

HW3 - Machine Learning in Healthcare**1. Clustering**

- a. K-Medoids is more robust as compared to K-Means as in K-Medoids we find k as representative object to minimize the sum of dissimilarities of data objects whereas, K-Means used sum of squared Euclidean distances for data objects. And this distance metric reduces noise and outliers. [1]

- b. which μ minimizes the term $\sum_{i=1}^m (x_i - \mu)^2$?

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

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

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

for the 1D case the centroid (μ) which minimizes the term is the mean of m examples.

- c. which μ minimizes the term $\sum_{i=1}^m |(x_i - \mu)|$?

$$\frac{d}{dx} |x| = \text{sign}(x)$$

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

This term equals to zero only when the number of positive items equals the number of negative items. Therefore:

$$\mu = \text{median}\{x_1, x_2, \dots, x_m\}$$

1. Arora, P. and Varshney, S., 2016. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78, pp.507-512.

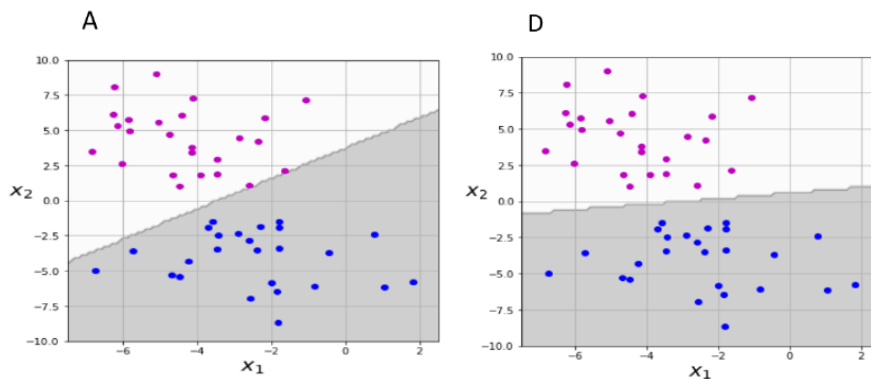
2. SVM

In both figures A and D it is a linear separation. C hyperparameter balances two different goals: maximizing the margin and minimizing the number of errors on the training data. When C is very small, the sum of error terms becomes negligible in objective function, the margin gets maximized and often more mistakes are made in classifying the training data. When C is very large, the sum of error terms dominates the margin term in objective function, the margin can be so small that it does not contain any points. We usually get fewer mistakes in classifying the data for training but the model tends to overfitting. [2]

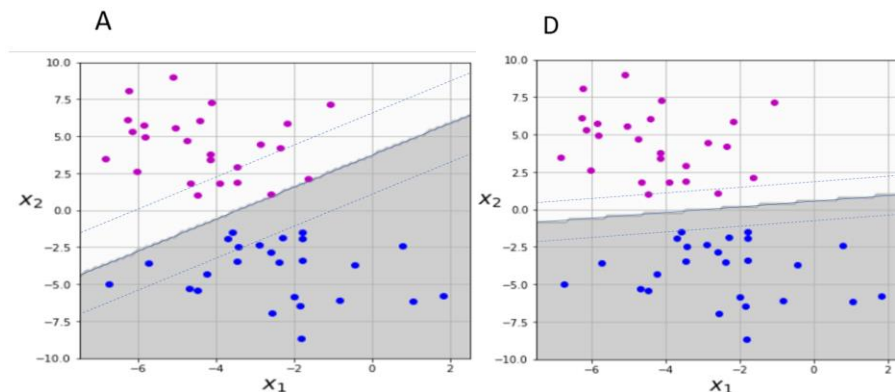
In both figures all of predictions were correct. In figure A, two points fall on, or very close to, the hyperplane, and it's seems that the margins contain more datapoints than in D, therefore it can be assumed that in figure A the value of C is smaller.

