ISMET OKAN CELIK CWID:10472265

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)

If there are irrelevant attributes in the data, these attributes reduce the accuracy of the KNN. Because when we calculate the distance in KNN, we consider all the attributes. This algorithm can be easily fooled by irrelevant attributes. Also, if there are a lot of attributes then computation complexity will be very high. In another word, making calculations with high dimension data will be complex and slow that is why PCA is important for KNN

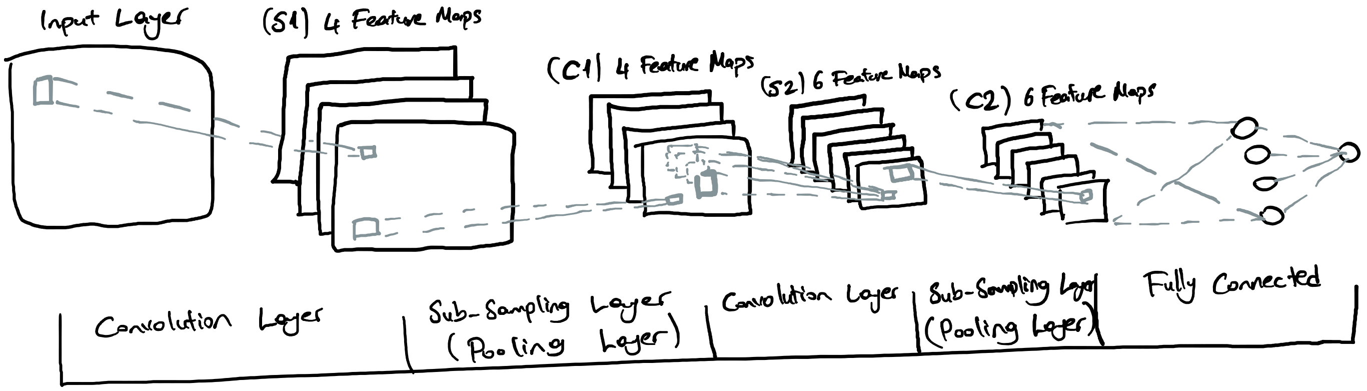
2) One limitation with K-Means is the variability issue. Explain how to address this problem. (5 points)

Running K-means multiple times results in different clustering and the problem is not knowing which one is the best clustering. The K-means algorithm can return clustering for any number k. Optimal k should have a good balance of inertia vs efficiency. To be able to solve this problem elbow rule or silhouette score can be used.

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

Gaussian Mixture Model is a probabilistic model that assumes that the instances were generated from a mixture of several Gaussian distributions. Anomaly detection means that we need to find the data which are statistically different from the rest of the data. Any data point located in a low-density region can be considered an anomaly. For example, in a manufacturing company that tries to detect defective products, let's say the ratio of the defective products is 3%, then we can set the threshold to be the value that results in having 3% of the instances located in areas below that threshold density. If we notice that there are too many good products marked as defective then we can lower the threshold or if we have too many defective products which are not marked as defective we can increase the threshold.

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



Input Layer: Input layer converts an image file (CNNs are usually used for images) to a matrix.

Convolution Layer: This layer is the first layer that is used to extract the various features from the input images. In this layer, the mathematical operation of convolution is performed between the input image and a filter of a particular size MxM.

Pooling Layer: In most cases, a Convolutional Layer is followed by a Pooling Layer. The primary aim of this layer is to decrease the size of the convolved feature map to reduce computational costs.

Fully Connected Layer: The Fully Connected (FC) layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers. These layers are usually placed before the output layer and form the last few layers of a CNN Architecture.

Latest Algorithms of CNN: LeNet-5, AlexNet, GoogLeNet, VGGNet ,ResNet-50, Inception-v4,SENet,YOLO, Capsule Networks

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

Vanishing gradient: Gradients get smaller and smaller toward lower layers and it tends to disappear in Backpropagation, which means the gradient gets too small.

Exploding gradient: Gradients grow bigger and bigger till the algorithm diverges. In the case of exploding gradients, the accumulation of large derivatives results in the model being very unstable and incapable of effective learning, The large changes in the model's weights create a very unstable network.

Solution for the problems:

- Improving weight initialization
- Using non-saturating activation functions (Using ReLU instead of Sigmoid Function)
- Batch Normalization
- Gradient clipping

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 ErrorD(h)?

$$error_S(h) = \frac{100-80}{100} = 0.2$$

$$z N(95\%) = 1.96$$

$$Error_D(h) = errors(h) \pm 1,96 \sqrt{\frac{0,2 \cdot 0,8}{100}}$$

$$Error_D(h) = 0,2 \pm 1,96 \cdot 0,04$$

$$Error_D(h) = (0.1216 \ , \ 0.2784)$$