Page No.
Date

NAME - SIDDHARTH PARAAG NILAKHE

A.ML HW - 5

Q1]

1) We all know that with an increase in the dimension, the size of the dataset tends to increase which will lead to an increase in the gap between the datapoints.
This causes a decrease in the efficiency of the algorithm.
During such a case use of PCA can help in improving the performance at a very low cost of accuracy. It has benefits such as ① producing independent, uncorrelated features of data and ② reduction in noise.
Some of the problems that are caused by large data are ① leads to overfitting ② Long computational time

2) The fundamental issue with k-means variability is that it assumes that every variables distribution variance is spherical. This implies that it is vulnerable to outliers & each of the k-clusters has the prob & the num of observation.
Ways to address the problem-
→ Inject some randomness in initialization pts
→ Along with random initialization try increasing the repeats for better avg of errors.

3) A GMM model is a probabilistic model in which all data points are generated by combining a finite number of Gaussian distributions with unknown parameters.

GMM → Generalization of k-means clustering. GMM contains info on the data's covariance structure and latent Gaussian centres.

$$f(x) = \sum_{m=1}^{M} \alpha_m \phi(x; u_m, \Sigma_m)$$ where,

$\phi(x; u_m, \Sigma_m)$ is a Gaussian fn with mean $= u_m$ and covariance matrix. $\sum_{m=1}^{m} \alpha_m = 1$

$$P(x_i \in C_m) = \frac{\alpha_m \phi(x_i; \hat{u}_m; \Sigma_m)}{\sum_{k=1}^{m} \alpha_m \phi(x_i; \hat{u}_k, \hat{\Sigma}_k)}$$

Anamoly detection:-

1) Univariate case - One - D - model

$$P(x) = \sum_{i=1} \phi_i N(x \mid u_i; \sigma_i)$$

$$N(x \mid u_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\left(\frac{(x-u_i)^2}{2\sigma_i^2}\right)}$$

Here u_x = mean

σ_k^2 = Variance for k^{th} element

ϕ_x = weight for cluster x

2) For multivariate case Multi Dimensional

$$P(\vec{x}) = \sum_{i=1}^{k} \phi_i N(\vec{x} \mid \vec{u_i}, \Sigma_i)$$

u_x = mean Σ_k = covariance matrix of x^{th} element

ϕ_x = weight of cluster k

Here, the algo trains upon the k clusters. Hence when we receive a new data pt the algo determines the dist from every distributⁿ & gives out the prob of belonging to each cluster.
If we have a cluster with the low prob then it tells us the data pt is an anomaly.

4) Input → Convolution layer → Pooling Layer → Fully connected Layer   o/p

Convolution layer - We perform convolution operation in order to extract features of the image.
& Activation layer is a sublayer of convolution layer which consists of activation functions like sigmoid, Relu, etc.

Pooling Layer - This is a sampling layer to reduce size of i/p & to try & keep overall value of the i/p same.

Fully connected layer - It is similar to a regular neural network. It is the final learning phase of the network which maps to extracted features of o/p.

CNN algos → AlexNet, ResNet, LeNet

5) 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 is the problem of gradient is gradually too large and by accumulation tend to disappear in Backpropogation like the number of gradient is greater than 1.

Method to address - ① Use ReLU activation fn
② Gradient Clipping ③ Weight initialization.

Q2] i - no. of samples classified incorrectly
$= 100 - 20 = 80$
n - sample size = 100

$\hat{p}$ - date for $Error_p(h)$

$$\hat{p} = \frac{20}{100} = 0.2$$

c %. confidence interval $= (\hat{p} - E, \hat{p} + E)$ where

$$E = z' \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ - margin of error

z' is critical z score for a particular confidence interval

$$\alpha = 1 - C = 1 - 0.95 = 0.05$$
$$\frac{\alpha}{2} = 0.025$$
$$z' = z_{0.025} = 1.96$$

$$E = 1.96\sqrt{\frac{0.2(1-0.2)}{100}} = 0.0784$$

95% = confidence interval for true error rate

for $Error_p(h) = (0.2 - 0.0784, 0.2 + 0.0784)$

$= (0.1216, 0.2784)$