HW3

Marina Tulchinsky, 342399862

1 Clustering (10%)

In the lecture we saw the K-means algorithm for clustering. It tries to minimize the Euclidian metric between the examples and some point in space which is named "centroid". Other methods try to minimize dissimilarities between the pairwise data examples. A classical algorithm which was designed to handle pairwise data is the K-medoid. This algorithm seeks to find a set of cluster representatives (named medoid) in the dataset and assign other examples to them. The algorithm randomly picks a k-set of medoids from the data and assigns points to each medoid based on their L1 distances to that medoid. Then, it iteratively tries to improve the assignment by swapping assigned medoid points with non-medoid points until the energy of the entire system (which is measured by the sum of distances between medoid points and their assigned data points) is minimized.

a. Is K-medoid more robust to noise (or outliers) than the K-means algorithm? Explain your answer.

ANSWER: K-medoid algorithm is based on L_1 distances ($\sum_i | x_i | = \| x \|_1$). So, if we have noise or outliers, then it is not squared, as it happens in L_2 ($\sqrt{(\sum_i x_i^2)} = \| x \|_2$), which means it has less effect and do not increase the noise. Thus K-medoid is more robust to noise than K-means algorithm.

b. Prove that for the 1D case ($x \in R1$) of K-means, the centroid (μ) which minimizes the term $\sum_{i=1}^{m}(x-\mu)^2$ is the mean of m examples.

ANSWER: To find the μ which minimizes the term we need to take the derivative of the term $\sum_{i=1}^{m}(x-\mu)^2$ with respect to μ and to equate the term to zero:

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

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

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

$$\mu = \frac{\sum_{i=1}^{m}x}{m}$$ - mean of m examples.

To prove that μ is minimum and not maximum we need to check that the second derivative is positive: $-2 \cdot (-m) = 2m > 0$ => the mean of m examples minimizes the term $\sum_{i=1}^{m}(x-\mu)^2$.

c. Bonus: Prove that the centroid (practically, the medioid) which minimizes the term $\sum_{i=1}^{m}|(x_i-\mu)|$ is the median of m examples given that μ belongs to the dataset.

ANSWER: First, we must distribute x so that $x_1 < x_2 < ... < x_m$. Assume that m is odd and μ is between $x_{\frac{m-1}{2}}$ and $x_{\frac{m+1}{2}}$ (otherwise, μ will not be cancelled).

Let us pull out x_1 and x_m ,and look at the expression we get:

$$\sum_{i=1}^{m}|(x_i-\mu)| = \sum_{i=2}^{m-1}|(x_i-\mu)| + |x_1-\mu| + |x_m-\mu| = \sum_{i=2}^{m-1}|(x_i-\mu)| + (\mu-x_1) + (x_m-\mu) = \sum_{i=2}^{m-1}|(x_i-\mu)| + (x_m-x_1)$$

If we continue the process, we will get:

$$\sum_{i=1}^{m}|(x_i-\mu)| = (x_m-x_1)+(x_{m-1}-x_2)+...+(x_{\frac{m+3}{2}}-x_{\frac{m-1}{2}})+(x_{\frac{m+1}{2}}-\mu)$$

To find the centroid which minimize the term we need to find μ that will give zero derivative and positive second derivative.

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

From the expression above we see that minimal derivative is when $\mu = x_{\frac{m+1}{2}}$. The second derivation of the term is >0, so the μ is minimum.

Let us assume that m is even now. We need to find another method to prove that the centroid which minimizes the term $\sum_{i=1}^{m}|(x_i-\mu)|$ is the median.

When the m is even, median is usually calculated as $\mu = \frac{x_{\frac{m}{2}} + x_{\frac{m+1}{2}}}{2}$.

