using the above linear model to predict the y value when x=6. (10)

Answer:

From above linear equation y=Ax

y = 0.558x is the standard regression equation.



b)

Now, we got equation from 3.a

i.e., Linear (Standard regression.

$Y = A x.$

$Y = 0.558 x.$

Predict y for $x = 6$,

$y = 0.558 \times 6$

$Y = 3.348$

$\boxed{Y \simeq 3}$

Reference:

Machine Learning Models and Algorithms for Big Data Classification Thinking with Examples for Effective Learning Book by – Dr Shan