

1. Clustering (10%)

a) Yes! K-medoids algorithm is more robust to noise (or outliers) because it represents an actual point in the cluster, while K-means takes 'mean point' as a center of clustering, in other words it minimizes the "Sum of squared Euclidean distance" and not the Sum of general pairs of differences as K-medoids does, so we can say that K-mean is less robust to noise or outliers, it's more influenced by them, & can't correctly represent the center of the cluster as K-medoids does.

b) To prove that the centroid (μ) which minimizes the term $\sum_{i=1}^m (x_i - \mu)^2$ is the mean of examples, we start by ~~setting~~ ^{Setting}

$\frac{d}{dx} [\sum \dots]$ equal to 0.

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

$$2 \cdot \sum_{i=1}^m (x_i - \mu) = 0 \quad / :2$$

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

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

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

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

we proved that the centroid μ = mean of m examples $\left[\frac{\sum}{m} \right]$ means adding up all the x_i 's and dividing by how many numbers there are (m) = average = mean

$$\frac{d^2}{dx^2} [\sum \dots] = 2m > 0 \rightarrow \text{minimum}$$

2. SVM (30%)

① **Linear kernel** for images A and D. When we have a low value of C , the margins are wider and more errors will be obtained in the Data classification of the training. On the other hand, for a large value of C the margins will be narrow and fewer errors will be obtained, but we may have an overfitted model.

We can see in figure A that there are two points on the border of decision, so we can assume that A is for a low value of $C \Rightarrow$ So:

$$A \rightarrow C = 0.01$$

$$D \rightarrow C = 1$$

* C is the hyperparameter which balances between the margin widening around the decision boundary and between the number of the errors obtained in the Data classification.

② **BBF** for images B and E, we can see their "Radial shaped line". B is more overfitted than E, so B will have the bigger value of γ , so we can tell that \Rightarrow

$$B \rightarrow \gamma = 1$$

$$E \rightarrow \gamma = 0.01$$

* γ controls the influence of new features; so, bigger γ means more overfitted model.

③ Polynomial Kernel for images C and F, we can see that image C is similar to parabola - Polynomial Kernel with ~~low~~ degree = 2 (Lower value), because ~~Lower~~ ^{Higher} degree means more flexible decision boundary, and the nuclei may be over fitted - this case matches figure F) ~~So~~ So F has a greater tendency to become over fitted (it's more complex shaped than C)

* The degree parameter controls the flexibility of the boundary of decision.

3. Capability of generalization

- a) In machine learning, This 'balance' describes the balance between performance and the complexity of the model. According to Einstein, to make a model with good Performing we need to have enough complexity in our model, the term for this balance is generalization.

b)

$$AIC = 2 \cdot \underbrace{p}_{\substack{\text{complexity of model} \\ \text{number of learned parameters}}} - 2 \ln \left(\underbrace{\hat{L}}_{\substack{(-\infty, 0) \\ \text{estimated likelihood} \\ (0-1) \\ \text{*Performance of model}}} \right)$$

- The bigger $2p$ is → the bigger the complexity → the bigger AIC.
→ The bigger \hat{L} → the lower AIC.

c)

Option	Explaining
① overfitting	High value of 'p' → high complexity → a risk of overfitting!
② underfitting	low value of 'p' → low complexity → low Performance → not accurated

* In case of overfitting (option 1), even if we have a good performance with a dataset, high complexity would prevent the model from working well in another dataset.

- d) If we wanna have a simple model with good performance we would minimize ~~the~~ AIC, as we explained before, minimizing AIC means getting a model with low complexity ($2p$) and high performance (\hat{L} or $2 \ln(\hat{L})$).