



HW 3

Q 1. (a) Yes, K-mediod is more robust to noise (and outliers) than the K-means. Actually, the K-mediod's algorithm takes a random K-set of medoids from the dataset itself, and matches points to each one based on their L_1 distance to that specific medoids. So, it minimizes the distance from a point that is part of our sample. While, K-mean minimizes the L_1 distance to zero by selecting a point in a space which is not necessary part of the sample (a calculated mean - not necessarily find in the sample). As we know, the median is more robust to noise than mean. To sum, in case of outliers/noise, K-mediod is more stable, K-mean try to be close to a number that is very affected from outliers/noise wich is the mean, so it is not robust to noise/outliers.

1. (b) In 1D ($x \in \mathbb{R}^1$) of k-means, the centroid (μ)

which minimize $f(x) = \sum_{i=1}^m (x_i - \mu)^2$ is the μ which

gives us $f'(x) = 0$.

$$\Rightarrow f'(x) = \sum_{i=1}^m 2(x_i - \mu)$$

$$f'(x) = 0 \Rightarrow 0 = 2 \left[\sum_{i=1}^m x_i - \sum_{i=1}^m \mu \right] \quad | :2 \Rightarrow 0 = \sum_{i=1}^m x_i - m \cdot \mu$$

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



which is the mean of m samples.

Q2. * Let's start with linear kernel. Only A & D can be describe by linear kernel. When C is large the margin is low, and when C is low the margin is large. So, A is more suitable to have $C=0.01$, because, it has a larger margin (the support vectors is farther than in D).

$$\Rightarrow \begin{cases} 1 \rightarrow A \\ 2 \rightarrow D \end{cases}$$

⊕ Now, for the case of polynomial kernel, C & F is more suitable to polynomial kernel, (they have 2 unlimited regions). The more simple one, C, is suitable to 2nd order polynomial kernel, and the more complex one is suitable to 10th order polynomial kernel.

$$\Rightarrow \begin{cases} 3 \rightarrow C \\ 4 \rightarrow F \end{cases}$$

* RBF Kernel is the Gaussian Kernel, in 2D we will see 2 regions, one is limited the other one is above her

So, E & B is suitable to Gaussian (RBF) Kernel.

The parameter γ is associated with the limited region's "diameter". Large γ is suitable to a narrow closed (limited) region, and low γ is for large closed (limited) region. So, B is suitable to large γ ($\gamma=1$), and E is for the low ($\gamma=0.2$)

$$\Rightarrow \begin{cases} 5 \rightarrow E \\ 6 \rightarrow B \end{cases}$$

Q3. (a) Generalization is the scientific name, the aim of it is to get as simple as possible model but to be able to generalize it to classify new and larger data. So it matches what Einstein meant to in his say.

(b) $2 \ln(\hat{L})$: The larger the \hat{L} (estimated likelihood) is the larger goodness of fit we get, and vice versa. It is good for the generalization to have high goodness of fit but not too much high. Because, at some level, high goodness of fit may cause overfitting, which we aim to avoid. " \ln " is an increasing function, so what I explained is true also for $2 \ln(\hat{L})$.

constant

increasing function

$2\hat{P}$: \hat{P} is the number of learned parameters, so the larger the \hat{P} is the more the complexity is and vice versa.


\Leftarrow

(Q3 b continuation)

Complexity is not a desired thing in generalization, but on the other hand, and as we see in the AIC equation, large number of \hat{p} will compensate a large \hat{L} . So it may increase the complexity but decrease the overfitting.

We can conclude that we should find a balance between these parameters to get a desired result.

(c) The two options are the following:-

1. Over fitting -(when \hat{L} is large)
2. Under fitting -(when \hat{P} is large) 

(d) The AIC should be low

Explanation: AIC actually estimates the relative amount of information that were lost by a given model. So, if the AIC is high the model loses a lot of information, and that is not a desired thing. So

we want a low AIC, which means that our model loses less information, which increases the quality of our model.