Student Name: - Kiran Digambar Yadav.

Assignment Name: - 2 (Worksheet-2)

Internship: - 35

Machine Learning (Assignment No-2)

Question No	Answers
1)	c) 1,3
2)	d) 1,2,4
3)	a) True
4)	a) 1 only
5)	b)1
6)	b) No
7)	a) Yes
8)	d) All of the above
9)	a) K mean clustering algorithm.
10)	d) All of the above
11)	d) All of the above

Q12 Answer: -

The K means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together the average make sense. However, when you have outliers, this can affect the average calculation of the cluster. As a result, this will push your cluster centre closer to the outlier.

Let us understand it with example, the average salaries of following people are as below.

30K, 40K, 50K, 60K, 65K & 1000K

The average ends up being (30K+40K+50K+60K+65K+1000K)/ (6) = 207.5K

If we did not have the 1000K outlier, the average would have been (30K+40K+50K+60K+65K) / (4) = 49K.

Note that the two average results are widely different from one another. Given that K means clustering is an unsupervised algorithm it is up to the interpreter to determine whether this makes sense or not for a given dataset. Those we can conclude it is sensitive to the outliers.

Q13 Answer: -

K means the number of clustering and means implies the statistics mean a problem. It is used to calculate code vectors (the centroid of different clusters). It calculates the distance from all the code vectors and assign the index of the code

vector with the minimum distance to this value. K means clustering has been around since 1970s and fares better than other clustering algorithms like density based, exception-maximisation. It is one of the best robust methods especially for image segmentation and annotation projects. K means is used to learn feature representation for images (use k means to cluster small patches of pixels from natural images, then represent images in the basis of cluster centres, repeat this several times to form a deep network of feature representation) gives image classification results that are competitive with much more complex deep neural network models. In fact, a lot of k means applications are now done using support vector machines. Hence, we can say it is better and simple to implement model.

Q14 Answer: -

Clustering algorithm with steps involving randomness usually give different results on different executions for the same dataset. This non-deterministic nature of algorithms such as k means clustering algorithm limits their applicability in prediction. The non-deterministic nature of k means is due to its random selection of data points as initial centroids. Here we proposed an improved density-based version of k means, which involves a novel and systematic method for selecting initial centroids. The key idea of algorithm is to select data points which belongs to dense regions and which are adequately separated in feature spaces as the initial centroids. Here we compare proposed algorithm to a set of eleven widely used clustering algorithms and a prominent ensemble clustering algorithm which is being used for data classification. The proposed algorithm has shown better overall performance than the others. Hence, we can conclude, it is non-deterministic in nature but we propose some modifications to yield better results out of it.

Worksheet 2 -SQL (Assignment-2)

Question No	Answers
1)	D) Unique
2)	A) Primary Key
3)	A) Each entry in the primary key uniquely identifies each entry or row in table.
4)	D) All of the above.
5)	B) Foreign Key
6)	C) 2
7)	C) One to one
8)	C) One to one
9)	D) None of them
10)	D) 2
11)	C) One to one
12)	C) Table
13)	A) Insert Into
14)	B) Unique & C) Primary key
15)	B) A blood group can contain only characters. & C) A blood group cannot have null values.

Statistics Worksheet-1 (Assignment-2)

Question No	Answers
1)	C) Both
2)	C) 12
3)	D) All of the above
4)	C) Both of these
5)	D) All of these
6)	B) Data set
7)	A) Two or more
8)	B) Scatterplot
9)	D) Analysis of variance
10)	A) Z-score
11)	C) Mean
12)	D) 400005.2
13)	D) Mean
14)	A) Descriptive & Inferences
15)	D) H-L