

"Clustering algorithms"

1st Homework

Exercise 1:

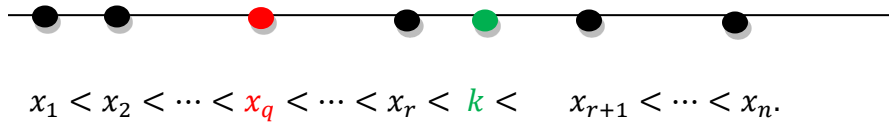
Let $x_1, x_2, \dots, x_n \in \mathbb{R}$, with $x_1 < x_2 < \dots < x_n$. Consider the quantity

$$A = \sum_{i=1}^n |x_i - \mu|.$$

Prove that A is minimized when μ is chosen as the median of x_1, x_2, \dots, x_n , i.e.

$$\mu = \text{med}(x_1, x_2, \dots, x_n).$$

Hints: (a) Let n be odd and $x_q \equiv \mu = \text{med}(x_1, x_2, \dots, x_n)$. Consider a number $k > x_q$ so that



Let

$$A_1 = \sum_{i=1}^n |x_i - x_q|,$$

and

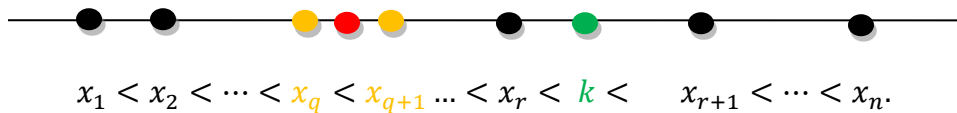
$$A_2 = \sum_{i=1}^n |x_i - k|.$$

The aim then is to prove that A_2 can be written as $A_2 = A_1 + \Delta$, where Δ is a positive quantity. To this end, express **each term** $|x_i - k|$ in A_2 in terms of $|x_i - x_q|$ and $|x_q - k|$, separating three cases: (a) $x_i < x_q$, (b) $x_q < x_i < k$, (c) $k < x_i$.

Finally, utilize the fact that $|x_q - k| > |x_q - x_i|$, for $i = q + 1, \dots, r$.

(b) The case where $k < x_q$ (n odd) is treated similarly.

(c) In the case where n is even, the median minimizes A but it is not the only minimizer.



More specifically, in terms of the above figure, **all the points** in the range $[x_q, x_{q+1}]$ are minimizers of A . Clearly, the **median** $\frac{x_q + x_{q+1}}{2}$ (denoted by the red dot) belongs to this range.

Exercise 2:

- (I) Derive the cost function optimization algorithm **hard k-medians** that belongs to the Generalized Hard Algorithmic Scheme (GHAS), where
- (a) the clusters are represented by point representatives
 - (b) the cost function that is minimized is

$$J(U, \theta) = \sum_{i=1}^N \sum_{j=1}^m u_{ij} \|x_i - \theta_j\|_1$$

where $x_i = [x_{i1}, \dots, x_{il}]^T \in R^l, i = 1, \dots, N$, are the N l -dimensional data vectors, $\theta_j = [\theta_{j1}, \dots, \theta_{jl}]^T, j = 1, \dots, m$, are the m l -dimensional point representatives of the clusters, and

$$\|x_i - \theta_j\|_1 = \sum_{r=1}^l |x_{ir} - \theta_{jr}|.$$

Give the algorithm in pseudocode (following the template given in the lecture slides).

- (II) Derive the cost function optimization algorithm **possibilistic k-medians** that belongs to the Generalized Possibilistic Algorithmic Scheme (GPAS), where
- (a) the clusters are represented by point representatives
 - (b) the cost function that is minimized is

$$J(U, \theta) = \sum_{i=1}^N \sum_{j=1}^m u_{ij} \|x_i - \theta_j\|_1 + \sum_{j=1}^m \eta_j \sum_{i=1}^N (u_{ij} \ln u_{ij} - u_{ij})$$

where $x_i = [x_{i1}, \dots, x_{il}]^T \in R^l, i = 1, \dots, N$, are the N l -dimensional data vectors, $\theta_j = [\theta_{j1}, \dots, \theta_{jl}]^T, j = 1, \dots, m$, are the m l -dimensional point representatives of the clusters, and

$$\|x_i - \theta_j\|_1 = \sum_{r=1}^l |x_{ir} - \theta_{jr}|.$$

Give the algorithm in pseudocode (following the template given in the lecture slides).

Exercise 3:

Let x_1, x_2, \dots, x_n be positive real numbers. Consider the quantity

$$A = \sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{\mu} \right)^2.$$

Prove that A is minimized when μ is chosen as the harmonic mean of x_1, x_2, \dots, x_n , i.e.

$$\mu = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

Hint: Take the derivative of A with respect to μ , set it equal to zero and solve.

Exercise 4:

- (I) Derive the cost function optimization algorithm **hard k-harmonic means** that belongs to the Generalized Hard Algorithmic Scheme (GHAS), where
- (a) the clusters are represented by point representatives
 - (b) the cost function that is minimized is

$$J(U, \theta) = \sum_{i=1}^N \sum_{j=1}^m u_{ij} d(\mathbf{x}_i, \theta_j)$$

where $\mathbf{x}_i = [x_{i1}, \dots, x_{il}]^T, i = 1, \dots, N$, are the N l -dimensional data vectors with positive coordinates, $\theta_j = [\theta_{j1}, \dots, \theta_{jl}]^T, j = 1, \dots, m$, are the m l -dimensional point representatives of the clusters (again with positive coordinates), and

$$d(\mathbf{x}_i, \theta_j) = \sum_{r=1}^l \left(\frac{1}{x_{ir}} - \frac{1}{\theta_{jr}} \right)^2.$$

Give the algorithm in pseudocode (following the template given in the lecture slides).

