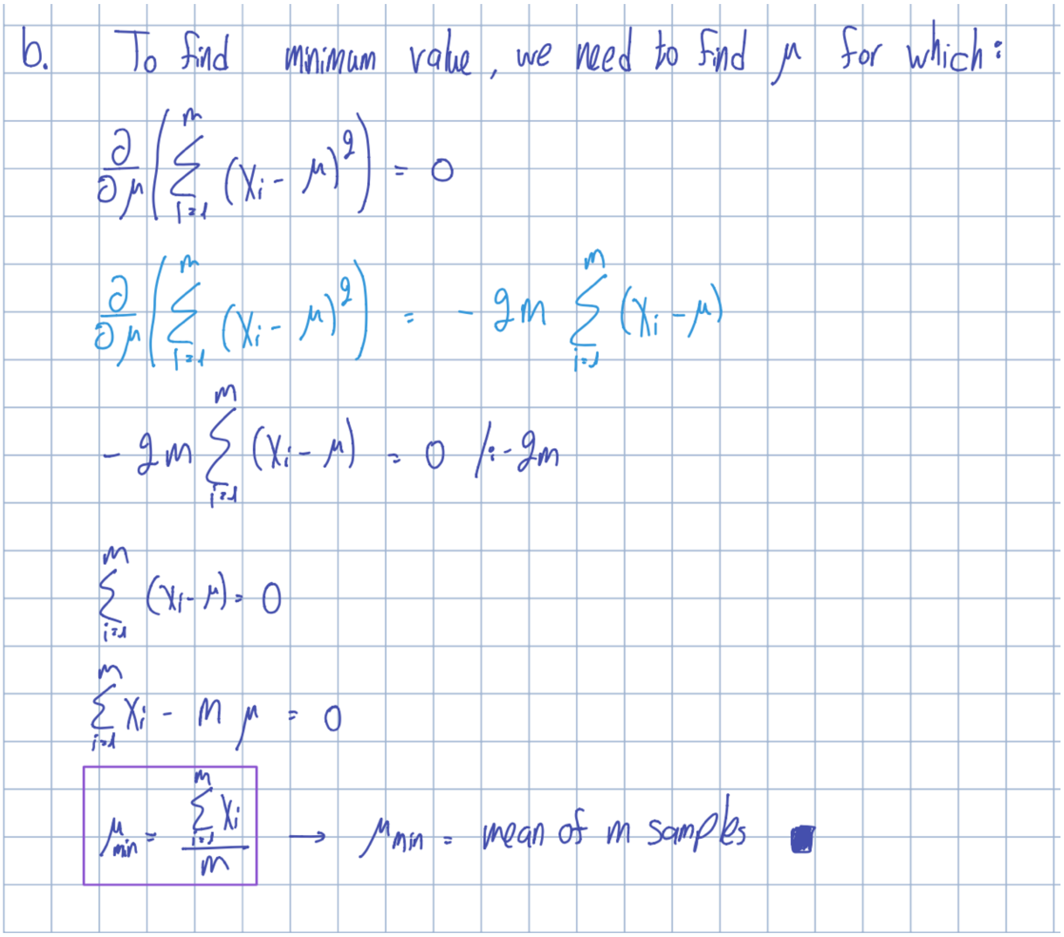
ML in healthcare 336546
HW3 Answers
Adi Alfassi 208929299

1. **Clustering**
   a. The goal of K-medoid algorithm is to minimizes a sum of pairwise dissimilarities, while K-means algorithm minimizes the sum of squared Euclidean distances. Therefore K-medoid algorithm would be more robust to noises and outliers than K-means algorithm.
   b.



b. To find minimum value, we need to find μ for which:

$$\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{m}(X_i-\mu)^2\right)=0$$

$$\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{m}(X_i-\mu)^2\right)=-2m\sum_{i=1}^{m}(X_i-\mu)$$

