HW3-Theoretical questions

Orel Shahadi, ID: 206092090

1 Clustering

a. The K medoid method is much more robust to noise than the K means method.

K means is much more affected by extreme data points (the outliers). It means that in the K means method, the center of the centroid is determined by squared Euclidean distances, so that the center of the cluster does not have to be a specific data point, thereby much more affected by noise. In contrast, the K medoid method select the center of the cluster according to specific points in the given data, the medoids, by calculating the minimum sum of distances of a point from other points in the cluster. The point that gives the minimal sum of distances will be chosen as the center of the cluster. This makes the method more robust t noise than K means method.

b. For the 1D case ($x \in R^1$), we will find the centroid (μ) that minimizes the term $d(x_i, \mu) = \sum_{i=1}^{m}(x_i - \mu)^2$:

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

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

$$\boxed{\rightarrow \mu_{min} = \frac{1}{m}\sum_{i=1}^{m} x_i}$$

c. Bonus:

Assumptions:

- $(n \, mod \, 2) = 0$
- $x_i < x_{i+1}$
- $x_{\frac{m}{2}} < x_{\frac{m}{2}+1}$

$For \;\; x_i < x_{i+1}, (n \, mod \, 2) = 0, the \; median \; \in \left(x_{\frac{m}{2}}, x_{\frac{m}{2}+1}\right)$

$Let \; m \epsilon \; \mathbb{R} \; that \; can \; be \; the \; median: x_{\frac{m}{2}} < \mu < x_{\frac{m}{2}+1}$

$So \; we \; need \; to \; proof \; that \; for \; every \; number \; a \in \mathbb{R}:$

$$\frac{1}{m}\sum_{i=1}^{m}|x_i - \mu| \leq \frac{1}{m}\sum_{i=1}^{m}|x_i - a|$$

$\rightarrow \frac{1}{m}\sum_{i=1}^{m}(|x_i - a| - |x_i - \mu|) \geq 0$

$We\ can\ define\ three\ groups\ of\ indexes:$

$A = \{i: x_i < a\}, \quad B = \{i: a < x_i \leq \mu\}, \quad C = \{i: x_i > \mu\}$

$\underline{for\ i \in A}:$

$|x_i - a| - |x_i - \mu| = a - \mu$

$\underline{for\ i \in B}:$

$|x_i - a| - |x_i - \mu| = 2x_i - a - \mu \geq a - \mu$

$\underline{for\ i \in C}:$

$|x_i - a| - |x_i - \mu| = \mu - a$

$\rightarrow \frac{1}{m}\sum_{i=1}^{m}(|x_i - a| - |x_i - \mu|) = \frac{1}{m}(\sum_{i\in A}(|x_i - a| - |x_i - \mu|) + \sum_{i\in B}(|x_i - a| - |x_i - \mu|) + \sum_{i\in C}(|x_i - a| - |x_i - \mu|))$

$$= \frac{1}{m}(\sum_{i\in A}(a - \mu) + \sum_{i\in B}(2x_i - a - \mu) + \sum_{i\in C}(\mu - a)) \geq$$

$$\geq \frac{1}{m}\left(\sum_{i\in A}(a - \mu) + \sum_{i\in B}(a - \mu) + \sum_{i\in C}(\mu - a)\right) = \frac{\mu - a}{m}[|C| - (|A| + |B|)] =$$

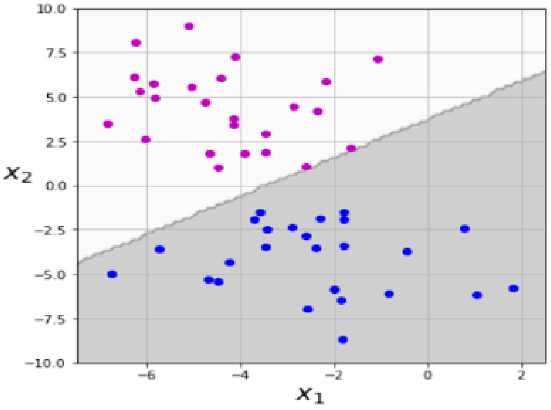
$$= \frac{\mu-a}{m}\left(\frac{m}{2} - \frac{m}{2}\right) = 0$$

