




Machine Learning in Healthcare – 336546

HW-3

Tania Assaf

I.D: 208655951

1) Clustering

- a. Both K-means and K-medoid are clustering algorithms that divides our data points into several groups or clusters by finding the minimal Euclidean distance from a certain value ($d(p, q) = \sum_i^n (q_i - p_i)^2$). 

K-means minimizes the Euclidean distance between our data and a point in space (not from the data) called a centroid. Thus, K-means relies on the mean value of the data in each cluster.

In K-medoids the Euclidean metric is minimized between some data and a data point called medoid (the medoid is a point from our data). Thus, K-medoid relies on the median value of each cluster.

As we know, median value is more robust to outliers than the mean value. So, we can conclude that K-medoid is more robust to noise and outliers than K-means.

- b. We need to prove that in the 1D case ($x \in R^1$) the centroid μ which minimizes the term $\sum_{i=1}^m (x_i - \mu)^2$ to is the mean of m examples.

The definition of mean value of m examples is: $\mu = \frac{\sum_{i=1}^m x_i}{m}$

On the other hand, to find a critical point of the term (minimum or maximum) we need to calculate the derivative and equal it to 0 :

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

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

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

To prove that this critical point is a minimum, we calculate the second derivative:

$$\frac{d}{d\mu} \left(-2 \sum_{i=1}^m x_i + 2m\mu \right) = 2m > 0 \rightarrow \mu \text{ is the minimum}$$

In conclusion, μ that brings the term to its minimum equals to $\boxed{\mu = \frac{\sum_{i=1}^m x_i}{m}}$ (the mean value of m examples.)

c. **Bonus:**

We need to prove that the centroid (medoid) μ which minimizes the term $\sum_{i=1}^m |x_i - \mu|$ to is the median of m examples.

The definition of median value of m examples that it is the value which the number of values above it (including it or not) are equal to the number of values under it (including it or not).

On the other hand, to find a critical point of the term (minimum or maximum) we need to calculate the derivative and equal it to 0:

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

As the second derivate is positive, the critical point we found is the minimum value.

This derivative is the sign function around μ . from sign's function description, the number of values above μ equals to number of values under μ which also means that μ is the median value of those points. (This is true both in odd and even number of values due to the median's deffention).

2) SVM

A-1: In figure A there is a linear separation line. linear kernels are characterized by a linear line separation in the features plane. C states the penalty on margin violation, low C means low penalty and low restriction on points who will be in the margin area. In figure A there are two purple points near the (almost on the line) linear separation line and there are not blue points at the same distance as the two-purple point on the other side of the line. This means that the support vectors are before those two purple dots and that these dots are in the margin area which means that C is small (the smallest between $C=1$ and $C=0.01$).

→ **A : 1) linear kernel with $C = 0.01$**

B-6: in RBF kernels data is projected to higher dimension planes and separated there by separation plane that intersect multidimensional Gaussian. After projecting (back) to a lower dimension plane we get a closed separation shape (circle like). Higher Gamma leads to narrow Gaussian. In figure B we can see the shape's limits, thus the shape in B is narrower than the shape in D and belongs to the higher gamma. → **B : 6) RBF kernel with $\gamma = 1$**

C-3: In polynomial separation, the separation line is described by a polynomial in the features plane. In figure C there is a polynomial separation kernel. As known, the separation line in C is a second order's polynomial line. Thus, the data in figure C is separated by a second order polynomial. → **C : 3) 2nd order polynomial kernel**

D-2: In figure D, the data is separate by a linear line. Which means that the data is separated by linear kernel. In addition, there are not dots that are close to the separation line thus as explained above (in figure A) C parameter is higher than the C parameter in figure A , which means that the penalty restriction on the margin area is higher in D and there are no (or less) points in the margin area .(explanation on C parameter and linear kernel in A) →

D: 2) linear kernel with $C = 1$

E-5: From the explanation in figure B, the Data in figure E is also separated by an RBF kernel. Lower gamma leads to wider gaussian. In figure E, after we project back to a lower dimension, we cannot see the shape's limits which means that the gaussian in figure E is wider than in figure B and gamma is lower. → **E : 5) RBF kernel with $\gamma = 0.2$**

F-4: as explained in figure C, in polynomial separation, the separation line is described by a polynomial in the features plane. Higher polynomial orders get's more complicated, thus in figure F the polynomial line is more complicated than the polynomial line in see and belongs to a high polynomial order function. This means that the data in figure F is separated by a high order polynomial kernel. → **F : 3) 10th order polynomial kernel**

3) Capability of generalization

- a. The scientific term of the balance that Einstein meant to in machine learning aspect is Generalization which means the ability to handle unseen data. To be more specific, the terms that determine how simple our system are the model's complexity and goodness of fit. So, we want to balance between the model's complexity and goodness of fit thus we will not get high complexity, but we still get a good fitness for the model.
- b. AIC is calculated by: $AIC = 2p - 2\ln(\hat{L})$. $2p$ is used as a penalty for the AIC, it increases when the number of learned parameters is higher. This leads to higher AIC and increases the model's complexity and reduces the overfitting. \hat{L} is the estimated likelihood given these parameters, higher \hat{L} means higher goodness of fit, thus high \hat{L} leads to lower AIC ($-2\ln(\hat{L})$).
- c. If the balance in generalization is violated; meaning balance between the model's complexity and the goodness of fitting this imbalance, may lead to overfitting or to underfitting. which leads to poor performance of our model.
- d. AIC identifies an estimator of out-of-sample prediction error. And it is calculated by: $AIC = 2p - 2\ln(\hat{L})$. Thus, the preferred model is the model with the minimum AIC value. But, we also aim for a low value of AIC that keeps the balance explained above.