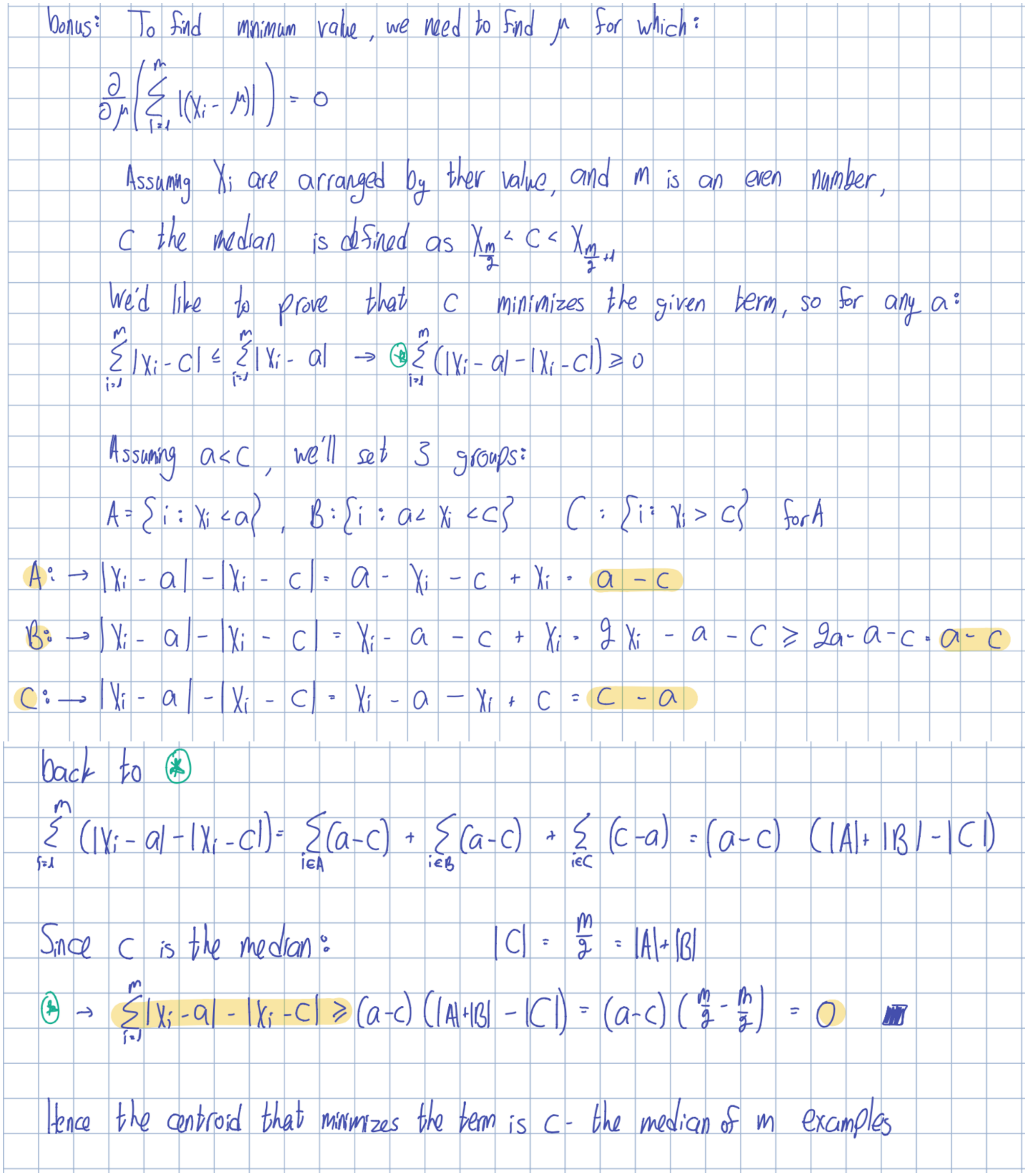
$$-2m\sum_{i=1}^{m}(X_i-\mu)=0 \quad /:-2m$$

$$\sum_{i=1}^{m}(X_i-\mu)=0$$

$$\sum_{i=1}^{m}X_i-m\mu=0$$

$$\boxed{\mu_{min}=\frac{\sum_{i=1}^{m}X_i}{m}} \rightarrow \mu_{min}=\text{mean of } m \text{ samples} \ ■$$

