Harsh Sanjay Agrawal

Homework 5

Q1a. 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?

PCA determines which axis accounts for the most variance in the training data, and it does so for each subsequent axis as well. These axes are at right angles to one another. Instance-based learning techniques like KNN now suffer from the Curse of Dimensionality. The volume of the feature space may be extraordinarily huge if there are many dimensions in that space, and as a result, the points we have in that space frequently represent a sparse and unrepresentative sample. For instance, if there are 20 characteristics used to describe the instance but only two of them are pertinent to the target function, the remaining attributes may be harmful to the target function and produce false results.

As a result, it's crucial to lower the dimension and eliminate pointless aspects of the data (using, for example, PCA or instance-based learning techniques like kNN).

Q1b. One limitation with K-Means is the variability issue. Explain how to address this problem.

When we run K-means numerous times, we get various clusters, which is the K-means variability issue. A performance indicator called inertia is primarily the total of the distances between instances and their respective centroids. Now, we can solve this problem by repeatedly running K-Means with various beginning centroids. The final clustering is chosen because it has the lowest inertia. Therefore, even though various clusters form each time, we identify and employ the clustering that has the centroid with the lowest inertia. This is how the variability problem with K-means can be solved.

Q1c. Please explain the technique of Gaussian Mixture and how it is used for anomaly detection.

A dataset anomaly is essentially an out-of-the-ordinary observation in the dataset. A method for unsupervised clustering is the Gaussian mixture model. In this method, we fit k Gaussians to the data and then calculate the mean, variance, and weight of each cluster's Gaussian distribution. Finally, we determine the odds of each point being a member of each of the three clusters.

1. One dimensional mode:

$$p(x) = \sum_{i=1}^{k} \phi i N(x|\mu_i, \sigma_i) \text{ and } N(x|\mu_i\sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{(x-\mu_i)}{2\sigma_i^2} \right)$$

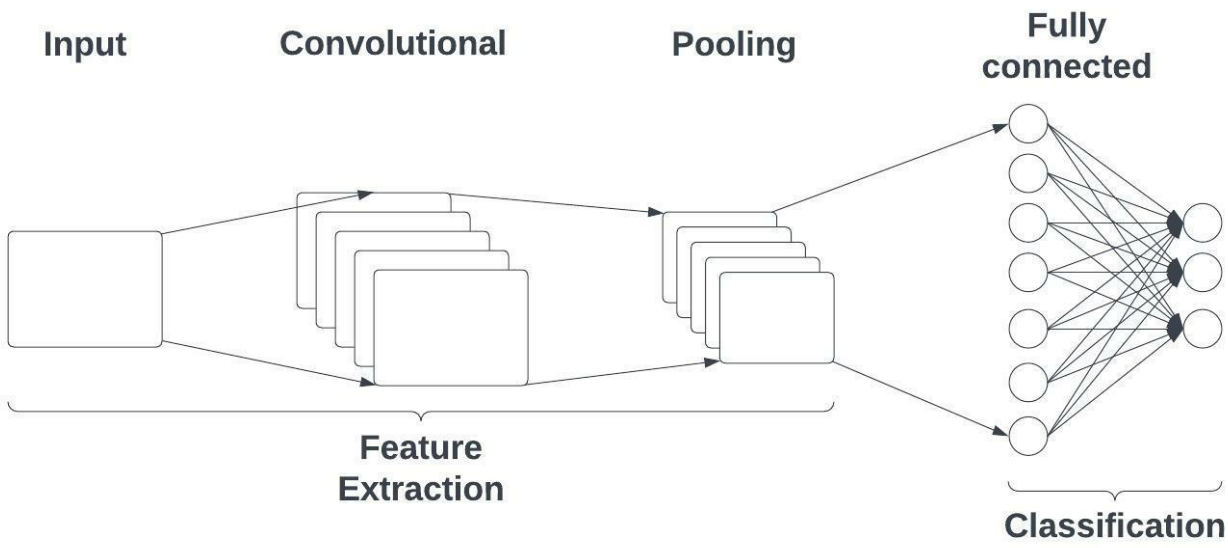
where, $\mu_k = mean$ and σ_k = variance for the k^{th} component ϕ = weight for the cluster 'k'

2. Multi-dimensional mode:

$$p(x) = \sum_{i=1}^{k} \phi_i N(x|\mu_i\Sigma_i)$$

A weighted average for the K gaussian distributions is described by the equations taken collectively. The program then uses these k clusters to train. Therefore, the method calculates the likelihood that a new point will belong to each cluster based on its distance from each distribution. As a result, if the likelihood is extremely low for a certain cluster, the datapoint is likely an anomaly.

Q1d. 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.).



There are three types of layers that make up the CNN.

1. Convolutional layers

2. Pooling layers

3. Fully connected layers

The activation function and the dropout layer are two additional crucial parameters in addition to these three layers. The different features from the input photos are extracted using the convolutional layer. The result is known as the feature map, and it provides details about the image, including its corners and edges. The pooling layer is mostly used to shrink the complicated feature map in order to save money on computation. It fills the space between the FC and convolutional layers. Weights, biases, and neurons between two separate layers make up the FC later. The flattened vector moves on to the output, which is utilized for classification, after undergoing the necessary modifications. The final classification uses the output layer. The overfitting of the training dataset is a problem that is solved through dropout. The use of the activation functions depends on the classification and how non-linearity is to be added. AlexNet, VGG, Inception, ResNet, Densenet, CBAM, and other recent CNNs are a few examples.

Q1e. What are the vanishing and exploding gradients problems in Backpropagation? Name several techniques to address these problems.

The gradients of ten get smaller and smaller and finally approach zero as a backpropagation algorithm moves backward from the output layer to the input layer, leaving the weights of the initial or lower layers essentially untouched. Because of this, the gradient decline never reaches the ideal state. A vanishing gradient problem is hence explained by as this process.

Contrarily, in some instances the gradients continue to grow as the backpropagation process continues to run. Due to this, the gradient descent diverges and results in very high weight updates. The exploding gradient problem is hence explained by as this process.

Some of the techniques to address this problem are:

1. Proper weight initialization

2. Using non-saturating activation functions

3. Batch normalization

4. Gradient Clipping

Q2. 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 ErrorD(h)?

For the given example, 20 examples are incorrectly classified on h.

Hence errors = 20 / 100 = 0.2

To achieve the 95% confidence interval for the true error rate

$$Error_D(H) = errors(h) \pm 1.96 \sqrt{\frac{errors(h)(1-errors(h))}{n}}$$

$$= 0.2 \pm 1.96 \sqrt{\frac{0.2(0.8)}{100}}$$

$$= 0.2 \pm 0.0784$$

Hence, the confidence interval is (0.1216, 0.2784)