CPE 695 - Applied Machine Learning

HW-5

Name: Hinal Patel
CWID: 10473912

Question 1 (25 points):

1) Explain why it is important to reduce the dimension and remove irrelevant features of data (e.g., using PCA) for Instance-Based Learning such as kNN? (5 points)

Ans:

K- Nearest neighbor generates frequent data points that improve the computer system's learning process in obtaining more precise results from its knowledge base. We are aware that data is growing at a rapid rate. If we do not minimize the dimensions and eliminate unnecessary data features, we may encounter a variety of challenges, including the following:

- The amount of data will grow exponentially in proportion to the number of dimensions, making it inefficient to manage.
- This will further increase the size of the data set, since the data set will expand exponentially along with the dimensions, causing a difficulty with algorithms such as kNN, where the data points will have a much greater distance between them.

Now imagine that the data point has a significantly larger gap between them. This will result in the following,

- The algorithm's efficiency will drop dramatically when more data points are collected to enable more precise decision making, since kNN requires the formation of neighbors for the data points.
- Even with the addition of newer data sets, the algorithm will no longer assist in making exact decisions, even if the data size does not expand dramatically and rapidly.

Thus, these are the reasons for the requirement to remove irrelevant data characteristics and minimize data dimensions.

2) One limitation with K-Means is the variability issue. Explain how to address this problem. (5 points)

Ans:

The fundamental problem with variability in K-means is that it assumes that every variable's distribution variance is spherical. This indicates that the variance of all variables is the same.

This also implies that each of the k clusters will have the same prior probability and number of observations. However, k-mean fails totally if any of the clusters breaches this assumption. Because of the high values, it is also susceptible to outliers.

This problem may be handled by utilizing various data points and improved initialization. To generate a large number of potential solutions, one may repeatedly initiate or start the process from various positions.

The best option for the result may be retained. This may be enhanced by adding some unpredictability. Variants in repeated k-means need this.

3) Please explain the technique of Gaussian Mixture and how it is used for anomaly detection. (5 points)

Ans:

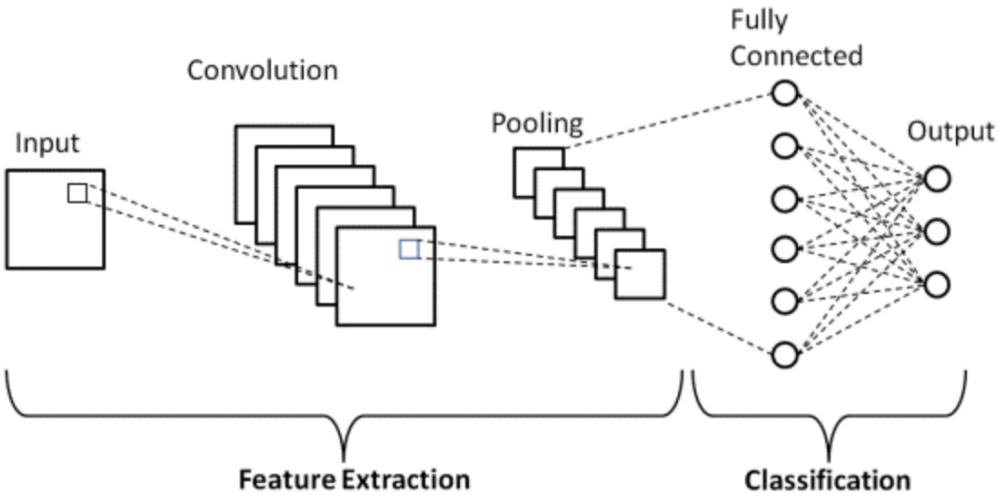
The Gaussian Mixture, It is a probabilistic model in which the data points are generated by a mixture of Gaussian distributions with certain parameters. These mixture models extend the k-means clustering method by include characteristics about the data structure, such as covariances.

The gaussian model object fits gaussian mixture models using the EM technique (expectation maximization). Additionally, it generates confidence ellipsoids for multivariate models and determines the total number of clusters in the data. A fit approach constructs a Gaussian Model from the training data.

If we have the test data, we may allocate it to all of the Gaussian samples and use the predict technique to find out. This gaussian mixture includes multiple options for constraining the covariance for difference estimated classes such as spherical, diagonal, and complete covariance.

To deal with anomaly detection, we must first fit a model to the data distribution. Let the training set be{ x (1)...., x (m)}, where x (i) ∈ Rn, n = 2), and we wish to predict the Gaussian distribution for all features. We must discover the mean for each characteristic, and with the aid of this mean, we can determine the variance of these features. Now that we have mean and variance, we must calculate the probability of these training instances so that we can determine which cases are anomalous.

