Machine Learning in Healthcare- 336546

HW3

Leen Ileimi, 208809715

1 Clustering

a. K-medoid is more robust to noise and outliers than the k-means algorithm. The reason is that in k-means, the centers of the clusters are points that seek to minimize the Euclidian metric between these centers and the data points from same cluster ("centroids"), whereas in k-medoids, the centers of the cluster are **points from the dataset** that seek to minimize the L1 distance between these centers and data points from same cluster ("medoids"). In other words, this is similar to comparing "mean" and "median". Since a mean or a "centroid" seeks to minimize Euclidian metric, adding an outlier to the data-set shifts the mean (centroid) closer to it and thus affects the result. On the other hand, since a median or a "medoid" is the middle data point of the arranged data set, assigning extreme values to points on the edges does not always change the answer. Therefore, outliers affect the k-means centers (centroids), while they do not affect the k-medoid centers (medoids) as much.

b. Proving that for the 1D case ($x \in R^1$) of K-means, the centroid (μ) which minimizes the term $\sum_{i=1}^{m}(x_i - \mu)^2$is the mean of m examples.
Proof:

$\underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{m}(x_i - \mu)^2$ = the μ that brings the derivative to zero:

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

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

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

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

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

Bonus:

Proving that centroid which minimizes the term $\sum_{i=1}^{m}|(x_i - \mu)|$ is the median of m examples, given that μ belongs to the dataset.
Proof:

$\underset{\mu\in\{x_i\}}{\operatorname{argmin}} \sum_{i=1}^{m}|(x_i - \mu)|$ = the μ that brings the derivative to zero:

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

$$\frac{\partial \sum_{i=1}^{m}|(x_i - \mu)|}{\partial \mu} = \sum_{i=1}^{m}(-1) * sign(x_i - \mu) = 0$$

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

Now we divide the sum to x's greater than μ ,x's smaller than μ, and equal to μ:

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

$$= \sum_{x_i<\mu} sign(x_i - \mu) + \sum_{x_i>\mu} sign(x_i - \mu) + \sum_{x_i=\mu} sign(x_i - \mu)$$

$$= \sum_{x_i<\mu}(-1) + \sum_{x_i>\mu}(+1) + \sum_{x_i=\mu} 0 = |x_i > \mu| - |x_i < \mu| = 0$$

Where $|x_i > \mu|$= number of x's greater than μ and $|x_i < \mu|$= number of x's smaller than μ.

$$\rightarrow |x_i > \mu| = |x_i < \mu|$$

Meaning: μ that minimizes the term $\sum_{i=1}^{m}|(x_i - \mu)|$ is μ where the number of x's greater than μ is equal to the number of x's smaller than μ. Or in other words, μ is the middle of an ascending series.

Therefore: μ is the median of the series {x}.

2 SVM

We will start with the linear kernels. As can be seen, the two linear classifiers (with a linear boundary line) are A and D. The difference is in the capacity hyperparameter C. Larger C means higher penalty to miss-classifications or data points between margins. 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 conclusion: A-1, D-2.

We will now move to the RBF kernel. RBF kernel is Gaussian kernel. And the classification results are similar to a topographic map, therefore B and E. The differences are in gamma. Gamma is inverse to the standard deviation of the gaussians. Higher gamma means better fit to the training set (very high gamma= overfitting). B boundary line is more fitted and more specific to the data, with smaller surface. Therefore, B is the RBF with the higher gamma.
In conclusion: B-6, E-5.

Now, for the polynomial kernel, it is obvious that C is of lower degree (2^{nd} degree), and F is of higher order (10^{th} order)- it is more fitted to data and of a more complex model. Meaning C-3, F-4.

To summarize:

A-1
B-6
C-3
D-2
E-5
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.

A model that is too simple leads to under-fitting. It doesn't describe the data well enough, and it is not useful to predict or classify new data based on a training set (or in unsupervised learning- it does not cluster the data correctly). On the other hand, a model that is too complicated leads to over-fitting. It fits the data too well, and leads to memorizing instead of learning. Thus, it is also not useful to predict or classify new data based on the training set (or in unsupervised learning, the data is "over clustered" and thus clusters might become meaningless).

Therefore, there should be a balance in the complexity of the model (or number of clusters) and in the extent to which the model fits the data (or variance inside clusters). This is where model complexity and generalization come in.

b. The term '2p' (the number of parameters) describes the model complexity (in unsupervised: number of clusters), and the '2ln(L)' (estimated likelihood given these parameters) describes the goodness of fit. So, in total, the AIC criterion describes the balance between model complexity and goodness of fit. 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. The balance can be violated in under-fitting or over-fitting:

Underfitting: the model is too simple with low '2p' (e.g., low number of clusters). It doesn't fit the data well (e.g., high variance in clusters), and therefore has low 'L' and therefore a very high '-2ln(L)'. In total, high AIC. In this case, the model doesn't fit the data well, and is not useful for new predictions.

Overfitting: the model is too complex (e.g., high number of clusters) and fits the data too well (e.g., low variance in clusters). Thus, it has high '2p', and high 'L'

therefore low ‘-2ln(L)’. In total, AIC is again high. In this case, the model isn’t generalized to new data. Again, it is not useful for new predictions.

d. We are aiming to a low AIC, where there is a balance between the model complexity (e.g., number of clusters) and the fitting of the data (variance in each cluster). As was mentioned above, 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.