



MLH-HW3

Shay Ohana 315800375

Q1-A

The K-medoids is more robust to noise and outliers than K-means because it tries to minimize the dissimilarities between the pairwise data examples i.e. to minimize the sum of distances between medoid points and their examples, while K-means tries to minimize the sum of squared Euclidean distances between μ and the observations. Therefore, outlier will affect the Euclidean distances more and K-medoids is more robust.


Bibliography:

http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

Q1-B

We want to prove that $C = \mu$ minimize the term: $\sum_{i=1}^m (x_i - C)^2$.

$$\begin{aligned}
S &= \sum_{i=1}^m (x_i - C)^2 = \sum_{i=1}^m (x_i - \mu + \mu - C)^2 = \\
&= \sum_{i=1}^m (x_i - \mu)^2 + 2 \cdot \sum_{i=1}^m (x_i - \mu)(\mu - C) + \sum_{i=1}^m (\mu - C)^2 = \\
&= \sum_{i=1}^m (x_i - \mu)^2 + 2 \cdot (\mu - C) \cdot \underbrace{\sum_{i=1}^m (x_i - \mu)}_{=0 (*)} + m \cdot (\mu - C)^2 \\
&= \sum_{i=1}^m (x_i - \mu)^2 + m \cdot (\mu - C)^2
\end{aligned}$$

 $\xrightarrow{\text{find } C \text{ that minimize}} \frac{dS}{dC} = m \cdot \frac{d(\mu - C)^2}{dC} = m \cdot (-2 \cdot (\mu - C)) = 0 \rightarrow C_{min} = \mu$

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

Q2-SVM

A, D separate the groups linearly, as expected from linear kernel. A allowing the examples to be closer to the vector (narrow margins, the error takes on less meaning) and therefore its param C is lower. So A is settings #1 and D is settings #2.

B, E separate in circle/ovals shapes, similar to the Gaussian projection on 2D and therefore they are from RBF kernel. In B the boundaries are closer to the data, this means that the variance of the Gaussian is smaller i.e. gamma (which is inversely related to the variance) is greater. So B is #6 and E is #5.


F, C having complex and undefined boundaries as expected from polynomial kernel. A higher level of complexity suggests on higher order of the polynom. So F is #4 and C is #3.

Q3-A

The scientific term is generalization, because we want to simplify the model as much as possible but it is impossible to simplify too much because then the algorithm will not be able to learn well and make a generalization from the learning phase to the phase when it gets new examples.

Q3-B

AIC is A measure that helps to choose the best model from several GMM models with different parameters. It balances the trade-off between a good fit of the model to the data and an over-fitting that limits the ability to generalize, the better the model the smaller the AIC. The meaning of a larger likelihood -L- means that the probability to belong to specific cluster is high and the probabilities to belong to the other clusters are low i.e. the level of confidence in the correctness of the classification is high and the model fits the data better and therefore we would like L to indicate that the model is good (remember that although the model is properly classified it is prone to overfitting) i.e. will reduce the AIC. When P is larger the model relies on higher number of learning parameters and is prone to overfitting so we would like it to indicate that the model is not so good, and increase the AIC. Therefore

 adds to AIC (positive sign) and L subtracts from AIC (negative sign).

Q3-C

If this balance is disturbed it is possible that:

1. Overfitting- The model is too adapted to the data it has studied and therefore its ability to generalize is limited and will not cope well with new data.
2. Underfitting- The model is too general and has not studied enough properties of the data (e.g. needs greater complexity / more features) and fails to classify at a sufficient level.

Q3-D

As I explained in section B, we are aiming for small AIC because small AIC means that the likelihood is high so the model fits the data good and the number of clusters really fits the data, so the chances to overfitting/underfitting are low.