Answers of the theoretical questions:

Elinoy Faibish 205738230

Q1:

a. Yes. Although K-medoid is clustering algorithm that similar to K-mean because both of them breaking the dataset up into groups. The K-medoid is more robust to noise, because it uses the actual point representative objects as reference points instead of using the mean value of the objects in a cluster as a reference point and k- mean is more sensitive because the mean is easily influenced by extreme values. i.e. K-medoid basically minimizes the sum of general pairwise dissimilarities, instead of minimizing the total squared error (like K-mean).

b. We want to minimize the following expression:

$f(\mu) = \sum_{i=1}^{m}(x_i - \mu)^2$ so we will derivate the expression and equal it to 0

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

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

$$* \frac{\sum_{i=1}^{m} x_i}{m} = mean\ of\ the\ examples\ (x_i)$$

We need to show that the function is convex, i.e. the second derivation is > 0:

$$\frac{d}{d\mu}\left(-2 \cdot \left(\sum_{i=1}^{m} x_i + \sum_{i=1}^{m} \mu\right)\right) = \frac{d}{d\mu}\left(2 \cdot \underbrace{\sum_{i=1}^{m} \mu}_{m \cdot \mu}\right) = 2m > 0$$

Bonus:

We want to minimize the following expression:

$f(\mu) = \sum_{i=1}^{m}|x_i - \mu|$ so we will derivate the expression and equal it to 0

$$\frac{d}{d\mu} f(\mu) = \sum_{i=1}^{m} \frac{d}{d\mu}|x_i - \mu| \underbrace{=}_{\frac{d\,|x|}{dx} = sign(x)} \sum_{i=1}^{m} \text{sign}(x_i - \mu) = 0$$

We notice that the expression is = 0 only when the number of positive samples are equal to the number of negative and it only happens when $\mu = median\ \{x_i\}_1^m$

*The median is not necessarily an item within the group.

Q2:

Figures	Settings	Explanation
A	1	The group are separated by a straight line so the kernel is linear. By definition the margins need to be equal, thus, some points are located inside the margins which mean we will neglect the outliers. And because of that C is probably smaller-> we will have more options to errors in the classifier
B	6	The classification is a closed shape a roughly circular. thus, it is a RBF kernel, the area of the shape is smaller than in E, which mean that B's variance is smaller so the γ is bigger $(\text{reminder: } \gamma = \frac{1}{2*(\sigma)^2})$
C	3	The group are separated by a parabolic line so the kernel is a second degree polynomial
D	2	The group are separated by a straight line so the kernel is linear. By definition the margins need to be equal, thus, there are no points located inside the margins which mean we won't neglect the outliers. And because of that C is probably bigger and the margins are smaller.
E	5	The classification is a closed shape a roughly circular. thus, it is a RBF kernel, the area of the shape is bigger than in B, which mean that E's variance is bigger so the γ is smaller $(\text{reminder: } \gamma = \frac{1}{2*(\sigma)^2})$
F	4	the line is not linear or circular. It's probably polynomial because the order isn't familiar and it looks more complex then C, it is a higher polynomial.

Q3:

a. Generalization. By that we will find the best trade-off between underfitting (A high-biased, low-variance) and overfitting (A low-biased, high-variance). Thus it agrees with Einstein because we want our model to be complexity (but not too much, simple as possible) and will yield good performance and Generalization is the term for that.

b. $\mathrm{AIC} = \overbrace{2\mathrm{p}}^{model's\ \text{complexity},} - \overbrace{2\ln(\hat{\mathrm{L}})}^{(-\infty,0)}$

$1. P - Number\ of\ learned\ parameters$

$2. log(L) - model's\ performance\ \ (0-1)$

↓

we would prefer to have the lowest AIC

AIC try to find the best trade-off between underfitting and overfitting like generalization. It is used to compare different models, estimate the quality of each model and determine the best model which is fit for the data.

When we increase the $2\mathrm{p}$-> we increase the complexity and AIC is increased

When we decrease the $\hat{L}$-> less number of parameters will bring us to better performance and AIC is increased

c. The 2 options are overfitting and underfitting:

Overfitting – High value of P which means the model is too complex. thus it's creating a bias towards the training dataset but it will lead to bad performances with a new data.

Underfitting- Low value of P which means the model is too simple. thus it will lead to lower performances and a lower accuracy.

d. We are aiming for the AIC to be lower if we want balance model that is less complex model (Low value of 'P') but still a good fit for the data (decreases log(L))

***I'm apologize if I have grammar mistakes**