HW 3 – ML in healthcare

Submitted by: Lidan Fridman 206201816

Part 1 – Clustering

a. K-medoid is more robust to noise and outliers than K-means because the classification of K-medoids relies on representative medoid elements of the data.
The medoids are the most centrally located elements within each cluster and by minimizing the dissimilarities of pairwise data examples (medoid and other object in the cluster). In this method we are less sensitive to errors caused by data distribution (like in k-means where we minimize squared distance between the mean of the cluster and another data point).
Generally speaking, average is more affected by data distribution than medians.

b. Let our lost function be $L = \sum_{i=1}^{m}(x_i - \mu)^2$, convex function and there is one global minimum.

Minimization:

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

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

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

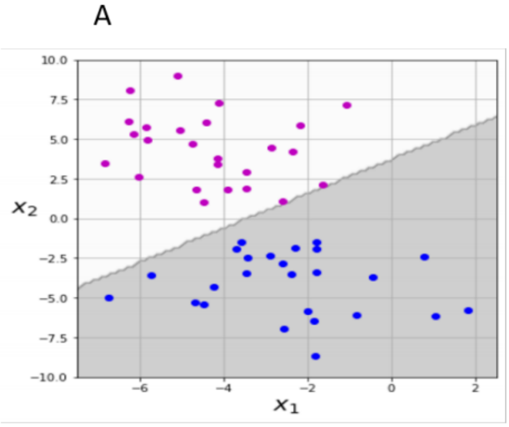
c. Bonus: Let our lost function be $L = \sum_{i=1}^{m}|x_i - \mu|$, also convex with single global minimum.

Minimization:

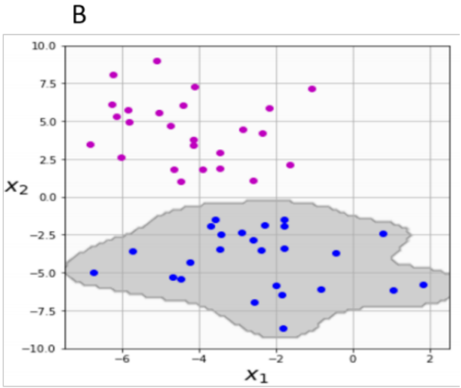
$$\frac{\partial L}{\partial \mu} = -\sum_{i=1}^{m} sign(x_i - \mu) = 0$$

$sign(x_i - \mu)$ is one when $x_i > \mu$ and is minus one when $x_i < \mu$. The derivative is zero when there is the same number of data points which are greater and lower than μ which means that μ is the median of x_i.

Part 2 – SVM



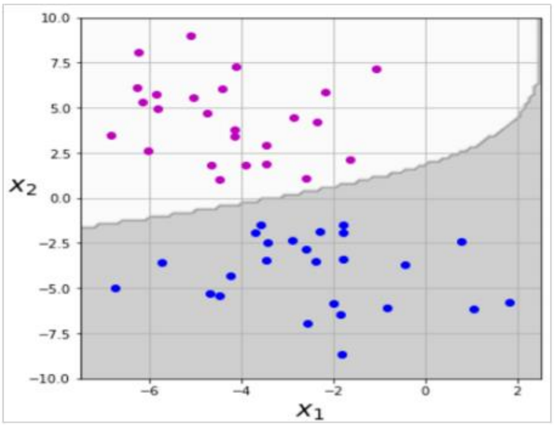
A-1: Linear $C = 0.01$, we can see a linear separation line. We can see that there are two purple samples inside the margins, close to the separation line, and there is a tendency for misclassifications. The purple dots are inside the margins and the margins not narrow because margins are symmetrical and there isn't any blue dot close enough to the separation line to be considered as the other margin. Lower value of C indicates more tolerance to misclassification and data points inside the margins as could be seen here.



B-6: RBF $\gamma = 1$. We can see complex wrapping margins characteristic of RBF. High γ leads to tighter margins and could also lead to overfitting.

