Question 1

1. To remove irrelevant features from the data, it can help increase the accuracy of our models as irrelevant features influence the model accuracy, and it need more time to fit the model.

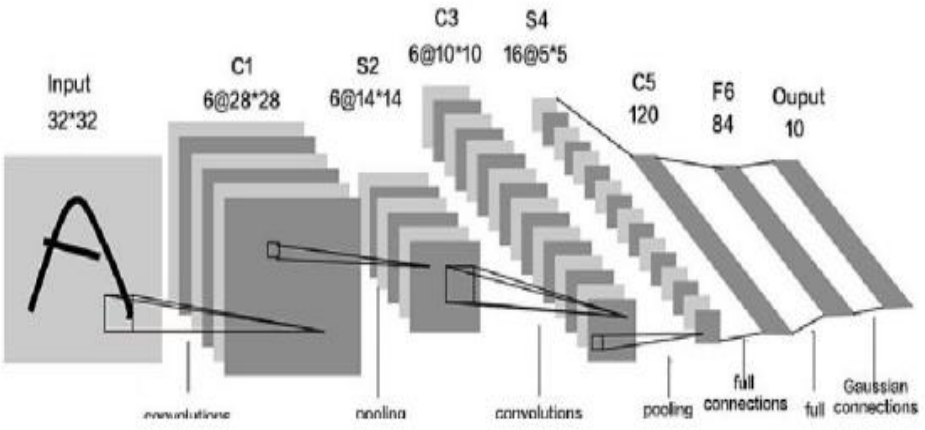
2. Variability issue is that user have to choose appropriate K, we can use elbow rule or silhouette score to help us choose the number of clusters K.

3. In anomaly detection, we try to identify observations that are statistically different from the rest of the observations.

The algorithm train upon these K clusters. Thus given a new data point, the algorithm finds its distance from every distribution & hence the probability of that point belonging to each cluster. Therefore if for a particular cluster, if the probability is very low that's an indication of the data point being an anomaly.

For multi-dimension anomaly detection, if the K cluster has lower probability, which is the anomalous cluster.

4.



1. input layer : For data input

2. Convolutional layer : Using convolution kernel to extract feature and feature mapping

3. activating layer : Add nonlinear mapping

4. pooling layer : Reduce the amount of data operation, the feature map is sparse processed by down sampling.

5. fully connected layer : Re fitting is used to reduce the loss of feature information

6. output layer : For result output

Example: Google Net, ResNet

5.

Vanishing gradients: Vanishing gradients is the problem of gradient is gradually too small by accumulation and tend to

disappear in Backpropagation, like the number of gradient is smaller than 1 .

Exploding gradients:Exploding gradients is the problem of gradient is gradually too large and by accumulation tend to disappear in Backpropagation like the number of gradient is greater than 1 .

Method to address :

Use ReLu activation function

Weight initialization (He initialization and Xavier initialization)

Gradient Clipping

Question 2

Error_S(h) = 20 / 100 = 0.2

Z-N(95%) = 1.96

Error_D(h) = Error_S(h) $\pm 1.96\sqrt{(0.2*0.8)/100}$