1. Clustering

a. K-medoids is more robust to noise and outliers than k-means is because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. And this distance metric reduces noise and outliers.

b. If we look at $\sum_{i=1}^{m}(xi - \mu)^2$

to prove the minimum we will do a derivative on the equation:

$\sum_{i=1}^{m} -2xi + 2\mu$

Next we will do another derivative to see if it's min or max: We received 2 → it's min.

We'll compare the first derivative to 0.

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

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

In other words:

$\sqrt{(x-\mu)^2}$ Is the 1D Euclidean distance.

For m examples, the mean of the distances is: $\sum_{i=1}^{m} \frac{\sqrt{(xi-\mu)^2}}{m}$

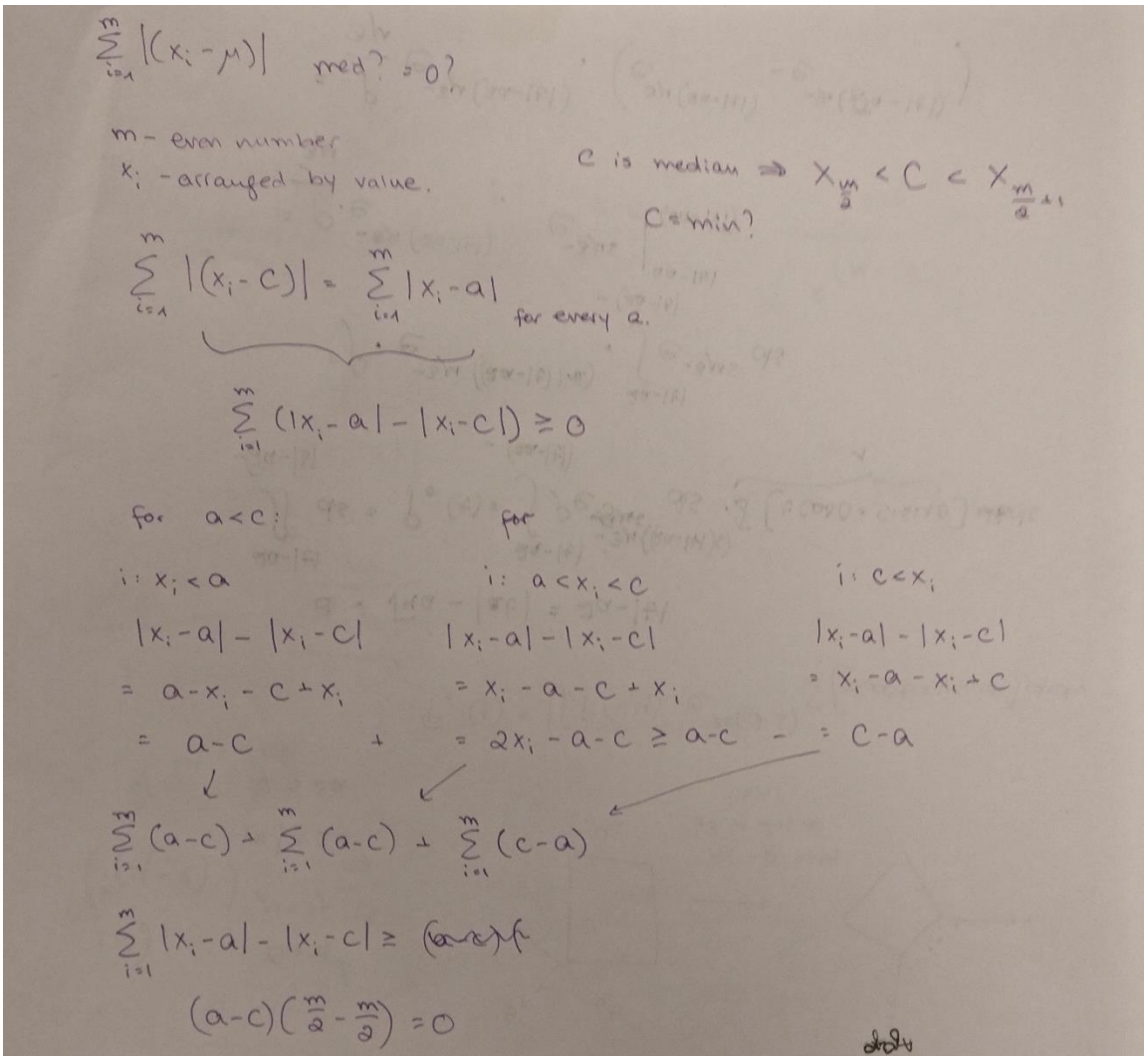
The μ that gives us the minimum of f(μ) will be the same μ that gives us the minimum of f(μ/m). Therefor we can remove m from the equation.

The min mean is $\sum_{i=1}^{m} \sqrt{(xi-\mu)^2}$

The μ that gives us the minimum of f($\sqrt{\mu}$) will be the same μ that gives us the minimum of f(μ). Therefor we can remove sqrt from the equation.

Meaning the smallest μ that minimizes the distance is also the mean $\sum_{i=1}^{m}(xi - \mu)^2$

c.



2. SVM

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

A and D have to be a linear kernel because the separator line is liner. C (the regulization) determines the margin of the classifier to the different points. If the C is higher the margin of the hyperplane will be smaller, and the hyperplane will classify the data correctly. If the C is lower the margin will be bigger and the hyperplane will mis-class more points. Since in A we can see some point that look mis-classified we can assume the C is lower – A represents 1. And D will have a high C, no points have been misclassified – representing 2.

In polynomial kernel the hyperplane is calculated as a polinom in the degree of d. That is why both C and F are plots with polynomial kernels.
In graph C the hyperplane looks represents the 2nd - 3. In polynomial, since we are powering the data by d (2), data that are in the same direction as the origin will get larger number – we are sqrt the value. The further they are in the direction the more positive they are. In F we see a hyperplane with a much higher degree so we can assume it represents 4.

RBF is based on the gaussian distribution and the sigmoid in the equation. Meaning we will have radial plots. B and E are radial. A higher value of γ causes the model overfit, so we assume B to have a higher γ. Therefor it represents 6. In E we have less fitting of the model so we can assume it has a smaller γ and therefor it represents 5.

3. Capability of generalization

a. The balance is a Machin learning aspect where the best-fit model is the one that explains the greatest amount of variation using the fewest possible independent variables - generalization. We are looking for simple – as little amount of parameters but not to simple that they explain the data.

We would like to achieve a balance where p is as small as possible, and L is big. Meaning we would like to use as little parameters as possible that still give us information – meaning have the largest likelihood for those parameters. This will determine the fitting of the model – that we didn't overfit.

We can rewrite the AIC model to explain the balance as:

$$AIC = 2ln\frac{e^k}{L}$$

Where e^k represents the p.

b. From the rewritten formula we can see that the AIC score is proportionate to the changes in both the p and the L. When we increase p or e^k the AIC score will increase. If the Likelihood L grows (performance), the AIC score will decrease.

c. over-fitting or under-fitting. If we have too many p we will have over-fitting – we will have a big AIC. Here we have an increase of complexity. -2ln(L), is negative, therefor if the likelihood is very high, this expression will be min. This "pulls" the function down, this will also cause over-fitting and lowering the complexity. If we have too little p we will have under-fitting, the model doesn't explain the data.

d. Lower AIC scores are better. If a model has a lower AIC score we can understand that the model is a better fit. 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. Meaning the lower score has less chance of over-fitting.