Aseel Nama
319112272

- Machine learning in Healthcare - Hw. 3

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 another 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 ~~see~~ say that K-mean is less robust to noise or outliers, its more influanced 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 $\frac{\partial}{\partial x}[\Sigma..]$ equal to 0.

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

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

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

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

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

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

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

$\frac{\partial^2}{\partial x^2}[\Sigma..]=2m>0 \rightarrow$ 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 ⇒ So:

A → C = 0.01

D → 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 RBF 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 ⇒

B → γ = 1

E → γ = 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 prabola – polynomial kernel with degree = 2 (Lower value), because Higher degree means more flexible decision boundary, and the nuclei may be over fitted – this case matches figure F)

So F has agreater tendency to become overfitted / 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 Einsten, to make a model with good Performing we need to have enough complexity in our model, the term for this balance is generalization.

b) complexity of model $(-\infty, 0)$

$$AIC = 2p - 2\ln(\hat{L})$$

- p: number of learned parameters
- $\hat{L}$: estimated likelihood $(0-1)$ * Performance of model

→ 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 $[2\ln(\hat{L})]$.