Machine Learning in Healthcare

HW no.3

Theoretical Questions

Ofek Aloush

307841742

Question 1:

Section a:

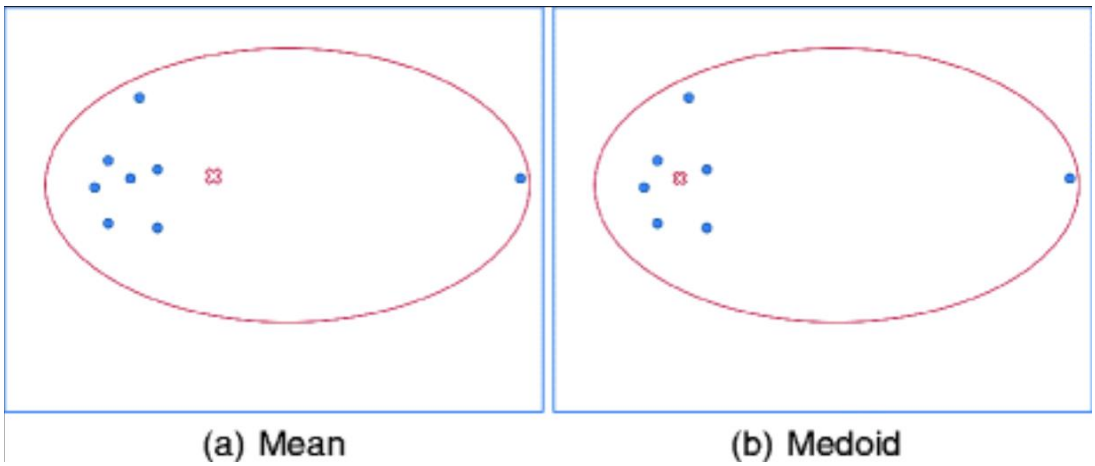
K-medoid is more robust to outliers than K-means algorithm.

As we know, mean is highly influenced by extreme values. That makes the K-means clustering algorithm to be pretty sensitive to outliers. On the contrary, K-medoids clustering is a variant of K-means that is more robust to noises and outliers.

Let's imagine a situation where we have data with 9 points, and we apply K-means on it. Then, we add a 10^{th} outlier point, one that it is very far away from the other 9 points. This extreme point is going to make a big difference on the mean value (as said, mean value is very sensitive to extreme values).

In contrary to K-means algorithm, instead of using the mean point as the center of a cluster, K-medoids uses an **actual point** in the cluster to represent it. Therefore, it won't matter so much whether the outlier 10^{th} point is close or 100 kilometers away from the other 9 points. The medoid will stay somewhere around the 9 central points. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points.

These images show the difference between mean and medoid in a 2-D example:



Section b:

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

We want to find the term that minimizes (*), and therefore we shall apply the differential operator, $\frac{d}{d\mu}$:

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

And then compare it with 0:

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

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

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

⇒ The centroid μ is the mean of the m examples.

Let's proof that the point we found is a minimum point. We'll do it by finding the second derivative:

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

⇒The relevant point is a global minimum (and as we found, it is the mean of x_i).

Bonus:

It is given that μ is a part of the dataset, and we want to find the minimum of:

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

Median is defined as the sample in which the number of samples above it, is equal to the number of samples underneath it. So basically, if we want to show that μ is the median, we need to show that it is correlated with the median's definition.

First, we divide the whole samples into three parts: μ_{low}, μ_{high} and μ. The first two are the number of samples under μ and the number of samples above it, respectively. That means, that if we show that $\mu_{low}=\mu_{high}$ then indeed μ is the median. (*)

$$f(\mu)=\sum^{\mu_{low}}|x_i-\mu|+\sum^{\mu_{high}}|-x_i+\mu|+|\mu-\mu|$$

For the sake of ease, I will mark: $\alpha = \mu_{low}, \beta = \mu_{high}$

$$f(\mu) = \sum_{1}^{\alpha}(x_i - \mu) + \sum_{1}^{\beta}(-x + \mu) = \sum^{\alpha} x_i - \sum^{\beta} x_i + (\beta\mu - \alpha\mu)$$

Using the differentiator operator: $\frac{d}{d\mu}$ and comparing the result with 0:

$$\frac{d}{d\mu} f(\mu) = \beta - \alpha = 0 \Rightarrow \quad \beta = \alpha$$

$\frac{d}{d\mu}\frac{d}{d\mu} f(\mu) = 0$ => we can't tell whether it is a minimum or maximum by the second derivation. However, we can see that the function $f(\mu)$ is nonnegative for any value of μ, therefore the extremum is a minimum point. We achieved that:

$$\alpha = \beta$$

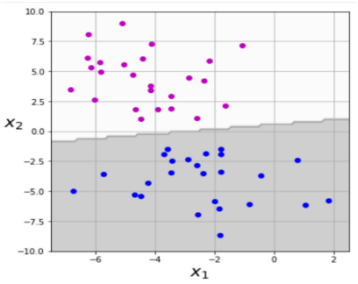
And considering the definition of $\alpha \; and \; \beta$, and as said in (*) – we can tell that μ is the median.

