Machine Learning

In Healthcare

336546

HW3

Noy Parti 312387798

Question 1:

a. The K-Means algorithm is sensitive to outliers – if some data point has an unusual value (in compare to other data points) it would be affected, that is why it is sensitive to the noise However, K-Medoids clustering algorithm does not take the mean value of the object, it replaces the mean of clusters with medoids which is the most centrally located object in a cluster, it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. And this distance metric reduces noise and outliers .
That is why **K-Medoids is more robust to noise.**

b.

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

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

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

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

Now we will check if we got a minimum by the second derivative:

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

Bonus:

First, we will divide the whole samples into two parts: the number of samples under μ: μ_{low} and the number of samples above it: μ_{high}, according to the definition of the absolute value:

$$f(\mu)=\sum_{i=1}^{m}|x_i-\mu|=\sum_{i}^{\mu_{low}}(x_i-\mu)+\sum_{i}^{\mu_{high}}(-x_i+\mu)=\sum_{i}^{\alpha}x_i-\sum_{i}^{\beta}x_i+(\beta\mu-\alpha\mu)$$

When $\alpha=\mu_{low}$ & $\beta=\mu_{high}$

Now we need to find the minimum: $\frac{d}{d\mu}f(\mu)=0\rightarrow\beta-\alpha=0\rightarrow\boldsymbol{\beta=\alpha}$

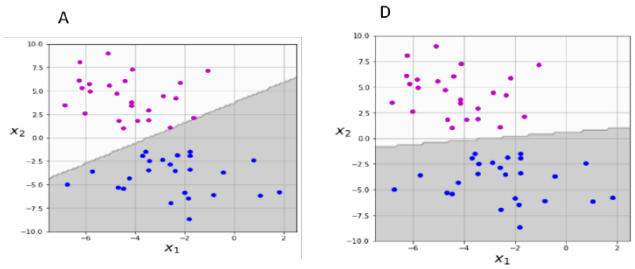
We can't use the second derivative to verify that this is a global minimum, however, $f(\mu)\geq 0\ \forall\mu$ and therefore, the extremum is a minimum point.
considering the definition of α,β we can tell that μ is the median- in order to show the μ minimizes the term we need to substitute:

$$f(\mu)=\sum_{i}^{\alpha}x_i-\sum_{i}^{\beta}x_i-\mu+\mu=0$$

Because the function is nonnegative for every value of μ- μ(which is the median) minimizes the given term.

Question 2:

Linear Kernels



First, we can see linear lines that separate between the two groups (linear SVM), C parameter determines the tradeoff between smooth decision boundary and accurate classification.

Higher C→ higher penalty for misclassification→ smaller margin.

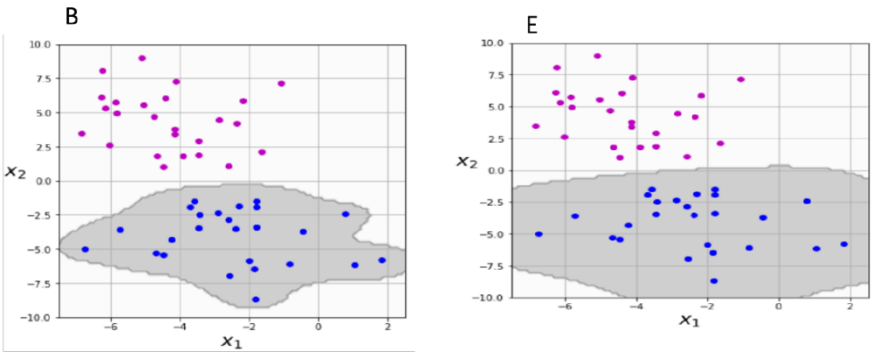
D-2: Linear kernel with C=1

in D we can't see any clear misclassification and that is correlated with higher C value

A-1: Linear kernel with C=0.01

the smaller the C is- the smaller the penalty. It allows more misclassifications which allow a larger margin- in A we can see two points that are almost misclassified- that indicates that the penalty for misclassification is not so high.

RBF Kernels



In B & E we can see RBF kernels with different gamma value. (RBF uses normal curves around the data points- matches both figures)

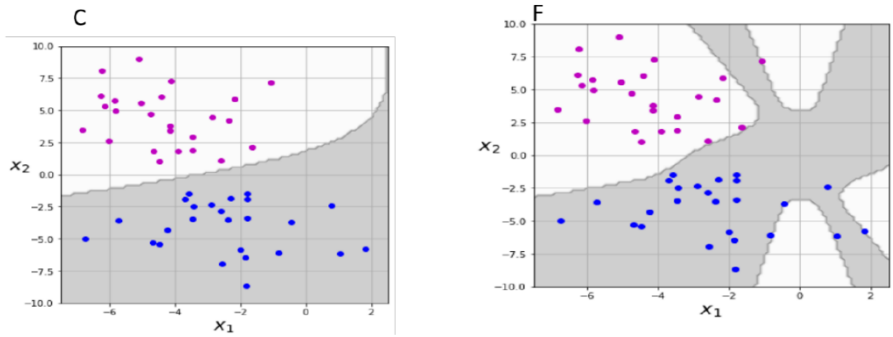
When Gamma is very small $\gamma = 0.2$ the model is too constrained and cannot capture the complexity or "shape" of the data. The resulting model will behave similarly to a linear model with a set of hyperplanes that separate the centers of high density of any pair of two classes.

That is why E matches SVM with RBF kernels $\gamma_E = 0.2$

However when gamma receives an intermediate values ($\gamma = 1$) the models perform better, creating a defined shape, if gamma will be too large the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with c will be able to prevent overfitting.

That is why B matches SVM with RBF kernels $\gamma_B = 1$

Polynomial Kernel



C- we can see a separation curve that resembles to 2nd order polynomial, while in F we can see overfitting that is very common in high (10th) order polynomial kernel- the system memorizes instead of learning.

Moreover, the degree parameter controls the flexibility of the decision boundary- higher kernel yields a more flexible decisions boundary as shown in F while C have less flexibility in the decision boundary.

To sum up:

A-1 , B-6, C-3, D-2, E- 5, F-4

Question 3:

a. The scientific term of balance in machine learning aspect is the **Generalization**- Generalization describes the ability of the model to react to new data after being trained on a training set, a good model will allow an accurate prediction (i.e – the simple solution is to train the model well), however, if a model has been trained too well on training data it would be unable to generalize. It will make inaccurate predictions when given new data, making the model useless even though it is able to make accurate predictions for the training data. This is called overfitting (this is an example of the simpler part in Einstein sentence- this is the balance of training the model).
to sum up- the aim is to find a balance between the complexity and the performance of the model.

b. $AIC = 2p - 2\log(\hat{L})$

two parts of the AIC:

- 2p- increases as we increase the number of parameters in our model-this penalizes us for building models that are complicated.
- 2log(L)- increases as our model gets better at explaining our data this rewards us for building models that fit out data well (rewards us for building accurate models)

c. If the balance was violated, we are facing two options:
- **Overfitting**- A simpler hypothesis representation is less prone to overfitting (that means that if the balance is violated and our model is too complex- we might risk overfitting).
- **Underfitting**- the model is not trained enough on the data- the model is useless is not capable of making accurate predications (I.e- the model is too simple, and the predictions are not accurate)

d. A low AIC value will indicate a better fit- from the one hand it will not be too complex (and we can avoid over-fitting of the data) and from the other hand it will explain our data good enough to get good and accurate predictions.