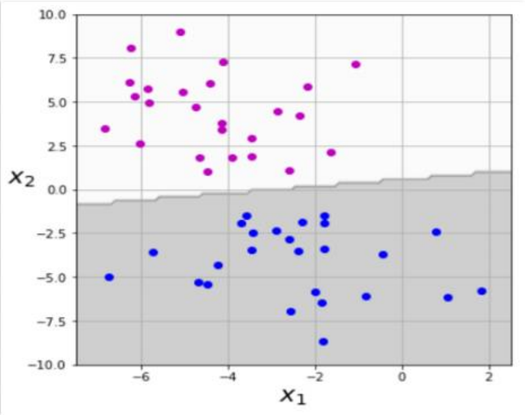
C-3: 2^{nd} Polynomial. We can see mildly curved polynomial decision boundary. The low polynomial degree gives us almost similar result to linear SVM and reduces the risk of overfitting.

C



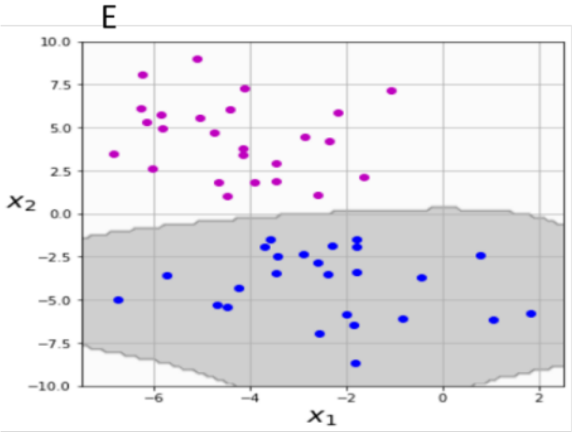
D-2: Linear $c = 1$. We can see a linear separation line. There are blue and purple dots located at equal distances from the separation line and we could assume they are on the margins. In this case there would be less or none cases of misclassifications or datapoints inside the margins, and therefore we have higher value of C (more penalization to misclassified/leaking samples)

D

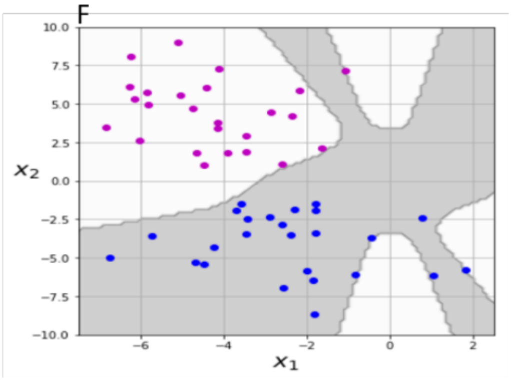


E-5: RBF $\gamma = 0.2$. We can see complex wrapping margins characteristic of RBF. Low γ leads to broader margins and reduces the risk of overfitting.

E



F-4: 10^{th} Polynomial. We can see severe overfitting with distinct complex margins indicating overfitting. This could be the result of a projection of a high order polynomial hyperplane. Therefore the degree of the polynomic kernel is higher than in picture C.



Part 3 – Capability of generalization

a. Generalization balances between adequate data fitting to previously unseen data and model complexity. Bad generalization could stem from too complex or too simple a model and it is assessed by performances tests.

b. **2p** – higher p meaning more parameters/features to learn leading to more complex model which could lead to overfitting and bad performances.
Lower p meaning fewer features, simpler model, could lead to underfitting and bad performances.
In both cases the generalization is bad and the model would fit poorly to the test set.
according to the AIC criteria, we would like fewer features in order to achieve a better model, but not too much because then the likelihood would decrease.

$2\ln(\hat{L})$ – if the log-likelihood is higher, the total logarithm of the value will also be higher (monotonic function). Higher log-likelihood meaning better model fitting which would yield better generalization and probably moderate complexity.
Lower log-likelihood values could indicate under or overfitting.

c. Overfitting with too complex a model and underfitting with too simple a model. Both cases lead to bad performances (likelihood) and bad generalization.

d. AIC should be low, we want high likelihood > high log(likelihood) > low minus of the log-likelihood as well as fewer features (p) as possible, without damaging the likelihood.