$therfore: \frac{1}{m}\sum_{i=1}^{m}(|x_i - a| - |x_i - \mu|) \geq 0$

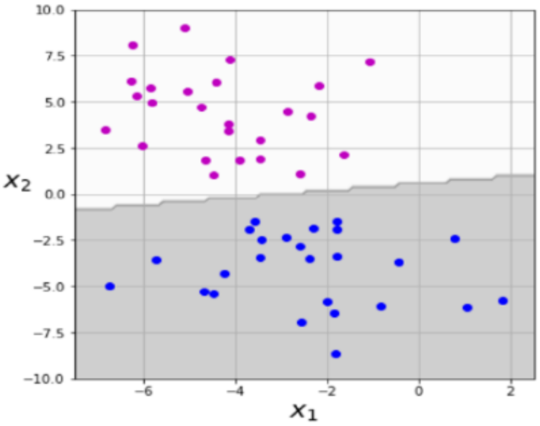
$\frac{1}{m}\sum_{i=1}^{m}|x_i - \mu| \leq \frac{1}{m}\sum_{i=1}^{m}|x_i - a|$

2 SVM

A



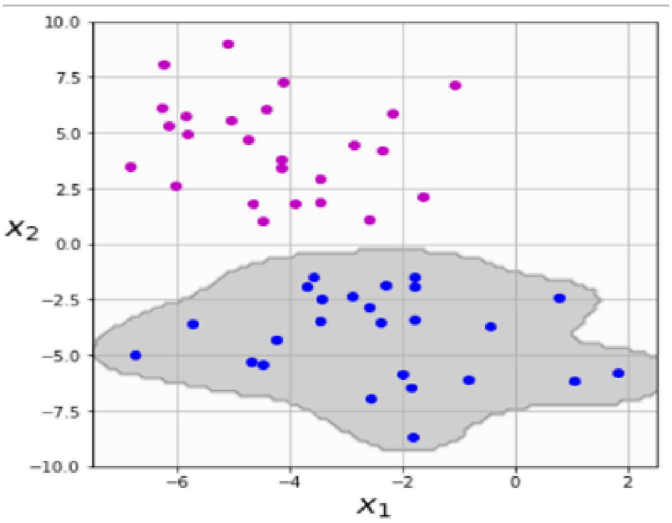
D



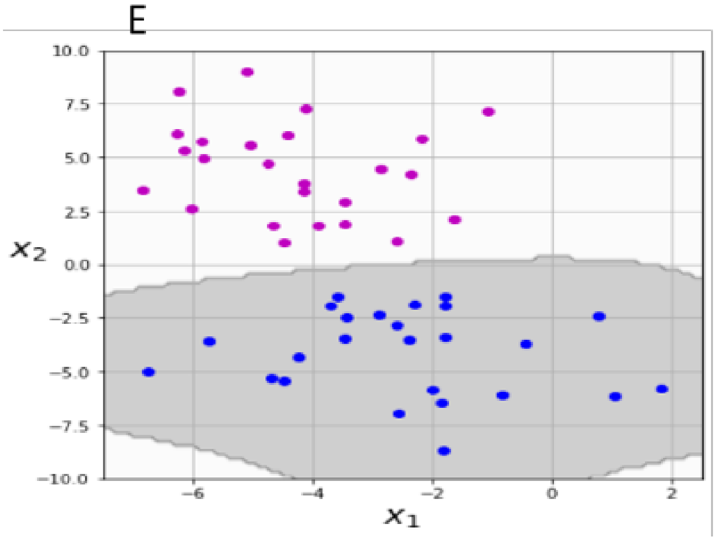
In both images there is a need of linear classification → linear kernel.

The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For a higher C value, we will get a smaller margin hyper plane. For a smaller C value, we will get a larger-margin separating hyperplane, risking in misclassifying data points. Therefore, image A has C=0.01 **(1)**, image D has C=1 **(2)**.