1-A $C=0.01$

2-D $C=1$



Optionaly margins:

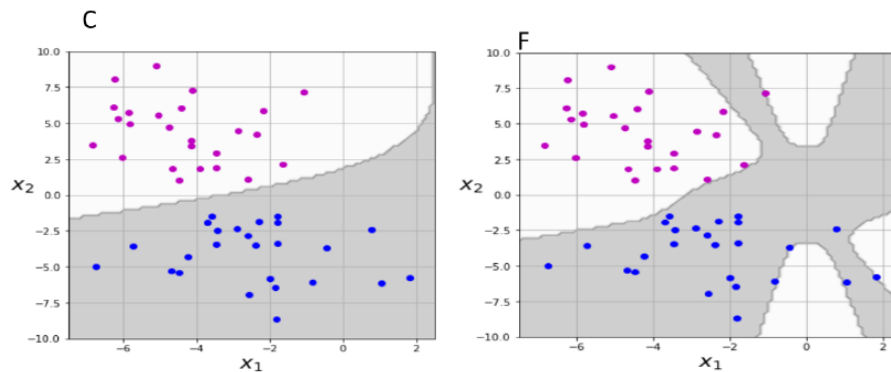


2. Nefedov, A., 2016. Support Vector Machines: A Simple Tutorial. svmtutorial. online.
URL: <https://svmtutorial.online>.

In figures C and F we can see a shape obtained from a polynomial kernel. In figure C, a low-order polynomial and in figure F, a high-order polynomial, Complex shape, and overfitting of the data.

3-C 2nd order polynomial kernel.

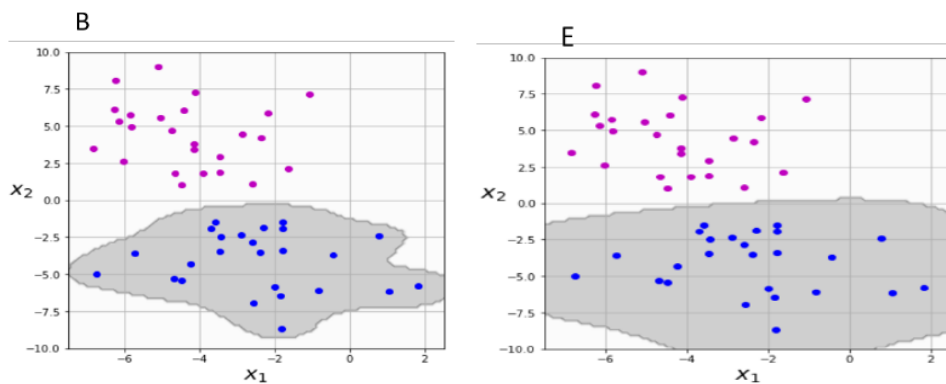
4-F 10th order polynomial kernel.



Figures B and E show close shape boundaries obtained as a result of Gaussian. The larger the γ , the better the approximation of the data for training and the model tends to overfitting.

5-E RBF kernel with $\gamma = 0.2$.

6-B RBF kernel with $\gamma = 1$.



3. Capability of generalization

- a. In machine learning aspect, there is balance between bias and variance. when increasing the bias we decrease the variance and when increasing the variance we decrease the bias. The bias reflects the hypothesis function complexity and the variance reflects how the model generalizes to new observations. We tradeoff between the two by selecting an appropriate classifier and adjusting the hyperparameters.

- b. AIC is an estimator used for model selection, the formula for AIC is:

$$AIC = 2p - 2\ln(\hat{L})$$

p is the number of independent variables, and therefore indicate model complexity, and \hat{L} is the log-likelihood estimate and reflects how the model generalizes to new observations.

- c. If the variance-bias balance is violated, the model tends to overfit or underfit the data. If the model is oversimplify we tend to underfitting and if the model is too complex we tend to overfitting.
- d. We aim to get AIC value as low as possible. A simple model will contain fewer parameters therefore the value of p will be low. And as the model will be more accurate the value of $-2\ln(\hat{L})$ will decrease.