HW3

Guy Hamburger

Q1)

a) K-medoid is more robust to noise and outliners than K-means. This is because the affect Euclidian distance has over outliners (if some point is really far the "error" value we will get is increased by power of 2. We calculate distance using l2). But in K-medoid we calculate the distance using l1 which means an error will get only the medoid error (absolute value and not increased by power of 2)

b)

$$E(\mu) = \sum_{i=1}^{m} (x_i - \mu)^2 \overset{min}{\Longrightarrow}$$

$$Min\big(E(\mu)\big) \rightarrow 2 * \sum_{i=1}^{m} (x_i - \mu) * (-1) = 0$$

We want to find the min point. we need to prove that is a Min point and not Max

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

$$\rightarrow \sum_{i=1}^{m} (x_i) = \sum_{i=1}^{m} (\mu) = m * \mu$$

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

The Equation we have is the mean distribution of Xi.

Since the expression $\sum_{i=1}^{m} (x_i - \mu)^2$ is "Convex function" the opt point is Min of the total energy is the expression accordingly.

Q2)

1) **D** - We know that it's a linear classify, therefore only A and D have linear separator line. Here we have C = 0.01 therefore using small value for C penalizes for small margins between classifier and data points. That is, a classifier with smaller C will have a large gap between data points and the classification line. Therefore, **D** is the best match here.
2) **A** - Like in answer (1) we have linear classifier and only A is an option. Because we have C larger than (1) we can see that **A** has smaller gap which aligns with (1), therefore, the answer is B
3) **C** - We can see that C and F are polynomial classifiers since they contain polynomial curves. F contains rather sophisticated curves compared to C, hence the answer is **C**
4) **F** - Like the answer in (3) we can see that F has a sophisticated curve that isn't 2^{nd} degree. Therefore, F is the answer
5) **E** – we have only 2 graphs left and they are RBF, gamma parameter decides how much a datapoint may affect other areas in the data. Small gamma indicated long term influence and causes decision boundary to be large and loose. On the other hand, large gamma allows only short-term influence and causes the decision boundary to be smaller and closer to the data points. Therefore, the answer here with the smaller gamma is **E**.
6) **B** – like the answer in (5) the only graph left is **B** and with the explanation I did in answer (5)

Q3)

a) The scientific term of balance that Einstein meant in machine learning is generalization.
He said that a system should be simple but not simplest this means, A system should have simple amount of features/data etc. which means we have to make the system simple (learning) but we don't want to make the system too much simple (simpler part) which will make as memorize and give us overfitting.

b) Each term in AIC $(2p, 2\ln(\hat{L})$ affect this balance.
2P – represents the number of features that we are going to learn. The more we increase P the higher AIC will get (and we want AIC to be lowest as possible) this is the "Simple part".
On other hand, the $2\ln(\hat{L})$ the lower we get P the higher value we will get "Simpler part".
We need to find the min between these 2 edges which is somewhere in the middle between them.

c) If the balance is violated which means we have two options as discussed in (b). If we have high number of features (p) the overall AIC will be really high (increases linearly) which means the system is out of balance. Another option is we have really low number of (p) which means the 2ln(L^) will be really high and the overall AIC will be high (the function increases logarithmic)

d) As discussed in previous answer, we are aim at the LOWEST value we can get with AIC. High value of AIC represents imbalanced system. We want to find the min between these 2 sides pulling to each other $(2p, 2\ln(\hat{L})$