

Name- Mahizebin shams-F- Mofiz

ID - 1712639692

Section- 05

" Assignment - 03 "

Date: 25 May, 2021

(A)

Ans to the Qus NO-01

Q1.

PCA is considered only as a variance to the features, but it is not considered as a relationship between features and labels while doing feature reduction.

But in case of regularization acts directly on the relationship between features and labels. Thus, regularization helps to develop models that are better at showing predictions for a given set of features.

As we apply PCA to reduce features in a data set so regularization makes the process better for this reason we still need regularization.

(2)

Ans to the Ques NO-02

Q2.

F₁ score:

F₁ score is the balance between the precision and the recall.

$$\therefore F_1 = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

I would use F₁ score as a performance metric for classification algorithms and gives a better measure of the incorrectly classified cases than the accuracy metric.

The methods I use to handle missing or corrupted data in a dataset are —

(3)

Method-01

Deleting rows or columns:

We usually use this method when it comes to empty cells. If the majority of our data is missing for a column or for a row, we can simply delete them.

Method-02

Replacing the missing data with

aggregated values:

We can calculate the aggregated value based on the rest of the values we have in the column and put the received numbers to the empty spot.

Method-03

(1)

Creating an unknown category:

Categorical features have a number of possible values, so we can create one more category for the missing values. This way the lower variance by adding new information to the data, will create. This also could be used when the original information is missing or cannot be understood.

Method-04

Predicting missing values:

Where we have no missing values, we can train a machine learning algorithm in order to predict the missing values since the samples for which this training

is performed. There are missing values, it is necessary to replace initially using one of the simplest methods for recovering gaps.

This way it will create and give us better performance unless a missing value should have a high variance.