In the case when m is even, we start the formulation as in the previous case (when m was odd).
We again assume that x_i is sorted from the smallest to the largest and the assumption for the μ is the same as with the odd m (due to the same reason):

$$\sum_{i=1}^{m}|(x_i-\mu)| = \sum_{i=2}^{m-1}|(x_i-\mu)| + (\cancel{\mu}-x_1)+(x_m-\cancel{\mu}) = ... = const+(\cancel{\mu}-x_{\frac{m}{2}})+(x_{\frac{m+1}{2}}-\cancel{\mu}) = const_+$$

We will get the upper μ cancellation only if $x_{\frac{m}{2}} < \mu < x_{\frac{m+1}{2}}$.

So, knowing that $x_{\frac{m}{2}} < \mu < x_{\frac{m+1}{2}}$ we will show that $\mu = \frac{x_{\frac{m}{2}} + x_{\frac{m+1}{2}}}{2}$:

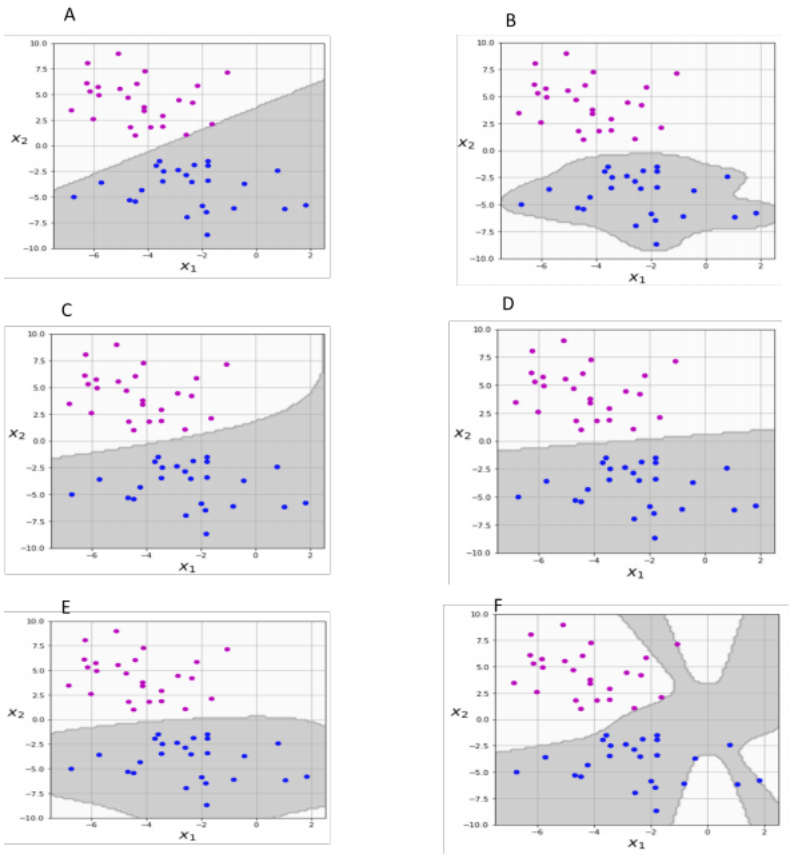
Derivative of the expression $\sum_{i=1}^{m}|(x_i-\mu)| = const + |x_{\frac{m}{2}}-\mu| + |x_{\frac{m+1}{2}}-\mu|$:

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

The derivative equals zero if $x_{\frac{m}{2}}-\mu = \mu - x_{\frac{m+1}{2}} \Rightarrow \mu = \frac{x_{\frac{m}{2}} + x_{\frac{m+1}{2}}}{2} = median$.

2 SVM (30%)

In the following figures you can see a visualization of SVM running with different settings (kernels and parameters) as follows:



The settings that were used are as following:

1. Linear kernel with C = 0.01.

2. Linear kernel with C = 1.

3. 2nd order polynomial kernel.

4. 10th order polynomial kernel.

5. RBF kernel with γ = 0.2.

6. RBF kernel with γ = 1.

Match every image (labeled by a capital letter) to its' setting (number). Explain each of your answers.

