



HW3

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

K-medoid and K-means are methods of data partition that attempt to minimize the distance between points in the cluster and the center of the cluster. K-medoid, chooses the actual data points as cluster centers. As a result, it is easier to interpret the cluster centers in this method, and therefore, K-medoids is better at clustering data when there are outliers. On the other hand, K-means chooses any center that minimizes the sum of squares, so it is more influenced by outliers. So, in conclusion, K-medoid is more robust to noise (or outliers) than the K-means algorithm

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

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



$$\frac{\partial D}{\partial \mu} = -2 * \sum_{i=0}^m (x_i - \mu) = 0$$

$$\sum_{i=0}^m x_i = m * \mu$$

$$\mu = \frac{\sum_{i=0}^m x_i}{m} = \bar{x}$$

Bonus:

Assumptions

- m is even
- x vals are sorted by size

$$x_{\frac{m}{2}} < x_{\frac{m}{2}+1}$$

$$\text{median} \sim [x_{\frac{m}{2}}, x_{\frac{m}{2}+1}]$$

$$a < \text{med}$$

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

$$\forall i \in A : x_i < a < \text{med}$$

$$|x_i - a| - |x_i - \text{med}| = a - x_i - \text{med} + x_i = a - \text{med}$$

$$\forall i \in B : a < x_i \leq \text{med}$$

$$|x_i - a| - |x_i - \text{med}| = x_i - a - \text{med} + x_i = 2x_i - a - \text{med}$$

$$\geq 2a - a - \text{med} = a - \text{med}$$

$$\forall i \in C : x_i > \text{med} > a$$

$$|x_i - a| - |x_i - \text{med}| = \text{med} - a$$

$$\frac{1}{m} \sum_{i=1}^m (|x_i - a| - |x_i - \text{med}|) =$$

$$\frac{1}{m} \left(\sum_{i \in A} (|x_i - a| - |x_i - \text{med}|) + \sum_{i \in B} (|x_i - a| - |x_i - \text{med}|) + \right.$$

$$\left. + \sum_{i \in C} (|x_i - a| - |x_i - \text{med}|) \right)$$

$$= \frac{1}{m} \left(\sum_{i \in A} (a - \text{med}) + \sum_{i \in B} (\text{med} - a) + \sum_{i \in C} (|x_i - a| - |x_i - \text{med}|) \right) \geq$$

$$\frac{1}{n} \left(\sum_{i \in A} (a - \text{med}) + \sum_{i \in B} (a - \text{med}) + \sum_{i \in C} (\text{med} - a) \right) =$$

$$\frac{1}{m} ((\text{med} - a)|C| - ((\text{med} - a)(|A| + |B|))) = \frac{\text{med} - a}{m} (|C| - (|A| + |B|))$$

$$|C| = \frac{n}{2} = |A| + |B|$$

$$\frac{1}{m} \sum_{i=1}^m (|x_i - a| - |x_i - \text{med}|) \geq \frac{\text{med} - a}{m} (|C| - (|A| + |B|))$$

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

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

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

Therefore, the median will give us the minimal value for this expression.

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

The C parameter is responsible for the trades off of correct classification of training examples and maximization of the decision function's margin. For larger C values, we will accept smaller margins if the decision function is better at classifying all training points correctly. A lower C will encourage a larger margin, and therefore a simpler decision function, at the cost of training accuracy.

1. Linear kernel with C = 0.01. **A**- the data is separated by a linear line and soft margins are accepted (2 dots on the separating line).

2. Linear kernel with C = 1. **D**- the data is separated by a linear line and hard margins are accepted. The accuracy is better.

As we saw in the previous HW, RBF kernel gave us the highest accuracy and has more oval shape. The model is very sensitive to the gamma parameter. If it is too large, the radius of the area of influence of the support vectors only includes the model will be overfitted.

If it is too small, the model is underfitted. The region of influence of any selected support vector would include the whole training set. The resulting model will behave similarly to a linear model with a set of hyperplanes that separate the centers of high density of any pair of two classes.

5. RBF kernel with $\gamma = 0.2$. **E**- pretty accurate model yet underfitted and more linear shaped.

6. RBF kernel with $\gamma = 1$. **B**- pretty accurate, yet overfitted model.


Polynomial kernel is another method of nonlinear svm , it is less accurate than RBF in certain cases. In higher polynomial orders it tends to overfit.

3. 2nd order polynomial kernel .**C**- the data is separated by a parabola and the data is not overfitted.

4. 10th order polynomial kernel. **F**- the data is polynomial and the data is overfitted and has larger margins than in C.


3.

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

 The scientific term is generalization. We aspire to get a model with high performance and low complexity. As mentioned in the lecture, A simpler, less complex hypothesis representation is less prone to overfitting. If a model is more complex, it will be easier to generalize from new unlearned data and the performance will improve with the risk of overfitting. If the model is too simple it will result the risk of underfitting and the performance will decrease. Therefore, a balance is needed.

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

The total number of learned parameters affects the problems complexity. More parameters will provide us additional training examples that will increase the problems complexity. With that been said, there is a risk of irrelevant data that will affect the learning and additional risk of overfitting. On the other hand, low number of learned parameters will create a simpler hypothesis with a risk of underfitting. The log likelihood effects on the performance. L is the estimated likelihood, when it close to one the performance is nearly 100% and when it is closer to zero the performance is low. From the logarithmic identities we get that the as L increases, the $\ln(L)$ decreased and as a result AIC. So, to sum up, we can see that low p means low complexity and low AIC and

 low L means high performance and high AIC so a balance is needed.

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

As implied above, the two options that are correlated with the loss of balance is overfitting and underfitting. If a model is more complex, the performance will improve with the risk of overfitting. If the model is too simple, it will result the risk of underfitting and the performance will decrease.

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

AIC is most often used for model selection. It allows us to compare and choose the model that is the best fit for the data. Lower AIC scores are better, because it requires less information to predict with almost the exact same level of precision. Additionally, it penalizes models that use more parameters.

Therefore, if two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model.