IML Tutorial
Review HW5

Jakob Jakob [1]

[1]ETH Zurich

K-means

Given: n points $x_i \in \mathbb{R}^d, i \in 1, ..., n$
Goal: Find the k clusters $\mu = (\mu_1, ..., \mu_k)$
Minimize

$$L(\mu) = \sum_{i=1}^{n} \min_{j \in 1,...,k} \|x_i - \mu_j\|_2^2$$

Algorithm : Assignment and refitting step

$$z_i \in \arg\min_{j \in \{1,...,k\}} \|x_i - \mu_j\|_2^2 \qquad \mu_j = \frac{1}{|\{i : z_i = j\}|} \sum_{i:z_i=j} x_i$$

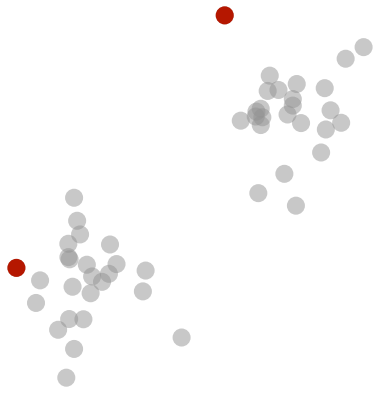
K-means — visualization

Given data

Example $k = 2, d = 2$

K-means — visualization

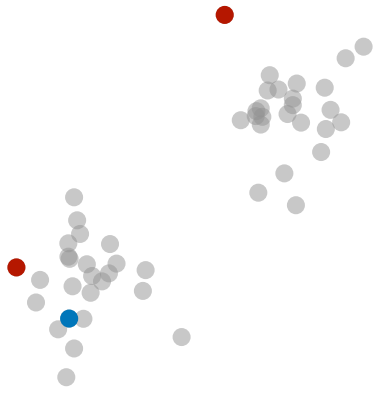
Initialization of μ_1 and μ_2



Example $k = 2, d = 2$

K-means — visualization

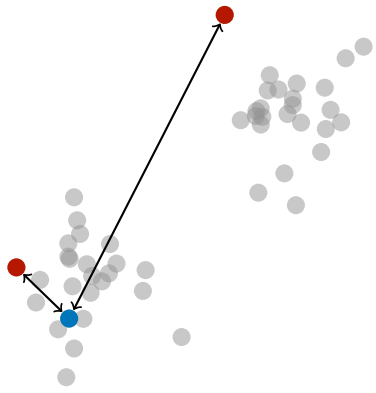
Assignment step



Example $k = 2, d = 2$

K-means — visualization

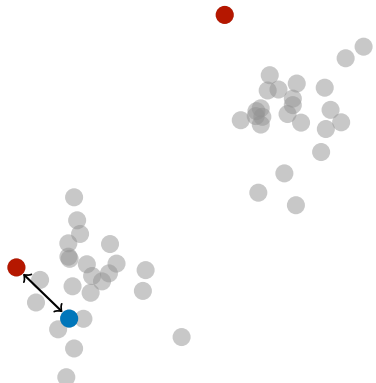
Assignment step



Example $k = 2, d = 2$

K-means — visualization

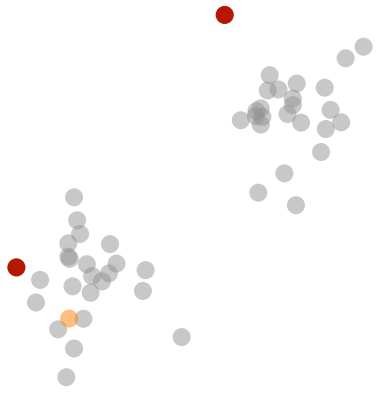
Assignment step



Example $k = 2, d = 2$

K-means — visualization

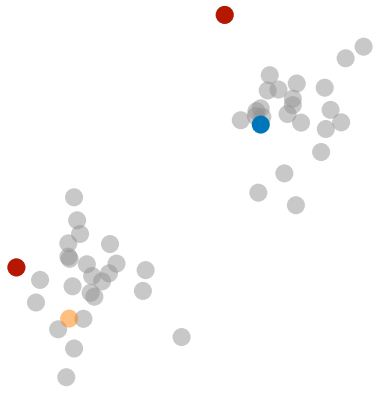