Question 2:

Linear Kernels:

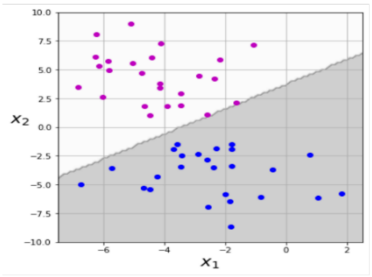
The C hyper-parameter in SVM algorithm controls the tradeoff between smooth decision boundary and classifying training points correctly.
In conclusion: Bigger C → Higher penalty for misclassification → smaller margin. And vise-versa.

D – 2: Linear kernel with $C = 1$.



The higher the C is, the bigger the penalty is for misclassifications, which makes the margin smaller. We can't see any clear misclassification points and that is more correlated with a bigger C (as explained).

A – 1: Linear kernel with $C = 0.01$.

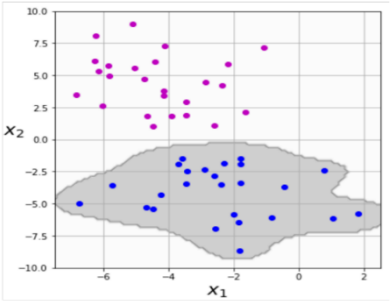


The smaller the C is, the smaller the penalty is. It enables the model to make more misclassifications, which enables a larger margin. plot A show 2 pink points which are "almost misclassified". That indicates that penalty for misclassification is small and that the margin can be large.
Now, if there was a blue dot that was correlated with one of the pink dots that I spoke about earlier, we might could tell that the margin is small, but this is not the case. In conclusion, we can see that plot A has "almost-misclassified" points which is more correlated with $C = 0.01$.

RBF Kernels:

The gamma hyper-parameter in SVM defines how far the influence of a single training example reaches.
Low values: every point has a far reach. High values: every point has a close reach.
In other words, as the value of Gamma increases, the model gets overfitted, and as the value of Gamma decreases, the model underfits.
In conclusion, SVM with RBF kernel has usually a circular characteristic, and the gamma parameter defines how roundish/curvy the decision boundary gets.

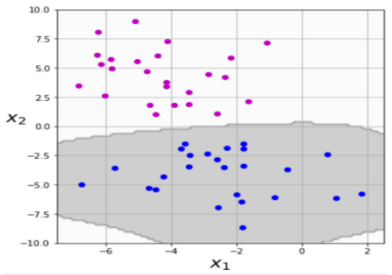
B – 6: RBF with $\gamma = 1$.



We can see that plot B describes an overfitted model, thus it has a bigger gamma variable.

E – 5: RBF with $\gamma = 0.2$.

The E plot has a RBF kernel characteristic form, and therefore we can tell that it was made with RBF kernel. However, it is not overfitted as the B plot, thus its Gamma parameter is: $\gamma = 0.2$.

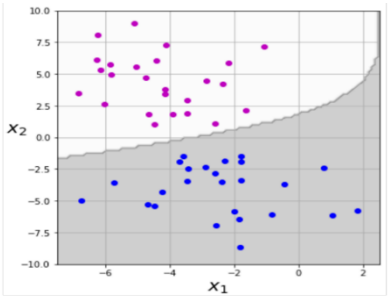


<u>Polynomial Kernels:</u>

Polynomial kernels doesn't form a round shape or a closed one. In addition, the higher the kernel's degree is, the more complex the model is.

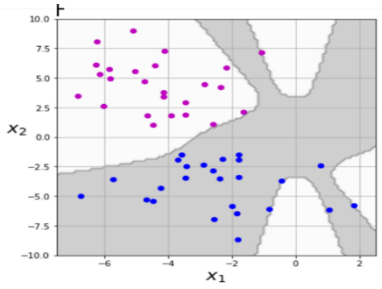
C – 3: 2nd order polynomial kernel.

The plot shows a polynomial-parabolic form.



F – 4: 10th order polynomial kernel.

Larger degrees models tend to be more overfitted and more complex. This figure shows a more complex shape.



Question 3:

Section a:

The scientific term that Einstein meant is **generalization.** Generalization is the model's ability to adapt properly to new data based on factors such as the complexity of the model and the model's performance. There are several assumptions regarding the data: it must be independently and identically (I.I.D), stationary, and the testing data must come from the same distribution as the training one. In practice, these assumptions can be violated sometimes.

In general, we can tell that the less complex a model is, the more likely that a good empirical result is not just due to the peculiarities of our sample. Generalization is based on ideas of measuring model simplicity / complexity. In conclusion, the aim is to find a balance between the complexity and the performance of the model.

Section b:

AIC is defined by: $AIC = 2p - 2\ln(\hat{L})$

We can break AIC into two parts: $2p$ and $2\ln(\hat{L})$

1. The first part, 2p, increases as we increase the number of parameters. This penalizes us for building models that are complicated. More parameters = more complexity.
2. The second part, $2\ln(\hat{L})$, increases as our model gets better at explaining our data. This rewards us for building models that fit our data well.

Section c:

Too much complexity will result in an overfitting model, while model which is not explanatory enough will result in under-fitting.

Section d:

The smaller the better. We prefer model that minimize AIC. AIC cannot tell us whether our model perform good or not, it can only tell us about if it strikes a better balance between model complexity and explanatory power than other models do. Therefore, as section b explains, smaller AIC means that the model is both less complex and both more explanatory than another model with a higher AIC.

Bibliography:

1. https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_426
2. Google's developers' school.
3. Class Lectures.