Question 1

a. Is K-medoid more robust to noise (or outliers) than the K-means algorithm? Explain your answer.

The K-medoid is more robust to noise and outliers than the k-means
The K-medoid minimizes the sum of common paired dissimilarities instead while k-means minimizes the sum of the squared Euclidean distances. And this distance metric reduces noise and outliers.
In K-medoid Instead of using the mean point as the center of a cluster, *K*-medoids uses an actual point in the cluster.

b. Prove that for the 1D case ($x \in R^1$) of K-means, the centroid (μ) which minimizes the term $\sum_{i=1}^{m}(x_i-\mu)^2$ is the mean of *m* examples.

Let's do first derivative to the tem then to solve it when it's 0

$$\frac{\partial \sum_{i=1}^{m}\left(x_i-\mu\right)^2}{\partial\mu} = -2\sum_{i=1}^{m}\left(x_i-\mu\right) = -2\sum_{i=1}^{m}x_i + 2m\cdot\mu$$

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

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

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

Second derivative to make sure it's a minimum⇒ $2m > 0\ \ so\ it's\ a\ min$

c. Prove that for the 1D case ($x \in R^1$), the centroid (practically, the medioid) which minimizes the term is the median of *m* examples given that *μ* belongs to the dataset. You can add a verbal explanation to the cases where *m* is an even number if needed.

$$\frac{\partial \sum_{i=1}^{m}\left|x_i-\mu\right|}{\partial\mu} = \sum_{i=1}^{m} sign\left(x_i-\mu\right) = 0$$

The sum of signs is eqal to 0 when the number of positive items(+1) equals the number of negative items(-1).

when the number examples that are bigger than μ is equal to the number of examples than are smaller from μ ,so μ in this case is the median of the m examples .

Question 2

1. **Linear kernel with *C* = 0.01.**

A

2. **Linear kernel with *C* = 1.**

D

Explanation

Since we have a linear border so the it's linear kernel svm

C parameter in SVM is **Penalty parameter of the error term**,which is a tradeoff between error and margin

For **greater values of C→ high penalty to classification** , there is no missclassified points or points inside the margin(D) . But for **Smaller C**, points start moving inside margin as we can see on (A)

3. **2nd order polynomial kernel.**

C

4. **10th order polynomial kernel.**

F

Explanation

C+F describe polynomial kernels The decision boundary is polynomial

The higher the polynomial degree the more complex decision boundary we have .

Thus F has a higher polynomial degree than C

5. **RBF kernel with *γ* = 0.2.**

E

6. **RBF kernel with *γ* = 1.**

B

Both E and B are RBF since they are having n2qearly closed decision boundaries and close to circular shape .
We can look at γ as the 1/radius of the boundary .
The bigger the area of the closed boundary the smaller the γ is
That's why E (bigger area)is an RBF with smaller γ
And B (smaller area) is an RBF with bigger γ

Question 3

a. What is the scientific term of the balance that Einstein meant to in machine learning aspect?

The trade-off between bias and variance.

Bias reflects the complexity of the model how and the variance reflects the ability of our model to generalize .

When variance is high the model tries to fit to the training data and thus it’s ability to generalize and to predict on a new data decreases.(overfitting and more complex model)

when bias is high it means that there is high error between training data and predictions (less complex model) and here variance will be low because model won’t be able to find the relationship between observations and prediction.(underfitting and more complex model)

A balance between bias and variance is important to find the best model .

b. How does each of the terms $(2p,2ln(\hat{L}))$ in AIC affect the terms of the balance you defined in (a)?

p is the total number of learned parameters in the model, indicates model complexity.

The estimated likelihood of the model $\hat{L}$, measures the goodness of fit of the model.

higher the number of parameters $\longrightarrow$ more complex model

higher $\hat{L} \rightarrow$ better goodness of fit.

The AIC score rewards models that achieve a high goodness of fit score and penalizes them if they become so much complex to avoid overfitting

c. What are the two options that are likely to happen if this balance was vio-lated?

This can lead to overfitting or underfitting and thus affect the generalization of the model as explained in a .

if model is too simple (less parameters) without enough data for it to understand the relationship between predictions and observations the model will underfit.

If there is more parameters and data, the model gets more complex, and it tries to fit all the data points, meaning it learns patterns along with noise, which causes overfitting.

d. What are we aiming for with the AIC? Should it be high or low?

AIC should be low .We want the model to have a high log likelihood(goodness of fit of the model is better) but at the same and to penalize over the high to number of parameters to avoid overfitting .