Machine Learning In Healthcare- HW3

Ofri Vizenblit- 205894348

1. Clustering

a. K-means and K-medoids are both partitional algorithms. K-means attempts to minimize the total squared error, while K-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In the K-Medoids algorithm, instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.
Thus, the K-mean algorithm is more sensitive to outliers, since an object with an extremely large value may substantially distort the distribution of the data.
In addition, K-medoids could be more robust to noise and outliers as compared to K-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances.

b. We saw in the lecture that the cost function of the K-mean algorithm is:

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

When μ is the centroid, C is a cluster in size m. we want to prove that the μ that brings this expression to a minimum is the mean of m examples (mean of the cluster). Let us first optimize with respect to μ:

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

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

We will check that this extreme point is indeed a minimum:

$$\frac{\partial^2 J(C,\mu)}{\partial \mu^2} = -2 \cdot (-1) = 0 \rightarrow min$$

That is, the μ that minimize the cost function is the mean of m examples.

Bonus:
The cost function in this case is:

$$J(C,\mu) = \sum_{i=1}^{m} |x_i - \mu|$$

μ is a point on the dataset.
Derivative of the function by μ and set it to zero to find an extremum point:

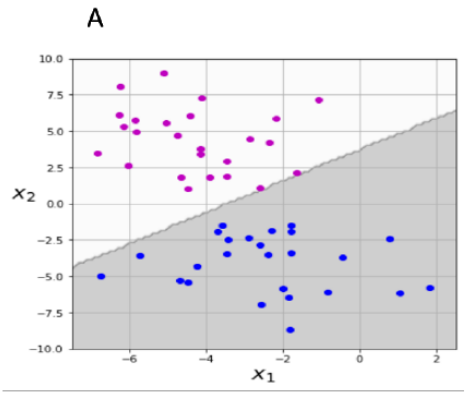
$$\frac{\partial J(C,\mu)}{\partial \mu} = -\sum_{i=1}^{m} sign(x_i - \mu) = 0$$

Assuming that there are n points greater than μ, this sum contains n ones and m-n minus ones.

$$n + (m-n) \cdot (-1) = 0 \rightarrow n = \frac{m}{2}$$

Hence that to get a minimum, half of the points must be greater than μ and the other half must be smaller than μ. Given that μ is a part of the dataset, μ is the median.

2. SVM



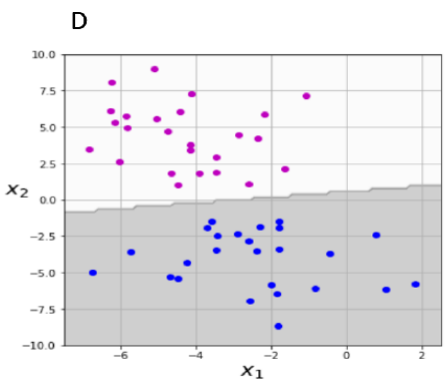
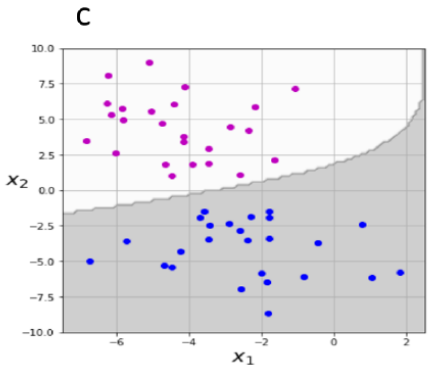


Figure A represents linear kernel with C=0.01 (1), and figure D represents linear kernel with C=1 (2).

First, we notice that in figures A and D the data is linearly separable and the SVM algorithm finds an optimal linear hyperplane that separates the classes, so these figures represent linear kernel. In addition, the hyperparameter C adds a penalty for each misclassified data point, so if C is small, the penalty for misclassified points is low and a decision boundary with a large margin is chosen. The margin is defined by the same distance of the boundary line from the two supporting vectors that are formed. In figure A we can identify that the distance of the points closest to the boundary line on both sides is not the same, in contrast to Figure D. means that in the case shown in A there are points in the margin so the penalty term must be small, that is smaller C.