ANSWER: Linear kernel will give linear decision boundary (linear line), so figures A and D fit for settings 1 and 2. C is a regularization parameter. If C is small more misclassifications can be. In the figure A we can see that there are 2 purple points very close to the decision boundary (line). Because there are not any

blue points on the same distance but from the other side of the decision boundary, the points are not support vectors. So, we understand that the points are closer to the decision line than real support vector/s and it means that the regularization parameter is small. In the figure D we see that there are no misclassifications and points are not on the line (like in figure A), so C must be bigger than in the A figure. Thus, figure A is suitable for Linear kernel with C = 0.01 and figure D – for Linear kernel with C = 1.

Second order polynomial kernel can split our classification problem according to quadratic functions like circle, parabola, ellipse, and hyperbola. Figure C looks like part of ellipse or hyperbola, so I think it suits to 2nd order polynomial kernel.

RBF kernel is exponential kernel. Gamma is hyperparameter that regulate influence of individual sample on the position of the decision boundary. If gamma is big the influence of individual cases will be relatively small, the decision boundary will be based mostly on support vectors, so the area of decision boundary will be small. Thus, we understand that figure B is RBF kernel with γ = 1, and figure E - RBF kernel with γ = 0.2.

From polynomial kernel ($K(x, y) = (x^T y + c)^d$), we can get complex shape of the decision boundary (bigger d -> the shape is more complex), like in the figure F, where d=10.

Summary:

Table 1

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

3 Capability of generalization (20%)

Ockham's razor states that "Non sunt multiplicanda entia sine necessitate". This is known as the law of parsimony and its' translation to English is "Entities are not to be multiplied without necessity". This concept, which is attributed mostly to the English Franciscan friar William of Ockham, basically means that the simplest explanation is usually preferred. A more modern variation of this concept (and a much more readable one) is attributed mostly to Albert Einstein and it states that "Everything should be made as simple as possible but not simpler". This balance is a major guideline in science in general and in data science in particular. We saw in the tutorial two methods of choosing a parsimonious model for K-means. In GMM, there is another criterion to do so and it is known as "Akaike information criterion" (AIC). It is composed of two terms and defined as follows: $\mathrm{AIC} = 2\mathrm{p} - 2\ln\left(\hat{\mathrm{L}}\right)$.

where p is the total number of learned parameters and $\hat{\mathrm{L}}$ is the estimated likelihood given these parameters.

a. What is the scientific term of the balance that Einstein meant to in machine learning aspect?

ANSWER: The term is generalization. It refers how well learning a model will generalize to new observations: it means that our model is not too simple and is not too complicated and fitted too much to the train data. (lecture C6 slide 5)

b. How does each of the terms (2p, 2ln ($\hat{L}$)) in AIC affect the terms of the balance you defined in (a)?

ANSWER: Maximum value of likelihood function ($\hat{L}$) relates to the "goodness of fit". The number of parameters gives "penalty". Because usually if the number is big, the $\hat{L}$ will be bigger too, and it leads to overfitting and high AIC. The similar situation is with underfitting: low number of parameters usually lead to low $\hat{L}$ (so the $-2\ln\left(\hat{L}\right)$ will be big). The balance is when the number of features is not too big but the $\hat{L}$ is big enough.

c. What are the two options that are likely to happen if this balance was violated?

ANSWER: The options are overfitting and underfitting.

d. What are we aiming for with the AIC? Should it be high or low? Explain.
ANSWER: We are aiming to get low AIC, because extreme cases will lead to high AIC (as said before). Additional example: if maximum value of likelihood =1, the overfitting occurred, so the AIC = 2p. Because usually the overfitting occurs when we have many features the AIC will be high.

References:

https://dzone.com/articles/using-jsonb-in-postgresql-how-to-effectively-store-1

https://www.kaggle.com/residentmario/l1-norms-versus-l2-norms

https://en.wikipedia.org/wiki/K-medoids

https://machinelearningwithmlr.wordpress.com/blog-feed/

https://en.wikipedia.org/wiki/Polynomial_kernel