




Machine Learning in Healthcare - HW3

Lior Drasinover - 206262099

1. Question 1

- a.  K-medoids is more robust to outliers than k-means because of the way it minimizes the sum of dissimilarities between 2 points (the centroid and another point) instead of squared Euclidean distances.
- b. First I will simplify the term

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



The term will be minimal for the μ that brings the derivation according to μ to 0:



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

$$\mu = \frac{1}{m} * \sum_{i=1}^m x_i = \text{mean}(X)$$

2. Question 2

- a. We can see that the 2 groups are separated by a straight line, this means we can assume that the kernel is linear, we can also see that the margin between the groups contains points and because of that we will assume C is smaller because smaller c gives more room for errors to the classifier and because of that **A- is 1- linear with C=0.01**
- b. We can see that the classification is a closed shape and because of that we will assume it is a RBF kernel, we can see that the shape is smaller than it is in E, γ controls how "far" is the influence and because of that γ is bigger, which means that **B is 6- kernel RBF with $\gamma = 1$** .
- c. We can see that the line separating the 2 groups is parabolic and because of that we will assume the kernel is a second degree polynomial which means **C is 3 -2nd order polynomial kernel**.
- d. We can see that the 2 groups are separated by a straight line, this means we can assume that the kernel is linear, we can also see that the margin between the 2 groups is bigger and contains no data points because of that we will assume C is larger and because of that **D- is 2- linear with C=1**
- e. We can see that the classification is a closed shape and because of that we will assume it is a RBF kernel, we can see that the shape is bigger than it is in B and we can't see the whole shape (it's more circular than F so I assume it's circular), γ controls how "far" is the influence and because of that γ is bigger, which means that **E is 5- RBF kernel with $\gamma = 0.2$** .
- f. We can see that the line here is not linear and not circular, so I assume it's polynomial, we can also see that it's not an order that I know and looks more complex, because of that I assume that **F is 4 - 10th order polynomial kernel**.

3. Question 3

- a. The term is Generalization, a generalized model balances the goodness of the fit with the complexity, this agrees with Einstein because we want the model to be as simple as we can but also to have good performances that satisfy our goal.

- b. P is the number of learned parameters and represents the model's complexity, $\log(L)$ represents the model's performance. We choose a model according to the minimal AIC, a more complex model and lower
 performance, which is smaller likelihood (likelihood is between 0-1) will give us a bigger AIC. This means that less parameters will give us better AIC score and better performance will also give us a better AIC score, but in a more significant way than the complexity because of the logarithmic nature. Because optimal AIC score is minimal, the complexity and performance both need to be balanced through the p and $\log(L)$.
- c. The 2 options are overfitting and underfitting:
Underfitting – if we choose a model with a small number of learned parameters (low complexity) it can lead to lower accuracy and lower performance, this is underfitting.
Overfitting – if the model is too complex then the performance might be better for the training set but too specific and lead to misclassification of new data, which means that the model was overfitted.
- d. We want a minimal AIC score. We want a balanced model that is not too complex (decreases p) yet still has a good performance (decreases $-\log(L)$) which means that we will get a small AIC.
It is also important to notice that because AIC score is more influenced by the performance (because of the nature of the logarithm) it prefers to give us a model with high performance even if it increases the complexity a little.