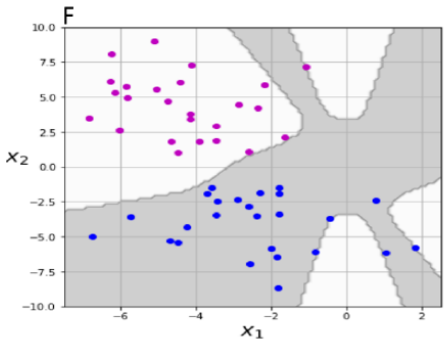


Figure C represent 2nd polynomial kernel (3) and figure F represent 10th polynomial kernel (4).

The SVM algorithm can deal with non-linearly separable data by adding new dimensions. These additional dimensions allow the data to be linearly separated. When the hyperplane is projected back onto the original two dimensions, it appears as a curved decision boundary. Because the hyperplanes in figures B, C, E, F are not linear, these figures represent non-linear kernels.

The polynomial kernel generates new features by applying the polynomial combination of all the existing features. That is, figures C and F represent polynomial kernels. The degree hyperparameter controls how “bendy” the decision boundary will be for the polynomial kernel, so figure F represent a larger order polynomial.

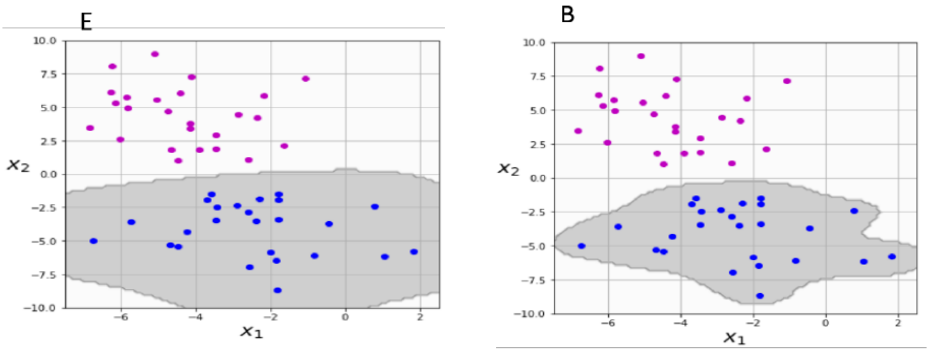


Figure B represent RBF kernel with γ=1 (6) and figure E represent RBF kernel with γ=0.2 (5).

Radial Basis Function generates new features by measuring the distance between all other dots to a specific dot/dots- centers, (RBF kernel is the Gaussian Radial Basis Function).

The hyperparameter γ controls the influence of new features on the decision boundary. The higher the γ, the more influence of the features will have on the decision boundary, more wiggling the boundary will be. We can see more wiggling boundary on figure B, means a higher γ.

3. Capability of generalization

a. The overfitting – underfitting balance (generalization vs simple models). In the context of overfitting, excessive complex models (These models have low bias and high variance) may tend to overfit the training data and therefore lead to reduced accuracy in the test set. Whereas simpler models may capture the underlying structure better and thus may have better predictive performance.
But, if the model is too simple, there is a risk for underfitting. underfitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance.

b. The AIC method is an information criterion for selecting the right complexity model.
the AIC term: $AIC = 2p - 2\ln(\hat{L})$
$\hat{L}$- estimated likelihood given these parameters (maximized likelihood). p- number of parameters.
For a larger p, the model has more parameters and therefore it is more complex and may tend to overfitting. Means that the larger the p, the larger the AIC term.
the maximized log likelihood of the model represents a measure of the quality of the fit to the data. That is, the more accurate the model, the larger the $\hat{L}$ and the AIC expression is smaller.

c. In a situation where the balance is disturbed, we can have an overfitted model of an underfitted one.
An underfitted models usually have high bias and low variance and can happen when there are not enough samples or if there are too many features, in that cases we cannot build an accurate model.
An overfitted models have low bias and high variance. These models excessively complex, therefore they may lead to poor predictions on unseen data and may be computationally expensive. Overfitting can happen when we train our model a lot over noisy dataset.

d. We want to find a model that is as simple as possible (small p) and as accurate as possible (large $\hat{L}$). Therefore, according to the equation given to the AIC, we would like to choose the model with the lowest AIC score.