Machine Learning – 336546

HW3

Taima Zoabi 318476868

Q1- Clustering

a) K-medoids uses an actual point in the cluster with minimum sum of distances to other points to represent the center of a cluster, instead of using the mean point as is the case in K-means. The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. Therefore, K-medoids is more robust to noises and outliers.

b) In order to find the centroid (μ) which minimizes the term

$$J(\mu) = \sum_{i=1}^{m} (x_i - \mu)^2$$

we need to differentiate and compare to zero-

$$\frac{dJ(\mu)}{d\mu} = -2 \sum_{i=1}^{m} (x_i - \mu)^2 = 0$$

$$\mu m - \sum_{i=1}^{m} x_i = 0$$

$$\boxed{\mu = \frac{1}{m} \sum_{i=1}^{m} x_i}$$

c) We need to differentiate and compare to zero in order to find the centroid (practically, the medioid) which minimizes the term, -

$$J(\mu) = \sum_{i=1}^{m} |x_i - \mu|$$

$$\frac{dJ(\mu)}{d\mu} = -\sum_{i=1}^{m} sign(x_i - \mu) = 0$$

$$\sum_{i=1}^{m} sign(x_i - \mu) = 0$$

In order for this term to equal zero we need exactly half of the x_i's to be bigger than μ and exactly half of them to be smaller, in other words μ has to be median of m examples.

Q2- SVM

In images A&D the data is separated using a single Line(linear boundary), which indicates that a linear kernel has been used. As for the C parameter, it tells the SVM

optimization how much we want to avoid misclassifying each training example. So a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. And for a larger C we get a hyperplane that correctly separates as many instances as possible, even if it was with a much smaller margin. Therefore, A has the smaller C value (0.01) it's two purple samples which are the most closer to boundary are in the margin area because they cannot be the support vector otherwise there will have to be at least one blue sample at the same distance from the boundary from the other side. And in the other hand, in D there is a safe distance between the samples of both classes and the boundary which means a higher value of C=1.

Clearly, C&F are with the polynomial kernel, because the data is separated with high order polynomial. And the higher degree polynomial order kernels allow a more flexible decision boundary. So we can see that the more flexible boundary is in image F so we assume it is with 10^{th} order polynomial kernel, and C has Quadratic function as a boundary (2^{nd}).

That leaves us with the Gaussian kernel (RBF) in images B&E, indeed they has closed Gaussian boundary. We learned in the lectures that the higher Gamma is the more we better fit the training data. So too high of a Gamma can cause overfitting. In addition, it can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself (small region), and vice versa for a small values of gamma the region of influence of any selected support vector would include the whole training set(large region). Therefore, B is with $\gamma = 1$ and E with $\gamma = 0.2$.

To sum the results up: A-1, B-6, C-3, D-2, E-5 and F-4.

Q3- Capability of generalization

a) Generalization, it refers to how well the concepts learned by a machine learning model will translate to new observations not seen by the model when it was trained. A properly generalized model will assume balance between how good it fits the training and the test sets and the complexity of the model. We seek simplicity because it is easier to perform and less prone to overfitting.

b) We always seek to maximize the likelihood ($\hat{L}$) given some parameters, therefore to maximize $\ln(\hat{L})$ cause it is a monotonic function. And the parameters which maximize the likelihood guaranty a best fitted therefore generalized model. As for the number of the parameters, it represents the complexity of the model, the higher p is (therefore (2p)) the more complex our model and likelihood function become.

c) If this balance was violated we might get overfitting or underfitting, as I said before complexity of a model could cause overfitting and with the goodness of fit comes a more complex model. In the other hand, with a very simple model it could less fit and therefore underfitted.

d) We are aiming for a low AIC, because as was mentioned before we want to maximize the likelihood so we will achieve a good fit model, also we do not want it to be a very complex one so we are in favor for an approximately low number of parameters. As a result, the AIC should be low.