336546 - Machine Learning in Healthcare

HW3

Daniel Sapir

308412113

1 Clustering

a. The K-Medoids algorithm is more robust as compared to K-Means. Because *k*-medoids minimizes a sum of general pairwise dissimilarities, while K-Means minimizes a sum of squared Euclidean distances. This distance reduces noise and outliers.

b. We started with the equation:

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

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

$$= \sum_{i=1}^{m}(x_i - \bar{x})^2 + 2\,(\bar{x} - \mu)\sum_{i=1}^{m}(x_i - \bar{x}) + n(\bar{x} - \mu)^2 =$$

$$\sum_{i=1}^{m}(x_i - \bar{x})^2 = 0$$

$$= \sum_{i=1}^{m}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

The left part is constant in relation to μ.

The value of μ that minimizes the whole equation, is when $\mu = \bar{x}$

Bonus:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}|x_i - \mu| = \sum_{i=1}^{A} x_i - \mu + \sum_{i=1}^{B}(\mu - x_i) + (\mu - \mu)$$

While A: $x_i > \mu, \quad B: x_i < \mu$

$$= \sum_{i=1}^{A} x_i - \mu A + \mu B - \sum_{i=1}^{B} x_i$$

$$\Rightarrow \frac{d}{d\mu}\left(\sum_{i=1}^{A} x_i - \mu A + \mu B - \sum_{i=1}^{B} x_i\right) = -A + B = 0$$

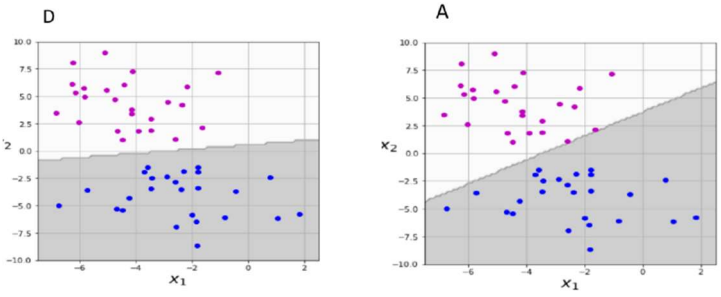
$$A = B$$

And we know that while μ that minimizes the term m i=1 is the median if A=B when the term is minimized.

2 SVM

$A \rightarrow 1$

$D \rightarrow 2$:



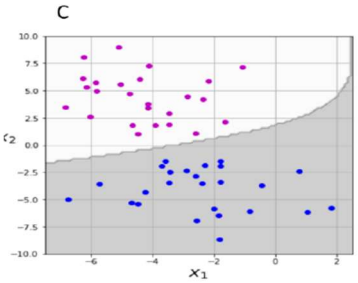
We can see that both A and D are linear model, So the options are 1 and 2.

When we increase the C value, the margin gets smaller. On the contrary decrease C value, the margin gets bigger. In graph A it is one can observe 2 purple points that are almost on the decision boundary, while in the blue class there are no points this close to the decision boundary. This indicates that some misclassification is in this model, which means C is relatively small.

Graph D shows that does not in misclassifications, so it include bigger C than the previous section.

$C \rightarrow 3$:

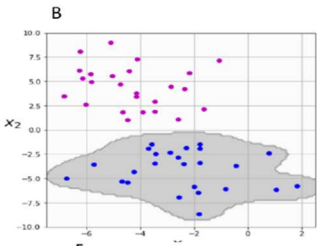
C is simple second order vector. So easily we can determine that is 3 as we can see the parabolic graph in C.

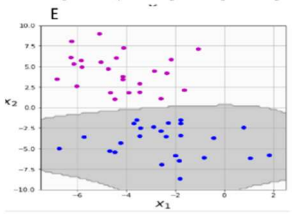


$B \rightarrow 6$

$E \rightarrow 5$:

E and B looks like RBF because they represent non-linear SVM. And One of the most used non-linear kernels is the radial basis function (RBF).



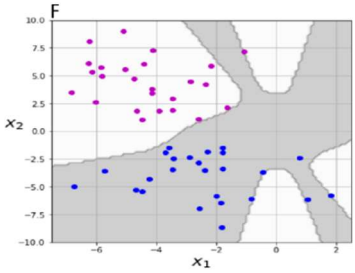


Low values of gamma indicate a large radius which results in more points being grouped together. And high values of gamma, the points need to be very close to each other to be considered in the same group. Therefore, models with very large gamma values tend to overfit.

We can see her that B represent some overfit, so it fits the value of 1 (bigger one). And E represent less overfitting rather than B, so it fit the value of 0.01.

$F \rightarrow 4$:

Simple clearly overfitting with ten order polynomials.



3 Capability of generalization

a. The term is Generalization. There are two options in ML models, the first is when oversampling and underfitting (high bias error), and the second is when overcomplicated and overfitting (high variance error). While Generalization will be in a large error.

b. On the one hand $2p$ present the model complexity (number of learning parameters), and on the other hand $-2\ln(\hat{L})$ represent the model performance (estimated Likelihood). So, the LOG-Likelihood of the model determine the variables which could lead to overfitting. Hence, $2p$ penalty term introduced which does not remove overfitting completely.

c. The balance could violate by underfitting (small parameters), the Likelihood will be small, and the AIC might be big. Additionally, if the balance is violated by too many parameters, and overfitting using too much complex model, the AIC is big (as we can see in the equation).

d. The AIC statistic penalizes complex models less, meaning that it may put more emphasis on model performance on the training dataset, and select more complex models. And as explained in the last section the AIC is large when overfitting or underfitting. So small AIC Clarify that not too many parameters are used. Additionally, we can see in the equation that low $2p$ will lead to simple model (no underfitting), and high $2\ln(\hat{L})$ will lead to high probability estimation.

4 EigenFaces

The code is attached in the Jupyter Notebook.