4) Please draw the diagram of Convolutional Neural Networks (CNN). Then explain the functionality of each layer of CNN. Name several latest algorithms of CNN (e.g., AlexNet etc.). (5 points)



The CNN is made up of three kinds of layers: convolutional layers, pooling layers, and fully-connected (FC) layers. A CNN architecture will be constructed when these layers are layered. In addition to these three layers, there are two additional significant factors to consider: the dropout layer and the activation function, both of which are detailed further down.

1. Convolution Layer: This is the first layer utilized to extract different characteristics from the input photos. This layer performs the mathematical action of convolution between the input picture and a filter of size MxM. By swiping the filter over the input picture, the dot product is calculated between the filter and the regions of the input image that are proportional to the filter's size (MxM). The result is known as the Feature map, and it contains information about the picture such as its corners and edges. This feature map is then supplied to further layers, which learn various different features from the input picture.

2. Pooling Layer: A Convolutional Layer is usually followed by a Pooling Layer. This layer's major goal is to lower the size of the convolved feature map in order to reduce computational expenses. This is accomplished by reducing the connections between layers and operating independently on each feature map. Pooling activities are classified into numerous kinds based on the approach utilized.

3. Fully Connected Layer: The Fully Connected (FC) layer, which includes weights and biases as well as neurons, is used to link neurons from various layers. These layers are often placed before to the output layer and constitute the last few levels of a CNN Architecture.

4. Dropout: When all of the characteristics are linked to the FC layer, the training dataset is prone to overfitting. Overfitting happens when a model performs so well on training data that it has a detrimental influence on the model's performance when applied to new data.

5. Activation Function: Finally, the activation function is a critical element in the CNN model. They are used to learn and estimate any kind of continuous and complicated connection between network variables. In other words, it determines which model information should be sent forward and which should not at the network's end.

CNN Algorithms are listed below,

- LeNet-5 (1998)
- AlexNet (2012)
- ZFNet(2013)
- GoogLeNet/Inception(2014)
- VGGNet (2014)
- ResNet(2015)

5) What are the vanishing and exploding gradients problems in Back propagation? Name several techniques to address these problems. (5 points)

Ans:

- Vanishing Gradients:

Vanishing Gradient happens when the derivative or slope becomes less and lower as we back propagate with each layer. When the weights update is extremely little or exponentially small, the training period is excessively lengthy, and in the worst situation, the neural network training is totally stopped.

Because the derivatives of the sigmoid and tanh activation functions are between 0 and 0.25 and 0–1, the sigmoid and tanh activation functions have a vanishing Gradient issue. As a result, the new weight values are extremely comparable to the old weight values, and the updated weight values are modest. This results in the Vanishing Gradient issue. We may circumvent this issue by utilizing the ReLU activation function, which has a gradient of 0 for negative and zero input and 1 for positive input.

- Exploding gradients:

During back propagation, an exploding gradient happens when the derivatives or slope get more and larger as we move backward with each layer. This is the inverse of the vanishing gradients issue. This issue arises as a result of weights, not the activation function. Because of the high weight values, the derivatives will similarly be high, resulting in a large difference between the new and old weights, and the gradient will never converge. As a consequence, it is possible that it will fluctuate near minima but never reach a global minimum point.

These are several techniques to address these problems,

- Redesigning the Network Model
- By using Weight Regularization Method
- By using Gradient Clipping Method
- Using Long Short Term Memory Networks

CPE 695 - Applied Machine Learning

HW-5

Question 2 (5 points):

Consider a learned hypothesis, h, for some Boolean concept. When h is tested on a set of 100 examples, it classifies 80 correctly. What is the 95% confidence interval for the true error rate for $Error_D(h)$?

Ans:

Let, x = Number of hypothesis classifies incorrectly.
n = sample size
Given, n = 100, x = 100 - 80 = 20
Let, P = rate for $Error_D(h)$

P = x / n = 20/100 = 0.2

C% confidence interval for population proportion is given by,

(P - E, P + E)

where, E is margin of error and it is given by,

$$E = z' * ((P * (1 - P))/n)^{1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

z' = critical z score for given confidence interval.
Given, Confidence level (c) = 0.95

α = 1 - C = 0.05 , α/2 = 0.025,
z' = z_0.025= 1.96

Put the values in equation 1,
E = 1.96* ((0.2 * (1 - 0.2))/100)^1/2 = 0.0784

Now,
P - E = 0.20 - 0.0784 = 0.1216
P + E = 0.20 + 0.0784 = 0.2784

95% confidence interval for the true error rate for $Error_D(h)$ is (0.1216, 0.2784)