HW3 Machine Learning in Healthcare (336546)
Omer Gottlieb - 302184064

1. Clustering

a. K-medoid is more robust to noise than K-means because it uses actual point from the cluster to represent its center, where K-means is sensitive to outliers because it calculates the center of the cluster as a mean of the data point and a mean is easily influenced by extreme values.

b. A centroid of a cluster Ci is defined as: $\mu_i = argmin_{\mu \in R^n}\{\sum_{x\in C_i} d(x,\mu_i)\}$
The K-means objective functions are: $J_{k-means}(C_1,..C_k,\mu_1..,\mu_k) = \sum_i^k \sum_{x\in C_i} d(x,\mu_i)$
In 1D case the Euclidian distance is: $\sum_{j=1}^m (x_j - \mu_i)^2$ thus in order to find a minimum for J we shall derive J according to μ and compare to 0:

$$\frac{\partial J}{\partial \mu_i} = 2\sum_{j=1}^{m}(x_j - \mu_i) = 2\sum_{j=1}^{m} x_j - m\mu_i = 0$$

$$\rightarrow \mu_i = \frac{\sum_{j=1}^m x_j}{m} = mean(x_j \epsilon C_i)$$

Bonus:
A centroid of a cluster Ci is defined as: $\mu_i = argmin_{\mu \in R^n}\{\sum_{x\in C_i} d(x,\mu_i)\}$
The K-means objective functions are: $J_{k-means}(C_1,..C_k,\mu_1..,\mu_k) = \sum_i^k \sum_{x\in C_i} d(x,\mu_i)$
In 1D case L1 distance is: $\sum_{j=1}^m |x_j - \mu_i|$ thus in order to find a minimum for J we shall derive J according to μ and compare to 0:

$$\frac{d|x|}{dx} = sign(x)$$

$$\frac{\partial J}{\partial \mu_i} = \sum_{j=1}^{m} sign(x_j - \mu_i)$$

J derivative by μ equals to zero only when the number of positive items equals the number of negatives which happens when $\mu_i = median\{x_1, \dots x_m\}$.

2. SVM

Parameter C in linear kernel tells SVM algorithm how much to penalize in order to avoid misclassifying in training set. For larger values of C SVM optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. For small values of C SVM optimizer will choose for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.
The margins have equal distance from the boundary line thus if we draw the margin in graphs A two purple points are inside the margin.
Thus A-1, D-2.

RBF kernel is defined as: $k(x,z) = e^{-\frac{||x-z||^2}{2\sigma^2}}$; $\gamma = \frac{1}{2\sigma^2}$
When γ increase σ decreases thus, the width of region of similarity is minimal and only points that are close are consider similar, and the model tends to overfit.
Thus B-6, E-5.

When we use polynomial kernel in SVM the higher the degree of the polynomial the more chance for overfitting of our model to the training set.
For polynomial with 2nd degree the boundary between regions shall have a quadratic shape and with degree of 10th the boundary probably will over fit.
Thus C-3 , F-4

3. Capability of generalization

a. The scientific term of balance Einstein meant in machine learning aspect is Generalization of a model. Einstein's quote refers to that in order to describe a phenomenon your model / theory has just enough "parts" / parameters to describe the recorded data and new data generated by the phenomenon.

b. When the number of learned parameters increase (complexity), p, the generalization of our model decreases. When the likelihood, $\hat{L}$, increases the goodness of fit increase which means the phenomenon is well represented by our model.

c. The two option that are likely to happen when model generalization is off balance are underfitting and overfitting.
When our model is too general underfitting will occur because it does not represent well the training data and thus the recorded phenomenon and when our model is not general enough overfitting will occur and it will not be able to handle new data that is generated from the phenomenon we are trying to module.

d. We aim to decrease AIC. AIC rewards goodness of fit and penalize increase in number of parameters in order to discourage overfitting, thus estimating the relative amount of information lost by a given model.
If we compare models in order to find the best to represent a phenomenon, the model with the lower AIC is a better model because it still represents the phenomenon and it has a lower risk of overfitting.