Taryn Michael

201800323

Advanced Machine Learning NAML841 Final Exam

26th July 2021

Question 1:

1.1. Firstly, the process of collecting or generating the data will be stochastic. This is due to the noise in the data or certain errors possibly made while collecting the data. Secondly, our y values (predicted values) will be stochastic due to the noise in the data being approximated to be distributed under a certain probabilistic distribution. Y can then be seen as a random variable distributed according to a certain distribution. For example, y may be normally distributed and represented as y ~ N(0,1). The ~ indicates the stochastic nature of the variable.

1.2. The normal distribution

Question 4

4.1 Weaknesses of using k-means for clustering elliptical data.

K-means has the disadvantage in its lack of flexibility in cluster shape, and its lack of probabilistic cluster assignments. This means it is unable to handle the uncertainty when a data point is close to more than one cluster centroid. K-means thus requires that cluster models must be circular and cannot account for oblong or elliptical clusters. It also requires clear separation of data points in the data. Thus, the cluster assignments end up becoming muddled if one tries to use k-means on elliptical cluster shapes.

4.2 Gaussian Mixture Models is an unsupervised clustering technique that forms clusters based n the probability estimations using the Expectation-Maximization function. Using the mean and covariance over K-means which only uses the mean, provides Gaussian Mixture Models with he ability to provide a better quantification measure of fitness per number of clusters. It thus accounts for uncertainty as it contains a probabilistic model under the hood, meaning it is possible to find probabilistic cluster assignments.

A gaussian mixture model can also fit to the oblong or elliptical data quite well by allowing for a full covariance type. The covariance is a hyperparameter that controls the degrees of freedom in the shape of each cluster. This provides flexibility to Gaussian Mixture Models as it can use its covariance type to account for any cluster shape. Specifying the ‘full’ covariance type in python simply allows each cluster to be modelled as an ellipse with arbitrary orientation.