HW3-MLH

Shachar Zigron 316356401

Theoretical Questions

1. Clustering:

a. K-medoid is more robust to noise than the K-means algorithm because K-means minimize the sum of squared Euclidean distances between μ and the observations whereas K-medoid minimize sum of pairwise dissimilarities (distance between two observations). so, an outlier in the observation will affect more on the sum of squared Euclidean distances and cause him to be higher which will result to change in μi.

In K-medoid the centres of the clusters is Medoids that they dissimilarity to all the objects in the cluster is minimal. And therefore, like with median, this algorithm isn't sensitive to outliers.

Example – data set (1,5,7,8,16,18,8000), the median(8) represents better the data set than the average=1150.714

b. $x \in \mathbb{R}^1$

$$\bar{X} = \frac{\sum_{i=1}^{m} x_i}{m} = mean\ of\ m\ exampels$$

$$\sum_{i=1}^{m}(x_i-\mu)^2 = \sum_{i=1}^{m}(x_i-\bar{X}+\bar{X}-\mu)^2$$
$$= \sum_{i=1}^{m}(x_i-\bar{X})^2 + 2\sum_{i=1}^{m}(x_i-\bar{X})(\bar{X}-\mu) + \sum_{i=1}^{m}(\bar{X}-\mu)^2$$
$$= \sum_{i=1}^{m}(x_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{m}(x_i-\bar{X}) + m(\bar{X}-\mu)^2$$
$$= \sum_{i=1}^{m}(x_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{m}0 + m(\bar{X}-\mu)^2$$
$$= \sum_{i=1}^{m}(x_i-\bar{X})^2 + m(\bar{X}-\mu)^2$$

That is, in order to minimize the expression $\sum_{i=1}^{m}(x_i-\mu)^2$ *its required to minimize the expression:* $m(\bar{X}-\mu)^2$.

So μ that minimizes this expression is the average of m samples.

$$\frac{d(m(\bar{X}-\mu)^2)}{d\mu} = -2m(\bar{X}-\mu) = 0$$
$$\mu = \bar{X}$$

Second derivative, to make sure it is the minimum:

$$\frac{d\left(-2m(\bar{X}-\mu)\right)}{d\mu} = 2m > 0$$
$$\Rightarrow \mu = \bar{X} \ \ minimum\ point$$

2. SVM

Linear kernel: for this type of kernel, I expect a straight separate line as can be seen in graphs A and D.
As is well known the parameter C tell us how much the error is tolerable, i.e. the higher C the more we punish the algorithm and give greater importance to the classification error.
For high C the margins of the SVM model will be narrower because they are determined by points closer to the separate line and therefore fewer errors in classification will be allowed. In this situation there is a tendency to over-fitting to the training data. When C is low, we get a wider margin because they are determined by more distant points and therefore more errors in classification will be allowed.
As stated above, graph A corresponds to the case of lower C, since it is known that the margins are equal on both sides and I see two points that are on the separate line i.e., the points are within the margin range. Therefor the margins were determined by more distant points i.e., a smaller C. And the opposite for D graph that corresponding to a higher C, so that the margins are determined by closer points and thus there should be fewer classification errors (this cannot be seen in the given graphs).
In conclusion: $1 \rightarrow A \quad and \quad 2 \rightarrow D$

RBF kernel- For this type I expected to obtain circular separate lines as can be seen in graphs E and B. The separate lines are determined according to two parameters which are variance and μ.

$$k(x,z) = \exp\left(-\|x-z\|^2\gamma\right)$$

- $\gamma = 1/(2\sigma^2)$

According to what is learned in the lectures, gamma is the opposite of variance. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'.
That is, the smaller the gamma, the variance of the Gaussian is greater and therefore we get a less specific classification of the data which corresponds to graph E. whereas when the gamma is large the variance of the Gaussian is smaller, and the separation obtained is more specific to the data and therefore corresponds to graph B.

In conclusion: $5 \rightarrow E \quad and \quad 6 \rightarrow B$

Polynomial kernel- The higher the degree of polynomial, the more complex separate lines we will get.
Graphs F and C remain therefore, graph F is suitable for a 10^{th} order polynomial because the high complexity and its shape Fits precisely to the data. So, if I put new data into the model there is a chance that it will not classify them correctly because the overfitting. Graph C corresponds to 2^{th} order polynomial, much smaller complexity, and less specific fit to the data.
In conclusion: $3 \rightarrow C \quad and \quad 4 \rightarrow F$

3. Capability of generalization

a. The scientific term of the balance that Albert Einstein meant to in machine learning aspect is **Generalization:** is a measure of how the model performs on predicting unseen data.
According to Einstein everything should be simple but not the simplest, i.e. in the case of a learning algorithm, one that it too simple will not yield good results and on the other hand one that it too complicated will result over-fitting to the training data and therefore there will be no generalization to new data.

b. $AIC = 2p - 2\ln(\hat{L})$

p- is the total number of learned parameters.

$\hat{L}$- the estimated likelihood given these parameters.

AIC is a criterion for determining the quality of the model we have built in GMM, compared to a model with other parameters (number of clusters, etc). And it deals with the trade-off between a good fit of the model and the simplicity of the model, basically what generalization try to do. The best model will be the one with the lowest AIC.
The term $2\ln(\hat{L})$ affects on generalization, our goal is to maximize the term. The larger $\ln(\hat{L})$ we get a better fit of the model to the data, the probability to belong to specific cluster is higher and the probabilities to belong to the other clusters are lower. So larger L the better the fit of the model and therefore for a very high expression we get over-fitting.
To maintain the capability of generalization of the algorithm there is the expression 2P which is a type of punishment factor to prevent over-fitting. That is, higher P (the higher the number of clusters in the model) causes an increase in AIC.
In conclusion, a large number of clusters (parameters for learning the model) high likelihood function, detracts from the generalization ability of the model.

c. The two options that are likely to happen if this balance was violated are overfitting to the training data or underfitting.
Underfitting meaning that the model is too general and doesn't fit specifically

to the type of data we are working with and therefore we will get unsatisfactory classification results.

d. We are aiming to the lowest AIC. Low AIC means that the $\hat{L}$ is high so we are in a good fit but the number of clusters that was selected is not too high or too low to avoid the cases we talked about in Section c.