

### HW3

1.

- a. The **K-means** clustering algorithm is **sensitive to outliers**, because a **mean** is easily influenced by **extreme values**. ... **Mean** is greatly influenced by the **outlier** and thus cannot represent the correct cluster center, while medoid is robust to the **outlier** and correctly represents the cluster center

- b. Will define:  $J = \sum_{i=1}^m (x_i - \mu)^2$

Will differentiate with respect to  $\mu$

And get:  $\frac{\partial J}{\partial \mu} = \sum_{i=1}^m -2(x_i - \mu)$

Will substitute  $\mu = \frac{\sum_{i=1}^m x_i}{m}$  and will get  $= \sum_{i=1}^m -2 \left( x_i - \frac{\sum_{i=1}^m x_i}{m} \right) =$

$(-2(mx_1 - x_1 - x_2 - \dots - x_m) + \dots + -2(mx_m - x_1 - x_2 - \dots - x_m))$  will divide by  $-2$  and then will get  $= (m(x_1 + \dots + x_m) - mx_1 - \dots - mx_m) = 0$

Therefore it's an extremum point will show it's a minimum rather than a maximum

Will differentiate again:  $\frac{\partial^2 J}{\partial \mu^2} = 2m > 0$  so the 2<sup>nd</sup> derivative is always positive and

thus the function is convex and has only minimum point.

- c. Will define  $J_2 = \sum_{i=1}^m |x_i - \mu|$   $f(x) = |x| \rightarrow f'(x) = \frac{x}{|x|}$  therefore

Will differentiate with respect to  $\mu$

And get:  $\frac{\partial J}{\partial \mu} = - \sum_{i=1}^m \frac{(x_i - \mu)}{|(x_i - \mu)|}$

Will substitute  $\mu = \text{median}(x)$  and will get  $\frac{\partial J}{\partial \mu} = - \sum_{i=1}^m \frac{(x_i - \text{median}(x))}{|(x_i - \text{median}(x))|} =$

Will open:  $-\frac{(x_1 - \text{median}(x))}{|(x_1 - \text{median}(x))|} - \dots - \frac{(x_m - \text{median}(x))}{|(x_m - \text{median}(x))|} =$   
 $-1 \left( \frac{(x_1 - \text{median}(x))}{|(x_1 - \text{median}(x))|} + \dots + \frac{(x_m - \text{median}(x))}{|(x_m - \text{median}(x))|} \right)$

We know that exactly half of the  $x$ 's bigger than the median and exactly half of them smaller therefore will get exactly half of the numerator are +1 and exactly half are -1

And all of the denominator are +1 therefore will get something like that:

$-1(1 + 1 + \dots + (-1) + (-1)) = -1(0) = 0$

Will differentiate again:  $\frac{\partial^2 J}{\partial \mu^2} = 0$  so the 2<sup>nd</sup> derivative is always positive (or 0) and thus

the function is convex and has only minimum point.

2. We have 2 SVM with linear kernel A and D. A as samples within the margins and thus have more soft margins and thus smaller penalty term C.

A-1

D-2

C and F are polynomial kernel, and it's clear F is with higher degree.

F-4

C-3

B and E are both rbf, and B boundary is tighter so its with bigger  $\gamma$ , which is correlated with tighter boundary.

B-6

E-5

3.

- a. The scientific term is generalization
- b.  $2p$  could imply about overfitting (lack of generalization) model if its 2 big.  
And  $2\ln(L)$  could imply about underfitting (lack of generalization) if its to small.  
Therefor a good AIC is if there is a balance between the 2 parameters.
- c. If the balance is violated there could be under/overfitting.
- d. AIC is a criterion that help you asses goodness of a model. Lower values of the index indicate the preferred model, that is, the one with the fewest parameters that still provides an adequate fit to the data.