HW3 – theoretical part

Submitting: Rakefet Rozen

1. Clustering

a. The K-medoid is more robust to noise and outliers. The K-medoid loss function minimizes the pair wised dissimilarity – the median sample of the set. While k-means minimizes the square of the mean Euclidian distances – the mean value of the set. This results in the K-means case that the noise and outlier will affect the mean to move toward the outliers. In the k-medoid the median value is not affected by outliers for it is chosen in the middles of the data set position and not influenced from values at the edges of the set. So, in most cases when the number of outliers and noise is neglectable compared to the regular population the median will not be influenced from noise and outliers.

b. Let $f(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2$, we are looking for the minimum of f
To find the minimum we will compare the derivative of f to zero:

$$f'^{(\mu)} = -\sum_{i=1}^{n} 2(x_i - \mu) = 0$$

$$-\sum_{i=1}^{n}(x_i) + n\mu = 0$$

$$n\mu = \sum_{i=1}^{n}(x_i)$$

$$\mu = \frac{\sum_{i=1}^{n}(x_i)}{n}$$

Which is the mean value of the array x =$(x_1, x_{21}, \dots x_n)$

Bonus:
Let $f(\mu) = \sum_{i=1}^{n}|(x_i - \mu)|$ and let $K_1 = \{k \in (1,2, \dots n) | x_k > \mu\}, K_2 = \{k \in (1,2, \dots n) | x_k \leq \mu\}$, So,

$$f(\mu) = \sum_{i=1}^{n}|(x_i - \mu)| = \sum_{i \in K_1}(x_i - \mu) + \sum_{i \in K_2}(\mu - x_i)$$

$$f'(\mu) = \sum_{i \in K_1}(-1) + \sum_{i \in K_2} 1 = size\ of(K_2) - size\ of(K_1)$$

To find the minimum we will compare the derivative of f to zero:

$$f'(\mu) = number\ of\ samples\ smaller\ than\ \mu - number\ of\ samples\ bigger\ than\ \mu = 0$$
$$number\ of\ samples\ smaller\ than\ \mu = number\ of\ samples\ bigger\ than\ \mu$$

The last is the definition of a median when the median is part of the group

2. SVM

1. Linear kernel with C = 0:01.

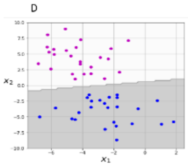
Figure D. Explanation: the hyper plane is linear. But the x dimension hyperplane distances from the points is higher than in option A. Meaning the regularization of the weights is lower

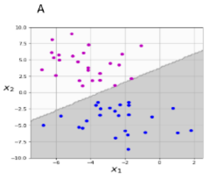
2. Linear kernel with C = 1.

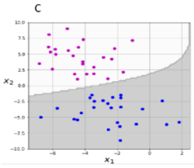
Figure A. Explanation: the hyper plane is linear. The x dimension hyperplane distances from the points is smaller than in option D. Meaning the regularization of the weights is higher

3. 2nd order polynomial kernel.

Figure C. Explanation: the hyper plane is the only one of second order







4. 10th order polynomial kernel.

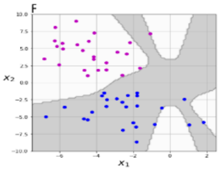
Figure F. Explanation: hyperplanes E and B seem to have close or almost close boundary which is typical to an RBF kernel. Figure F could be a 10th degree polynomial

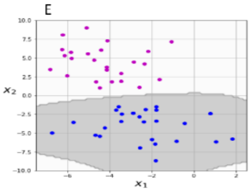
5. RBF kernel with γ = 0:2.

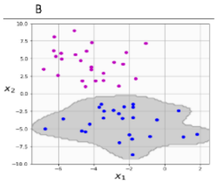
Figure E. The hyper plane has an almost close boundary which is typical to an RBF kernel. The gaussians here has a large spread which fits a small gamma

6. RBF kernel with γ = 1.

Figure B. The hyper plane has a close boundary which is typical to an RBF kernel. The gaussians here has a small spread which fits a large gamma







3. Capability of generalization

a. The term in machine learning that describes Einstein's quote "Everything should be made as simple as possible but not simpler" is Generalization. It is usually controlled by regularization. You want your model to fit the data well but not over fit. Over fitting will create a more complex model that fits the train set but will not be general enough for a new observation. By the proper regularization terms, the model will be adjusted to have a simpler representation to avoid over fitting and on the other hand does not give a too simple representation and avoid under fitting. Adjusting these parameters are exactly that threshold between making the model simple but not too simple.

b. The "Akaike information criterion" $AIC = 2k - 2\ln(\hat{L})$
The term $-2\ln(\hat{L})$ reflect the goodness of fit, encourage overfitting the model. This is caused because the nature of the likelihood function. By overfitting the model, the likelihood of detection reaches 1 and so the term goes the zero. For any value smaller than 1 the term increases. On the other hand, the term $2k$ is related to underfitting. It gives a penalty to the number of estimated parameters k.

c. The two options that might happen if one of these terms are violated are:
Overfitting if the term - $2\ln(\hat{L})$ is too low and $2k$ is too high
Underfitting if the term $2k$ is too low and - $2\ln(\hat{L})$ is too high

d. We are aiming for the $AIC = 2k - 2\ln(\hat{L})$ to be as small as possible letting - $2\ln(\hat{L})$ term take care of the complexity of the model (overfitting) and letting the second term $2k$ regulating that complexity be the penalty it creates for a complex model (underfitting).