Machine Learning in Healthcare – 336546

HW 3

Dmitry Rudman – 333814283

1. Clustering

a. The K-means algorithm is sensitive to outliers – since an object with an extremely large value may substantially distort the distribution of the data. Since a medoid defined as the most centrally located object in a cluster, K-medoids algorithm is more robust to noise than K-means algorithm, it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances.

b. Given term $\sum_{i=1}^{m}(x_i-\mu)^2$ represents a sum of squared errors (SSE). We can show that the centroid μ, which minimizes the term is the mean of m examples by differentiating the term:

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

setting it equal to 0:

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

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

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

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

Thus, we've proved that the centroid μ, which minimizes the term is the mean of m examples.

Bonus. Given term $\sum_{i=1}^{m}|(x_i-\mu)|$ represents a sum of absolute errors (SAE). We can show that the centroid, which minimizes the term is the median of m examples given that μ belongs to the dataset by differentiating the term:

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

setting it equal to 0:

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

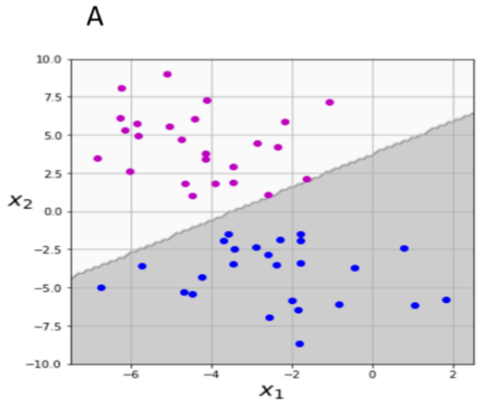
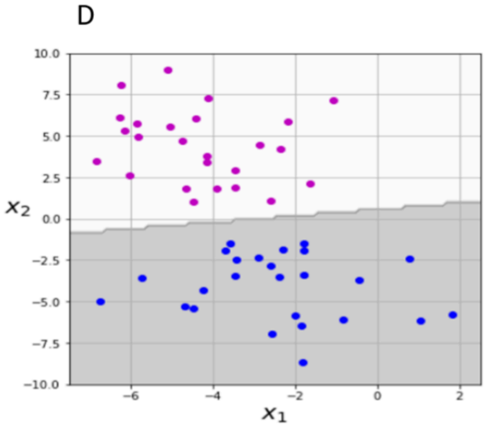
This equals to zero only when the number of positive items equals the number of negatives which happens when:

$$\mu=median\{x_1,x_2,\ldots x_m\}$$

Thus, we've proved that the centroid μ, which minimizes the term is the median of m examples and belongs to the dataset.

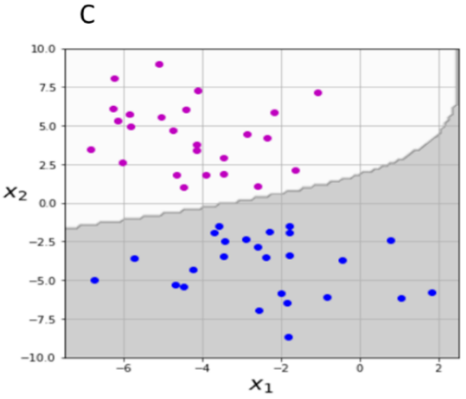
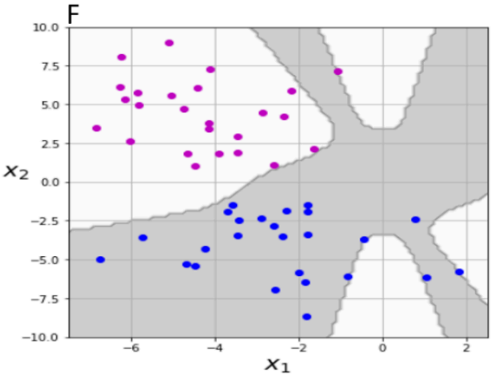
2. SVM

Linear kernels - A and D:

C=0.01	C=1
	

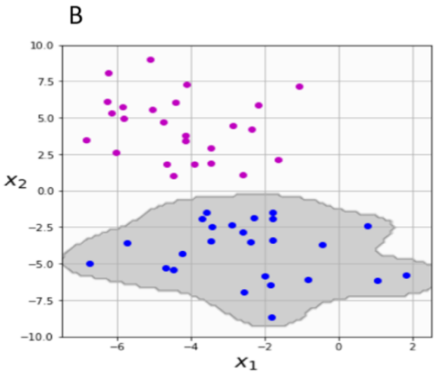
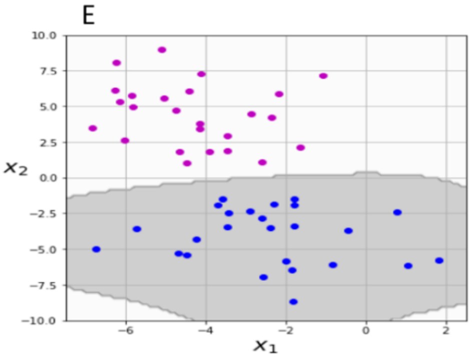
Explanation: In Sklearn term 'C' represents penalty, how much tolerance we want to give when finding the decision boundary. Bigger the 'C', the more penalty SVM gets when it makes misclassification. Therefore, the narrower the margin is and fewer support vectors the decision boundary depends on. In our case we don't see the margins, only the boundary. Small value of C (0.01) matches image A, because it causes the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. In case A there are 2 such points, which lie almost on the boundary line, this trade-off wouldn't be preferable for higher values of C.

Polynomial kernels – C and F:

2nd order polynomial kernel	10th order polynomial kernel
	

Explanation: Order parameters controls the flexibility of the decision boundary. Higher degree kernels yield a more flexible decision boundary. Since we observe much more flexibility in image F, we can conclude that the boundary matches 10th order polynomial kernel. According to this logic, image C matches the 2nd order polynomial kernel.

RBF kernels – B and E

RBF kernel with $\gamma = 1$	RBF kernel with $\gamma = 0.2$
	

Explanation: γ is a hyperparameter used with the Gaussian RBF kernel, which is used as a similarity measure between two points. γ is the inverse of the standard deviation of the RBF kernel, it sets the radius of the area of influence of the support vectors. Low values of γ lead to high variance and big area of influence, as shown on picture E. High values of γ lead to smaller variance and smaller area of influence, as shown on picture B.

3. Capability of generalization

a. In machine learning aspect the balance that Einstein meant to - is the balance between overfitting and underfitting, how well the concepts learned by a machine learning model will translate to new observations (generalization). By generalization, we find the best trade-off between underfitting and overfitting so that a trained model obtains the best performance.

b. $AIC = 2p - 2\ln(L)$, terms are:
$2\ln(L)$ term - measure the model fit. The higher the number, the better the fit.
2p term - penalizes the model for being overly complex
When given a set of candidate models for the data, the preferred model is the one with the minimum AIC value

c. When the balance is violated there are two options that are likely to happen:

- Underfitting – statistical model cannot adequately capture the underlying structure of the data, parameters are missing
- Overfitting – model obtains a high prediction score on seen data and low one from unseen data. When a model becomes too complex, it is usually prone to overfitting

d. When we want to compare between statistical models for a given set of data AIC would estimate the quality of each model, relative to each other models. The best model will be the one that neither under-fits nor over-fits, and it will have the lowest AIC value, which represents the minimal information loss of a model.