- (II) Derive the cost function optimization algorithm **possibilistic k-harmonic means** that belongs to the Generalized Possibilistic Algorithmic Scheme (GPAS), where
- (a) the clusters are represented by point representatives
 - (b) the cost function that is minimized is

$$J(U, \theta) = \sum_{i=1}^N \sum_{j=1}^m u_{ij} d(\mathbf{x}_i, \theta_j) + \sum_{j=1}^m \eta_j \sum_{i=1}^N (u_{ij} \ln u_{ij} - u_{ij})$$

where $\mathbf{x}_i = [x_{i1}, \dots, x_{il}]^T \in R^l, i = 1, \dots, N$, are the N l -dimensional data vectors with positive coordinates, $\theta_j = [\theta_{j1}, \dots, \theta_{jl}]^T, j = 1, \dots, m$, are the m l -dimensional point representatives of the clusters (again with positive coordinates), and $d(\mathbf{x}_i, \theta_j)$ is defined as in case (I).

Give the algorithm in pseudocode (following the template given in the lecture slides).

Exercise 5:

Consider the data set $Y = \{x_1, x_2, x_3, x_4, x_5\}$, where $x_1 = [4, 0]^T$, $x_2 = [7, 0]^T$, $x_3 = [4, 6]^T$, $x_4 = [4, 7]^T$, $x_5 = [1, 7]^T$.

- (a) Run the k-means, the k-medians and the k-harmonic means clustering algorithm, for two representatives, θ_1 and θ_2 , whose initial positions are $\theta_1(0) = [5, 6]^T$ and $\theta_2(0) = [4, 8]^T$, respectively. At each iteration report the (i) $U = [u_{ij}]$ matrix, (ii) θ_j 's and (iii) the formed clusters.
- (b) What would be the clustering result for the case where $\theta_2(0) = [4, 20]^T$?

How many clusters will be obtained if three representatives were employed?

Note: For this exercise use only paper, pencil and a pocket calculator (if necessary).

Exercise 6 (code):

- (a) Generate a data set consisting of 400 2-dimensional points. The 1st, 2nd, 3rd and 4th groups of 100 points stem from 2-dimensional normal distributions with means $m_1 = [10, 10]^T$, $m_2 = [20, 10]^T$, $m_3 = [10, 19]^T$, $m_4 = [19, 18]^T$, and covariance matrices $\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1.1 \end{bmatrix}$, $\Sigma_4 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.5 \end{bmatrix}$, respectively. Plot the data points using different colors for points stemming from different distributions.
- (b) Run the k-means, the k-medians and k-harmonic means¹ algorithms on the above data set and plot the data points that belong to different clusters with different colors.
- (c) Compare quantitatively the estimates of the representatives with the respective means of the distributions.

Hints: (a) To generate the data set use the following MATLAB code

```
randn('seed', 0)
m=[10 10; 20 10; 10 19; 19 18];
```

¹ You will need to write MATLAB code for the k-harmonic means algorithm, using as template the code for the k-means algorithm.

```

S(:, :, 1) = eye(2);
S(:, :, 2) = [1.0 .2; .2 1.5];
S(:, :, 3) = [1.0 .4; .4 1.1];
S(:, :, 4) = [.3 .2; .2 .5];
n_points = 100 * ones(1, 4);
X = [];
for i = 1:4
    X = [X; mvnrnd(m(i, :), S(:, :, i), n_points(i))];
end
X = X';

```

(b) To plot the data set use the following MATLAB code

```

figure(1), plot(X(1, :), X(2, :), '.b')
figure(1), axis equal

```

(c) To plot the points of different clusters with different colors, as well as to plot the cluster representatives, use the following MATLAB code

```

% Plot the clusters
figure(2), hold on
figure(2), plot(X(1, bel==1), X(2, bel==1), 'r.', ...
X(1, bel==2), X(2, bel==2), 'g*', X(1, bel==3), X(2, bel==3), 'bo', ...
X(1, bel==4), X(2, bel==4), 'cx', X(1, bel==5), X(2, bel==5), 'md', ...
X(1, bel==6), X(2, bel==6), 'yp', X(1, bel==7), X(2, bel==7), 'ks')
figure(2), plot(theta(1, :), theta(2, :), 'k+')
figure(2), axis equal

```

Exercise 7 (code):

- Implement in MATLAB the above derived algorithm.
- Generate a data set consisting of 500 2-dimensional points. The first 400 points are generated as in exercise 6. The 5th group of 100 points are noisy points that are spread among the points resulting from the above four distributions. Plot the data points using different colors for points stemming from different distributions.
- Run the k-means, the k-medians and the k-harmonic means algorithms on the above data set and plot the data points that belong to different clusters with different colors.
- Compare quantitatively the estimates of the representatives with the respective means of the distributions.

Hint: To generate the additional 100 noisy points add after the code generating the 400 points from the distributions, the following MATLAB code

```

% Generate the remaining 100 points
noise = rand(2, 100) * 16 + 6;
X = [X noise];

```

Exercise 8:

Discuss the results obtained in exercises 6 and 7.