Bonus:



bonus: To find minimum value, we need to find μ for which:

$$\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{m} |(x_i - \mu)|\right) = 0$$

Assuming x_i are arranged by ther value, and m is an even number,

c the median is defined as $x_{\frac{m}{2}} < c < x_{\frac{m}{2}+1}$

We'd like to prove that c minimizes the given term, so for any a:

$$\sum_{i=1}^{m} |x_i - c| \le \sum_{i=1}^{m} |x_i - a| \rightarrow \circledast \sum_{i=1}^{m} (|x_i - a| - |x_i - c|) \geq 0$$

Assuming $a < c$, we'll set 3 groups:

$A = \{i : x_i < a\}$, $B : \{i : a < x_i < c\}$ $C : \{i : x_i > c\}$ for A

$A: \rightarrow |x_i - a| - |x_i - c| = a - x_i - c + x_i = a - c$

$B: \rightarrow |x_i - a| - |x_i - c| = x_i - a - c + x_i = 2x_i - a - c \geq 2a - a - c = a - c$

$C: \rightarrow |x_i - a| - |x_i - c| = x_i - a - x_i + c = c - a$

back to $\circledast$

$$\sum_{i=1}^{m} (|x_i - a| - |x_i - c|) = \sum_{i \in A}(a-c) + \sum_{i \in B}(a-c) + \sum_{i \in C}(c-a) = (a-c)\ (|A| + |B| - |C|)$$

Since c is the median: $|C| = \frac{m}{2} = |A| + |B|$

$$\circledast \rightarrow \sum_{i=1}^{m} |x_i - a| - |x_i - c| \geq (a-c)\,(|A| + |B| - |C|) = (a-c)\left(\frac{m}{2} - \frac{m}{2}\right) = 0 \quad \blacksquare$$

Hence the centroid that minimizes the term is c - the median of m examples

2. SVM

- A and D are classified with a linear kernel SVM because of the linear line of the classifier.

  For large values of C, the classification line would have a small margin range.
  We notice that the margin in D is smaller. Furthermore, a small value of C may allow misclassifications, which occurs in A – there are 2 purple dots inside the margin range. Therefore: D=2, A=1
- Since RBF stands for Radial Basis Function, I would expect that the classification would be with a radial shaped line- Images B and E are relevant.
  The gamma value represents the influence of a single example, when high gamma value means "close influence", and low value means "far influence". These are reflected in the size of the radial shapes caused by the classifier.
  Therefore: B=6, E=5.
- The classification shape in C looks the most similar to a 2nd order polynomial function (parabola). Therefore, the kernel matches this image is 2nd order polynomial kernel: C=3.
- The classification shape in F doesn't resemble as a known/radial function and looks very complex. A complex shape for a classifier would imply a complex function for a kernel (and a risk of over-fitting), for example a 10th order polynomial kernel. F=4.

3. Capability of generalization

a. This balance in the aspect of machine learning is the balance between model complexity and performance. The term for this balance is Generalization- As Einstein mentioned, we'd like a simple model but with enough complexity to make a good-performing model.

b. The 2p term when p is the total number of learned parameters represents the complexity of the model. Therefore, and from the formula we can say that the bigger 2p is- the bigger the complexity and the bigger the AIC.

   L is the estimated likelihood (varies from 0-1) and represents the performance of the model. Mathematically, 2ln(L) can vary from ($-\infty$) to 0. From the formula we can see that the bigger the likelihood, the bigger 2ln(L) and the lower AIC.
   That's how and why these terms represent the balance mentioned earlier.

c. In case of high complexity and high performance, there is a risk of over-fitting:
   If the model learns from a large number of parameters it may cause a high complexity. That model may have a very high performance with the specific examples given but wouldn't work that way for any other dataset.
   In case of low complexity there is a risk of under-fitting: If the model learns from a small number of parameters it may cause a low complexity, which leads to a model that with a low performance and won't be accurate enough.

d. As mentioned in section b, low AIC means low 2p value- represents low complexity, and high 2ln(L) (high L value)- represents high performance. Since this is the balance we'd like to achieve (Best and most simple model), we would like to minimize the value of AIC.