1. Clustering

a) The K-means algorithm is sensitive to outliers, because **a mean is easily influenced by extreme values.** The K-medoids algorithm minimizes **a sum of pairwise dissimilarities** instead of a sum of squared Euclidean distances, so it is more robust to noise and outliers than K-means.

b) The expression we will search to minimize:

$$\boldsymbol{argmin} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} (\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{2}}$$

We will solve this by derivation and comparison to 0:

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

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

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

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

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

Which is the mean of m examples, by definition.

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

The value m being the number of examples (>0), the second derivative is obviously positive. Thus, this is a minimum.

Bonus :

Given that μ belongs to the dataset, we will cut the sum on his index k ($\boldsymbol{\mu} = \boldsymbol{x_k}$). Before the index k, the xi will be inferior to μ, after the index k the xi will be superior to μ. Knowing that we can simplify the expression:

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

$$\sum_{i=1}^{k} (\mu - x_i)\,, where\ xi < \mu$$

$$\sum_{i=k+1}^{m} (x_i - \mu), \qquad where\ xi > \ \mu$$

$$\sum_{i=1}^{k} (\mu - x_i) + \sum_{i=k+1}^{m} (x_i - \mu) = \ \boldsymbol{\mu k} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} (-\boldsymbol{x_i}) + \sum_{\boldsymbol{i=k+1}}^{\boldsymbol{m}} \boldsymbol{x_i} - \boldsymbol{\mu}(\boldsymbol{m} - \boldsymbol{k})$$

We can choose μ, k, m in order for this expression to be minimal.

$$\mathrm{\mu k - \mu(m - k) = 0}$$

$$\mathrm{\mu(2k - m) = 0}$$

$$\mathrm{(2k - m) = 0}$$

$$\mathbf{k = \frac{m}{2}}$$

The index k being on m/2, μ is the median value.

2. SVM

Linear kernels

The C parameter tells **how much we want to avoid misclassification.** For large values of C, we will choose a smaller margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, for a very small value of C, we will choose a larger margin hyperplane, even if that hyperplane misclassifies some points or admits some points inside the margins.
In the image A, the hyperplane prioritizes a large margin, even if that hyperplane allows data points between the margin.

A: Linear kernel with C = 0.01
D: Linear kernel with C = 1

Polynomial kernels

The higher is the dimension of the polynomial kernel, the more the hyperplane will have a complex shape.

C: 2nd order polynomial kernel
F: 10th order polynomial kernel

RBF kernels

For a high gamma, the model would consider only the points close to the hyperplane for modeling. Conversely, for a low gamma, the model would consider only the points far to the hyperplane for modeling. So, the higher is gamma, the more we fit to the training data.

B: RBF kernel with $\gamma = 1$
E: RBF kernel with $\gamma = 0.2$

3. Capability of generalization

a) The scientific term that Einstein meant to in machine learning aspect is **generalization**. Generalization refers to your model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model. A proper generalized model **deals with the balance between goodness-of-fit and the simplicity of the model.**

b) p : the number of estimated parameters in the model.
L : the maximum value of the likelihood function for the model.
AIC **rewards goodness of fit** (as assessed by the likelihood function) and it also includes a penalty that is an increasing function of the number of estimated parameters **(penalty for a too high complexity).**

c) The two options that are likely to happen if this balance is violated are underfitting/overfitting.
On the one hand, too much goodness of fit and a too high number of estimated parameters in the model will lead to an **overfitting.** One the other hand, a too simple model will be **underfitted**. Thus, it is important to preserve the balance.

d) AIC estimates the quality of each model. Thus, AIC provides a means for model selection. AIC deals with both the risk of overfitting (penalty for high number of parameter) and the risk of underfitting (use of likelihood function that assure a good fit). Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.