HW3 Answers:

1.

a. K-medoid is more robust to noise than K-mean.
K-mean looks for a point in the center of the data while giving all the point the same weight, meaning that noise can really shift the center of the data.
K-medoid uses a point that is in the data which the distances from that point to the other points is minimized which is less affected by noise (since the distance between the other points to the mediod hasn't changed because of the noise)

b. We want to minimize the expression: $\sum_{i=1}^{m}(x_i - \mu)^2$
We can also write the expression like this:
$\sum_{i=1}^{m}(x_i - \mu)^2 = \sum_{i=1}^{m}((x_i - c)^2 + 2(x_i - c)(c - \mu) + (c - \mu)^2)$
In the term $\sum_{i=1}^{m} 2(x_i - c)(c - \mu)$, $2(c - \mu)$ is not dependent on the index i, therefore we can write it like this: $2(c - \mu)\sum_{i=1}^{m}(x_i - c)$
In order to minimize the initial equation, we want this sum to be equal to zero:
$2(c - \mu)\sum_{i=1}^{m}(x_i - c) = 0 \ => \ \sum_{i=1}^{m}(x_i - c) = 0 \ => \ c = \frac{\sum_{i=1}^{m} x_i}{m} = \bar{x}$
The term $\sum_{i=1}^{m}(c - \mu)^2$ is not dependent on the index I, so we can write it as $m(c - \mu)^2$
Again, in order to minimize the initial equation, we want this term to be equal to zero:
$m(c - \mu)^2 = 0 => \ c = \mu => \ \mu = \bar{x}$
In conclusion: $\sum_{i=1}^{m}(x_i - \mu)^2 = \sum_{i=1}^{m}(x_i - \bar{x})^2$

2.

A->1, D->2
Using the C parameter and linear SMV classifier, therefore images A and D.
The C parameter balances the goals of increase the distance of decision boundary to classes and maximize the number of points that are correctly classified. Models with low C value tend to be more generalized and have more misclassification. A have 2 dots near/on the separator line, therefore its C value is lower than D's.

C->5, F->6
RBF decides the classification of a new data based to its distance from the training data. The Gamma parameter controls the similarity radius. Low gamma means a large similarity radius, therefore more points being group together. Low gamma looks similar to linear SMV, therefore image C goes with the lower gamma value.
High gamma requires the points to be very close for them to be in the same class. Model with large gamma values usually overfit, therefore image F goes with the high gamma value.

B->4, E->3
Polynomial Kernel takes the data to a higher dimension in order to linearly separate the data, therefore the separation looks round in compare to other separations. The higher the order, the more fitted the separation will be to data, therefore B has the higher order and E the lower.

3.

a. The scientific term of the balance in machine learning that Einstein meant to is the balance between generalization and fitting the data. We want the model to be simple enough to handle new data (in other word, to not be overfitting), but if it will be to simple, it won't fit the data well (underfitting).

b. 2ln(Lˆ) is a measure of how the model fits the data. The higher the log-likelihood, the best the model fits the data.
2p is a measure of the generalization of the model. The higher the learned parameters are, the less general the model is.

c. If there is no balance in the model, the model will either be overfitted (high fit, low generalization), or underfitted (low fit, high generalization).

d. Since we want to have the best fitted model using the minimum number of parameters (so we won't have overfit), we are aiming to low value of AIC.