HW3

1 Clustering (10%)

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

The K-means algorithm is more affected by outliers. K-means algorithm uses the mean of all the xi's in the cluster in order to calculate the C and as we know, mean is very sensitive to outliers, one unusual event will change it completely. Alternatively, K-medoid chooses the centroid to be one of the existing data points depending on which one has minimum loss (difference between each point in the cluster to the centroid point) therefore it is less sensitive. Meaning, the k-medoid is more robust to noise.

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

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

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

$$0 = -2\sum_{i=1}^{m} x_i + 2 * m * \mu$$

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

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

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.

Let's assume that Xi is sorted from smallest to largest, Lets prove this the other way around, we will assume that $\mu = x_{\frac{m+1}{2}}$ which is the median, and show that the derivative is equal to zero.

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

If we assume that μ is the median and x is sorted, the values from x_1 till $x_{m/2}$ will be smaller than μ (we can remove the absolute value and add a minus) and the values from $x_{\frac{m+1}{2}}$ till x_m will be greater than μ.

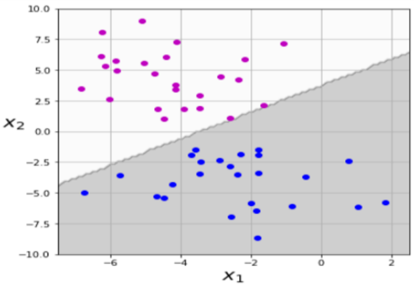
$$\frac{d}{d\mu}(\sum_{i=1}^{m}|x_i-\mu|)=\frac{d}{d\mu}(-\sum_{i=1}^{\frac{m}{2}}(x_i-\mu)+\sum_{i=\frac{m+1}{2}}^{m}(x_i-\mu))=$$

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

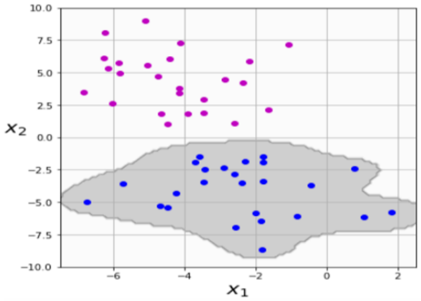
2 SVM (30%)

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

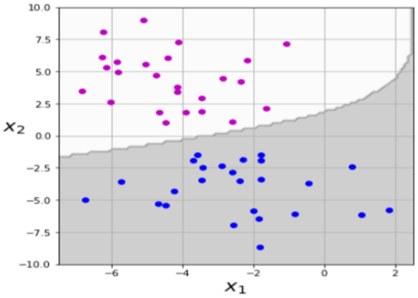
A



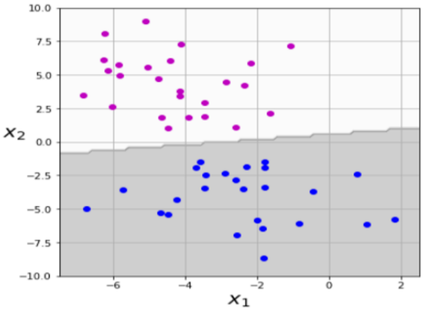
B



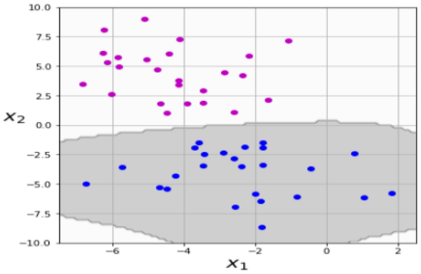
C



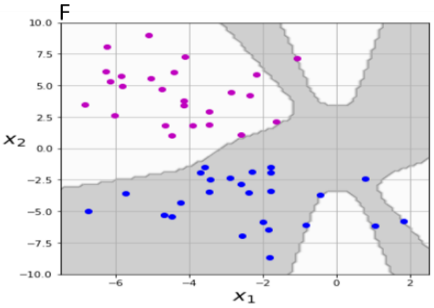
D



E



F



The settings that were used are as following:

1. Linear kernel with $C = 0.01$. **A**

   We can see that A is linear, we chose a small C because the fitted line is rather close to two of the purple observations, meaning the purple observations are inside the margins and there is less "punishment" for misclassification.

2. Linear kernel with $C = 1$. **D**

   Here, as opposed to in A, the fitted line is very much in the middle of the two colors, it is a better fit to the data so we can assume that the C is larger, punishes for misclassification.

3. 2^{nd} order polynomial kernel. **C**

   In this graph we can see a sort of parabola which fits with second degree polynomial kernel. The fit of the observation is rather general (not overfitted).

4. 10^{th} order polynomial kernel. **F**

   This graph matches a high order polynomial kernel because it is extremely over-fitted, the shape has branches which is a common thing that happens with overfitting because the area comes to cover the training data perfectly.

5. RBF kernel with $\gamma = 0.2$. **E**

   RBF kernels are similar to a Gaussian distribution. In this case, the kernel is not overfitted but quite general, it has a general shape without any branches so we can infer that gamma is small.

6. RBF kernel with $\gamma = 1$. **B**

   The data looks over-fitted. The shape is conformed to match the observations (not an ellipsoid but a funky shape). This indicates a high gamma (represents goodness of fit).

3 Capability of generalization (20%)

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

   The term is generalization. The ability of the model to adapt to new data from the same distribution as the trained data and to find a balance between goodness of fit and complexity.

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

   P is the number of learned parameters meaning if p is too large you could be looking at overfitting which is why p is with a positive sign in the AIC equation. L on the other hand, has a negative sign and then put into ln. This means that larger numbers of L are "rewarded" (will make the AIC smaller). This fits with our theory that L represents goodness of fit and the better the fit, the smaller the AIC.

c. **What are the two options that are likely to happen if this balance was vio-lated?**

   Overfitting and underfitting.

d. **What are we aiming for with the AIC? Should it be high or low? Explain.**

We would like AIC to be low. We would like a balance between goodness of fit and overfitting therefore the closer the AIC is to zero the better (perfect balance).

4 EigenFaces (40%)

In the eigenfaces notebook