

1.

- a. K-medoid is more robust to noise and outliers because it uses an actual data point as the cluster center rather than the cluster mean – which may be very far away from the actual cluster center when using data with large outliers, furthermore K-medoid allows for using other evaluation metrics rather than the Euclidean distance, which may be better for clustering (for example – using some feature correlation between 2 data points).
- b. We can find the minimum of the term by derivation and equating to zero:

②

$$b. \quad L(\mu) = \sum_{i=1}^m (x_i - \mu)^2$$
$$\frac{dL(\mu)}{d\mu} = -2 \cdot \sum_{i=1}^m (x_i - \mu)$$
$$= -2 \left[\sum_{i=1}^m x_i - m\mu \right] = 0$$
$$\Rightarrow \sum_{i=1}^m x_i = m\mu \Rightarrow \mu = \frac{1}{m} \sum_{i=1}^m x_i$$

Bouns : *** attached at the end of document ***

2.

A – 1

D – 2

C – 3

E – 4

B – 5

F – 6

Explanation: A & D are the linear kernels therefore correspond to labels 1 & 2, we can see that in A the decision boundary passes through 2 of the data points and in D it doesn't touch any data point, therefore we understand that A's SVM had more "slack" in its training meaning lower higher impact of ξ therefore lower C.

B & F are the RBF kernels because of the decision boundary shape which can only be achieved using radial functions, F is overfitted thus has the higher Gamma.

C & E are the polynomial kernels, C is the 2nd order because its decision boundary is a parabola, therefore E is the 10th must be the 10th order.

3.

- a. The scientific term referred to is bias-variance tradeoff.
- b. The $\ln(L)$ term correlates to the bias as it states how well does our data fit with our model given a set of chosen parameters, the P term refers to the variance – because it counts the number of parameters in our model.
- c. Violation of the bias-variance tradeoff can result in either overfitting or underfitting.
- d. We want to minimize the AIC because we want best fit for our model (meaning high $\ln(L)$) while having the simplest model – which means fewer parameters as possible – meaning low P.

Bonus:

$$L(\mu) = \sum_{i=1}^m (|x_i - \mu|)$$

If we take the first and last terms out of the sum we get:

$$L(\mu) = \sum_{i=2}^{m-1} (|x_i - \mu|) + |x_n - \mu| + |x_1 - \mu|$$

$$= \sum_{i=2}^{m-1} (|x_i - \mu|) + (x_n - \mu) - (x_1 - \mu)$$

$$= \sum_{i=2}^{m-1} (|x_i - \mu|) + (x_n - x_1)$$

we repeat this process until we have 1 or 2 elements left in the sum (depending on m if it's odd or even)

if m is odd then we get:

$$L(\mu) = |x_{\frac{m+1}{2}} - \mu| + (S_n - S_1) + (S_{n-1} - S_2) + \dots$$

$$= |x_{\frac{m+1}{2}} - \mu| + \text{Const}$$

In this case the minimum is given

by $\hat{\mu} = x_{\frac{n+1}{2}}$ which is the group median

If n is even we get:

$$L(\mu) = |x_{\frac{n}{2}} - \mu| + |x_{\frac{n+2}{2}} - \mu| + \text{const.}$$

In this case the minimum is received
for the mean of the two samples:

$$\hat{\mu} = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$$

which is the median
as well

Finally we get:

$$\hat{\mu} = \begin{cases} x_{\frac{n+1}{2}} & , n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2} & , n \text{ is even} \end{cases}$$