Amit Parizat – 207652009

Machine Learning in Healthcare – 336546 – Homework 3

Clustering

a. K-medoid is more robust to noise (outliers) than the K-means algorithm. This is because generally speaking; the median of a population is less affected by outliers than the mean of the population. Since the center of each cluster is determined by those values in every iteration of those algorithms, the algorithm using the medians will be more robust to outliers than the one using the means.

b. To find the minimum of the term $f(X,\mu)=\sum_{i=0}^{m}(x_i-\mu)^2$ for constant values of X I will differentiate it with respect to μ:

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

To find extremums I will compare it to zero:

$$-2\sum_{i=0}^{m}x_i+2m\mu=0$$

$$\mu=\frac{1}{m}\sum_{i=0}^{m}x_i=\bar{X}$$

Since $\frac{\partial^2 f}{\partial \mu^2}=2m>0$ this is the minimum of f.

c. First, notice that the derivative of the function $|x_i-\mu|$ with respect to μ is $sign(x_i-\mu)$ for all $\mu\neq x_i$. Thus:

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

This expression is zero only when the number of x_i larger than μ and the number of x_i smaller than μ are equal. This is exactly the definition of the median of a population. Thus when μ is the median the term $\sum_{i=0}^{m}|x_i-\mu|$ is minimized.

SVM

In the models of graphs A and D, linear kernel was used since the separation curve between the classes is linear. For linear kernel SVM, when the parameter C is large, the margin is smaller to classify more training examples correctly. In graph A there are (almost) misclassified examples, and the margin is larger than in graph D, and thus model A has a smaller C parameter. Therefore, model A is a linear kernel with $C = 0.01$ and model D is linear kernel with $C = 1$.

In the models of graphs B and E, RBF kernel was used since the separation curve between the classes is sort of circular. In graph B the separation curve is “tighter” around the blue class than it is in graph E. For larger γ values the influence of the examples is shorter and thus the separation curve is tighter. Therefore, model B is an RBF kernel with $\gamma = 1$ and model E is an RBF kernel with $\gamma = 0.2$.

In the models of graphs C and F, polynomial kernel was used since the separation curve between the classes looks like a linear combination of the features and their higher orders. The higher the order of the polynomial kernel, the higher the overfitting of the model. Graph F is more overfitted towards the training set and thus model C is a 2nd order polynomial kernel and model F is a 10th order polynomial kernel.

To summarize:

$$A \rightarrow 1\,, B \rightarrow 6\,, C \rightarrow 3\,, D \rightarrow 2\,, E \rightarrow 5\,, F \rightarrow 4$$

Capability of Generalization

a. The scientific term of the balance that Einstein meant to in machine learning aspect is generalization. When a ML model is oversimplified and underfitted it leads to high bias errors. When a ML model is overcomplicated and overfitted it leads to high variance errors. In both cases generalization of the model by testing it on test group will result in large errors. A good ML model should balance between those extremes.

b. The term $2p$ refers to the number of learned parameters. The higher this term, the more complex the model and thus contributes to overfitting of the model. On the other hand, the term $2\ln(\hat{L})$ refers to estimated likelihood given those parameters. For higher values of this term, the model is fitted better to the training set and thus it contributed to the goodness of fit.

c. When this balance is violated by overfitting using too complex model with too many parameters, the AIC is large. When using too little parameters for an underfitted model, the likelihood will be small (since the model does not fit well to the data) and AIC will be again large.

d. As I explained in the previous section, when the model is either overfitted or underfitted, the AIC is large. Thus, one should try to obtain small AIC value for his model which indicates that not too many parameters were used to obtain a decent fit of the model to the data.