Assignment step



Example $k = 2, d = 2$

K-means — visualization

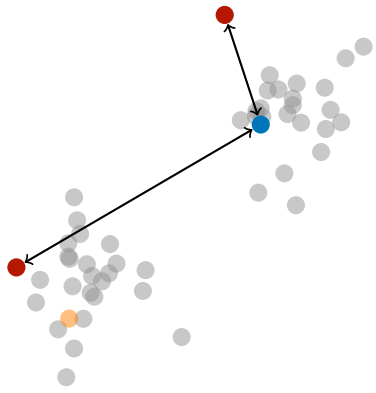
Assignment step



Example $k = 2, d = 2$

K-means — visualization

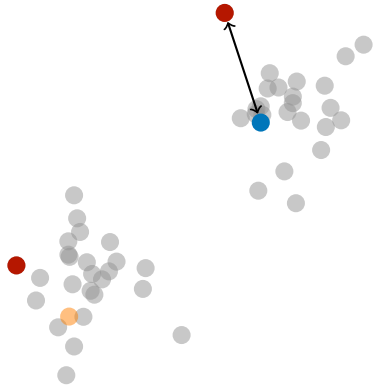
Assignment step



Example $k = 2, d = 2$

K-means — visualization

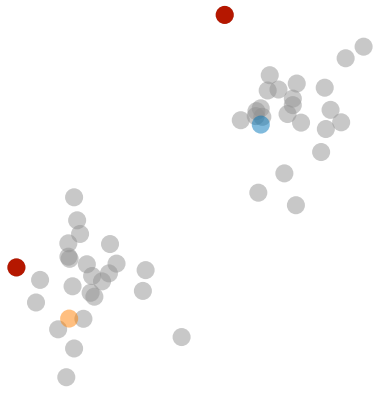
Assignment step



Example $k = 2, d = 2$

K-means — visualization

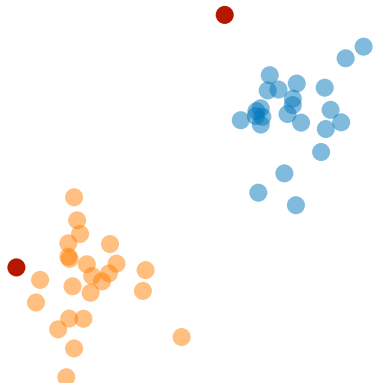
Assignment step



Example $k = 2, d = 2$

K-means — visualization

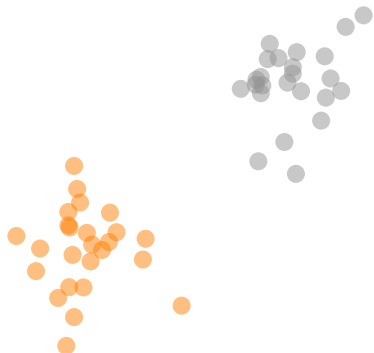
Assignment step



Example $k = 2, d = 2$

K-means — visualization

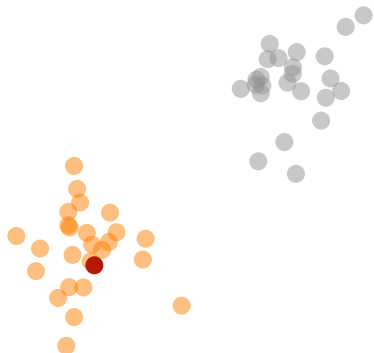
Refitting step



Example $k = 2, d = 2$

K-means — visualization

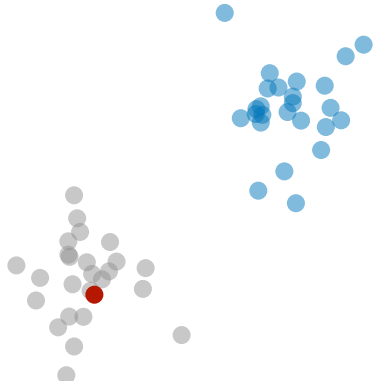
Refitting step



Example $k = 2, d = 2$

K-means — visualization

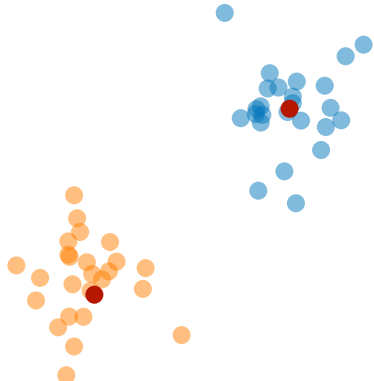
Refitting step



Example $k = 2, d = 2$

K-means — visualization

Refitting step



Example $k = 2, d = 2$

Review Q6

Explanation of Q6 (switch to solutions)