B

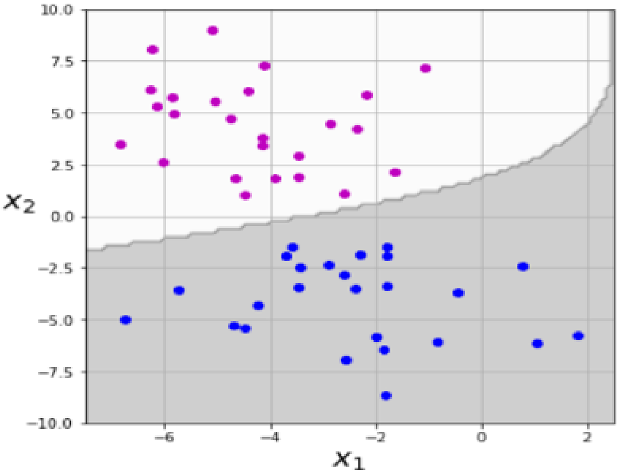


E

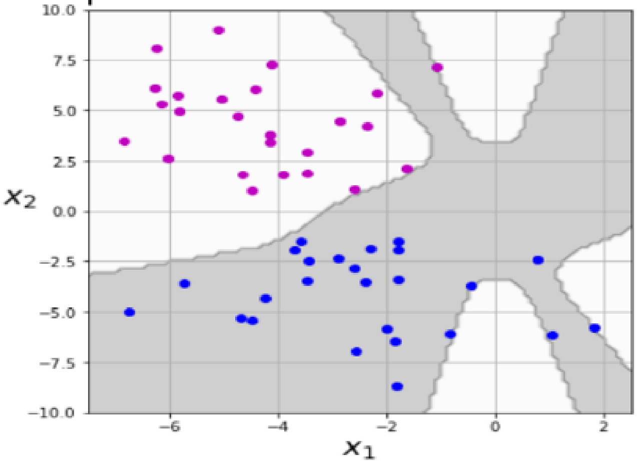


RBF kernel has similarity to Gaussian distribution, and we can see here a typical shape of this classifier. γ sets the "spread" of the kernel: $\gamma = \frac{1}{2\sigma^2}$.
As γ increases, σ decreases and hence, only if points are extremely close they are considered similar. Also, As γ decreases, farther away points can be considered to be similar. Therefore, image B has lager gamma: $\gamma = 1$ **(6)**, and image E has the smaller gamma: $\gamma = 0.2$ **(5)**.

C



F



In these images polynomial classifier was used. We can see that because of the separation line between the different labels: It looks like a linear combination of features. Also, we can tell that image F is much more overfitted than image C, therefore image F is given for 10^{th} order polynomial kernel **(4)**, and image C is given for 2^{th} order polynomial kernel **(3)**.

In total:

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

3 Capability of generalization

a. Einstein means the ability to generalize. When a model is too simple, it leads to errors and underfitting. On the other hand, a too complexed model, will lead to overfitting and high variance as we have seen in lectures and tutorials.

b. 2p- represents the number of learned parameters. For its high value, the greater the chance to get overfitting of the system and lose generalization capability.
 $-2ln\ (L)$ expresses the estimated likelihood given by the learned parameters. For its high value, there is a good fit of the model to the data.
In addition, for a complexed model with a high number of learned parameters, it adds penalty to avoid overfitting of the model. Our ambition is to increase the $-2ln\ (L)$ for which we will get a lower AIC.

c. According to the explanation in section a, a violation of the balance may cause overfitting in case there are many learned parameters but low $-2ln\ (L)$ term, leading to a high AIC. On the other hand, when we use a small number of learned parameters (too simple a model that will lead to underfitting), the likelihood function will also decrease and we will get a high AIC again.

d. Our goal is to reduce the AIC as much as possible. As stated above, small AIC expresses a balance between the two extremes, thus prevents underfitting or overfitting. A small AIC expresses a balance between the fit of the model and its generalization capability.