HW3 -Machine Learning in Healthcare 336546

Marah Matar 208235994

1 Clustering

a. K-medoid is more robust to noise and outliers than the k-means algorithm. Because it uses points from the dataset in the cluster with minimum sum of distances to other points to represent the center of a cluster, instead of using the mean point as is the case in K-means. For the mean adding an outlier to the dataset shifts the mean closer to it and thus affects the result. On the other hand, since a median is the middle data point of the arranged data set, assigning extreme values to points on the edges does not always change the answer, so the K-means clustering algorithm is sensitive to outliers because a mean is easily influenced by extreme values.

b. To find the centroid (μ) which minimizes the term:

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

we need to differentiate and compare to zero:

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

$$\mu m - \sum_{i=1}^{m} x_{i=0}$$

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

second derivation:

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

so, it is a minimum .

c. We need to differentiate and compare to zero to find the centroid (practically, the medioid) which minimizes the term

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

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

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

For this

We need exactly half of the xi 's to be bigger than μ and exactly half of them to be smaller to that term be equal zero .

So μ must be median of series x (m examples).

2 SVM

The linear kernels the images A,D the data is separated using a linear boundary, the difference is in the capacity hyperparameter C that tells the SVM optimization how much we want to avoid misclassifying each training example. Larger C means higher penalty to miss-classifications or data points between margins, and very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. Given that the support vectors of the two sides of the boundary line should be at equal distance from the boundary line, the two purple points at A at the boundary line are probably not support vectors, but data points inside the margin, since there are no blue data points on the other side with same distance. Therefore, in A there were more data points inside the margins. Meaning that the penalty of miss-classifications or data inside margin was smaller, and thus smaller C. In D there is a safe distance between the samples of both classes and the boundary which means a higher value of C=1.
In conclusion: A-1, D-2.

The RBF kernels- Gaussian kernel in images B,E that has closed Gaussian boundary the classification results are like a topographic map ,The differences are in gamma Gamma is inverse to the standard deviation of the gaussians. that the higher Gamma is the more we better fit the training data but too high of a Gamma can cause overfitting. A small value of gamma the region of influence of any selected support vector would include the whole training set. B boundary line is more fitted and more specific to the data, with smaller surface so B is the RBF with the higher gamma(1).
In conclusion: B-6, E-5.

The polynomial kernel in images C,F because the data is separated with high order polynomial. And the higher degree polynomial order kernels allow a more flexible decision boundary. So, we can see that the more flexible boundary is in image F so we assume it is with $10th$ order polynomial kernel, and C has Quadratic function as a boundary (2 nd).
In conclusion: C-3, F-4.

3 Capability of generalization

a. In machine learning aspect, the scientific term of the balance that Einstein discussed is the balance between the model complexity and the goodness of fit or generalization as we learn as we increase the model complexity the variance will increase, and the bias will be reduced , this balance called the bias-variance tradeoff which measured by Overfitting (low bias and high variance)and underfitting(high bias and low variance).

b. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and model complexity. By looking at AIC value : AIC=2p-2ln(L),
we can say that: The term **'2p'** (the number of parameters used to build the model) describes the **model complexity** (in unsupervised: number of clusters),The more complex the model is more variables needed to get the estimate or prediction.
and the **'2ln(L)'** (estimated the maximum likelihood given these parameters of the model) describes **the goodness of fit** (how well the model reproduces the data). The function 2ln(L) contributes to the goodness of function, when it gets higher the AIC gets lower value and it means that the model is fitted better.
As the model complexity increases, '2p' increases. And as the model fits the data better, L increases and therefore '-2ln(L)' decreases. So, there is a fine point where the two terms are in balance.

c. If this balance was violated, we might get overfitting or underfitting:

The risk of underfitting- having a model that is so simple with low number of clusters and not trained enough, and by that doesn't fit the data well .
In this case : low 2p and L therefore high '-2ln(L)', In total, AIC is high the model doesn't fit the data well and is not possible to get accurate new predictions .

The risk of overfitting- the complexity of a model could cause overfitting and with the goodness of fit comes a more complex model (high number of clusters) and fits the data too well (low variance in clusters) .
In this case : high 2p and L therefore low '-2ln(L)', In total, AIC is high the model isn't generalized to new data, and it is not useful for new predictions.

d. We are aiming for a low AIC, because as was mentioned above we want to maximize the likelihood so we will achieve a good fit model, also we do not want it to be a very complex, so we are in favor for an approximately low number of parameters.
This is an optimization problem. Since '2p' increases with model complexity and '-2ln(L)' increases with goodness of fit, the good balance between the two can be identified when AIC reaches a minimum.