L2 Loss

$$L(\mu) = \sum_{i=1}^{n} \min_{j \in 1,\ldots,k} \|x_i - \mu_j\|_2^2$$

$$L(\mu) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2 \qquad z_i = \arg\min_{j \in 1,\ldots,k} \|x_i - \mu_j\|_2^2$$

$$\mu = \arg\min_{\mu} \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2$$

$$\mu_j = \arg\min_{\mu_j} \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2$$

L2 Loss

$$\mu_j = \arg\min_{\mu_j} \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2$$

$$\frac{\partial L}{\partial \mu_j} = \sum_{i:z_i=j} -2(x_i - \mu_j) = -2 \sum_{i:z_i=j} (x_i - \mu_j) = 0$$

$$= \sum_{i:z_i=j} (x_i - \mu_j) = \sum_{i:z_i=j} x_i - |\{i : z_i = j\}|\mu_j = 0$$

$$\implies \mu_j = \frac{1}{|\{i : z_i = j\}|} \sum_{i:z_i=j} x_i$$

L1 Loss

$$L(\mu) = \sum_{i=1}^{n} \min_{j \in 1,\ldots,k} \|x_i - \mu_j\|_1$$

$$L(\mu) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_1 \qquad z_i = \arg\min_{j \in 1,\ldots,k} \|x_i - \mu_j\|_1$$

$$\mu = \arg\min_{\mu} \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_1$$

$$\mu_j = \arg\min_{\mu_j} \sum_{i:z_i=j} \|x_i - \mu_j\|_1 = \arg\min_{\mu_j} \sum_{i:z_i=j} \sum_{q=1}^{d} |x_{i,q} - \mu_{j,q}|$$

L1 Loss

$$\mu_j = \arg\min_{\mu_j} \sum_{i:z_i=j} \sum_{q=1}^{d} |x_{i,q} - \mu_{j,q}|$$

$$\mu_{j,q} = \arg\min_{\mu_{j,q}} \sum_{i:z_i=j} |x_{i,q} - \mu_{j,q}|$$

$$L(\mu_{j,q}) = \sum_{i:z_i=j} |x_{i,q} - \mu_{j,q}|$$

$$= \sum_{i:z_i=j,\ x_{i,q} \le \mu_{j,q}} |x_{i,q} - \mu_{j,q}| + \sum_{i:z_i=j,\ x_{i,q} > \mu_{j,q}} |x_{i,q} - \mu_{j,q}|$$

L1 Loss

$$
\begin{aligned}
L(\mu_{j,q}) &= \sum_{i:z_i=j,\ x_{i,q}\le\mu_{j,q}} |x_{i,q}-\mu_{j,q}| + \sum_{i:z_i=j,\ x_{i,q}>\mu_{j,q}} |x_{i,q}-\mu_{j,q}| \\
&= \sum_{i:z_i=j,\ x_{i,q}\le\mu_{j,q}} (\mu_{j,q}-x_{i,q}) + \sum_{i:z_i=j,\ x_{i,q}>\mu_{j,q}} (x_{i,q}-\mu_{j,q}) \\
\frac{\partial L}{\partial \mu_j} &= |\{i : z_i = j,\ x_{i,q} \le \mu_{j,q}\}| - |\{i : z_i = j,\ x_{i,q} > \mu_{j,q}\}| = 0 \\
&\implies \mu_{j,q} = \text{median}\,(x_{i,q}, i : z_i = j)
\end{aligned}
$$

Mean vs Median — visualization

Example with an outlier $k = 1$, $d = 1$

Mean vs Median — visualization

Example with an outlier $k = 1$, $d = 1$, median in blue and mean in red

END OF REVIEW OF HW5