336546 - Machine Learning in Healthcare – HW3

By: Adi Waisman, 316407279

1. Clustering-

a. The K-medoid is more robust to noise (and outliers) than the K-means algorithm. K-means algorithm groups the data based on their closeness to each other according to the Euclidean distance. The mean value is taken as similarity parameter to form clusters. In a cluster with outliers, K-means will place the center of the cluster towards the outliers. K-medoid chooses actual data points as centers (medoids), whose average dissimilarity to all the objects in the cluster is minimal. Because K-medoid minimizes the sum of dissimilarities of data objects instead of the sum of squared Euclidean distances, it is more robust to noise and outliers than K-means.

b. Prove that for the 1D case of K-means, the centroid μ which minimizes the term: $\sum_{i=1}^{m}(x_i-\mu)^2$ is the mean of m samples, meaning: $\mu=\frac{\sum_{i=1}^{m}x_i}{m}$.

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

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

Therefore, the centroid which minimizes $\sum_{i=1}^{m}(x_i-\mu)^2$ is the mean of m samples.

Bonus: Prove that the centroid (the medoid) which minimizes the term $\sum_{i=1}^{m}|x_i-\mu|$ is the median of m examples given that μ belongs to the dataset.
I will show that μ minimizes the $E|X-\mu|$ for $X=\{x_i\}_{i=1}^{m}$:

$$\frac{\partial(E|X-\mu|)}{\partial\mu}=\frac{\partial}{\partial\mu}\left(\int_{-\infty}^{\infty}P(x)\cdot|X-\mu|dx\right)=\frac{\partial}{\partial\mu}\left(\int_{-\infty}^{\mu}P(x)|X-\mu|dx+\int_{\mu}^{\infty}P(x)|X-\mu|dx\right)=$$
$$=\int_{-\infty}^{\mu}\frac{\partial}{\partial\mu}(P(x)|X-\mu|dx)+\int_{\mu}^{\infty}\frac{\partial}{\partial\mu}(P(x)|X-\mu|dx)=$$
$$=\int_{-\infty}^{\mu}-P(x)dx+\int_{\mu}^{\infty}P(x)dx=0$$

$$\implies\int_{-\infty}^{\mu}P(x)dx=\int_{\mu}^{\infty}P(x)dx\implies P(X\le\mu)=P(X\ge\mu)$$

$$\underset{(*)}{\Rightarrow}P(X\le\mu)=1-P(X\le\mu)\implies 2P(X\le\mu)=1\,,\quad(*)\,P(X\le\mu)+P(X\ge\mu)=1$$

$$\implies\boldsymbol{P(X\le\mu)=\frac{1}{2}=P(X\ge\mu)}$$

Therefore, μ is the median that minimizes the term.

2. SVM-

The settings that were used are:

1. Linear kernel with $C = 0.01$ - matches image **D**.
2. Linear kernel with $C = 1$ - matches image **A**.
   Explanation: Images A and D match setting 2 and 1 because there is linear separation between the classes. When the cost hyperparameter C, is set to a low value ($C = 0.01$) the SVM classifier will choose a large margin decision boundary at the expense of a few misclassifications, which matches image D. This results in a simpler and better generalized model. When C is set to a higher value (in our case $C = 1$), the classifier will choose a low margin decision boundary and try to minimize the misclassifications, matching image A. This may result in an overfitted model.
3. 2nd order polynomial kernel - matches image **C**.
4. 10th order polynomial kernel - matches image **F**.
   Explanation: Images C and F match setting 3 and 4 because SVM with a polynomial kernel generate a non-linear decision boundary with polynomial features. The higher the degree of polynomial, the more "bendy" and complex the decision boundary will be. Therefore, image C matches 2nd order polynomial kernel and image F matches 10th order polynomial kernel. The higher the degree of polynomial, the more it tends to overfit the training set.
5. RBF kernel with $\gamma = 0.2$ - matches image **E**.
6. RBF kernel with $\gamma = 1$ - matches image **B**.
   Explanation: Images B and E match setting 6 and 5 because SVM with RBF kernel produces a ring-shaped decision boundary. Low values of gamma ($\gamma = 0.2$) indicate a large similarity radius, matching image E. For higher values of gamma, the similarity radius decreases, and the data points need to be very close to each other in order to be considered in the same class. Models with larger gamma values tend to overfit, matching image B.

3. Capability of generalization-

a. One of the most important consideration when training a model is how well it will generalize to new, previously unseen, observations. This is called generalization. Regularization is a technique to improve the generalizability of the model, it is a fundamental concept in machine learning. Its purpose is to penalize the complexity of the model while favoring a "more likely" explanation, because a simpler hypothesis representation is less prone to overfitting. To summarize, the term of balance is between the complexity of the model and goodness of fit.

b. In AIC, p is the total number of learned parameters in the model, it is a measure of the model complexity. The higher the number of parameters, the more complex is the model. The estimated likelihood of the model $\hat{L}$, measures the goodness of fit of the model. The higher $\hat{L}$ is, the better the goodness of fit. The AIC score rewards models that achieve a high goodness of fit score and penalizes them if they become overly complex.

c. If the balance between goodness of fit and complexity is violated the model can be overfitting or underfitting the data. When the model is too simple, and there isn't

enough data for it to understand the pattern, the model will be underfitting. When there are 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. We aim to get the model with the lowest AIC score, because it will have a better balance between its ability to fit the data set and its ability to avoid overfitting. We can look at the AIC formula this way:

$$AIC = 2p - 2\ln(\hat{L}) = 2\ln(e^p) - 2\ln(\hat{L}) = 2\ln\left(\frac{e^p}{\hat{L}}\right)$$

When p increases (i.e. the model is more complex), e^p increases and the AIC score is higher. When $\hat{L}$ increases (i.e. the goodness of fit of the model is better), the AIC score is lower. Therefore, we prefer lower AIC score model.