On the quality of k-means clustering based on grouped data

Meelis Käärik, Kalev Pärna

University of Tartu

OUTLINE

P – probability distribution on $\mathbb{R}$

X – random variable, $X \sim P$

DEF 1.1. We measure the quality of approximation of the distribution P by a set $A = \{a_1, \ldots, a_k\}$ by the following loss-function:

$$W(A, P) = \int \inf_{1 \le i \le k} (x - a_i)^2 P(dx).$$

P – probability distribution on $\mathbb{R}$

X – random variable, $X \sim P$

DEF 1.1. We measure the quality of approximation of the distribution P by a set $A = \{a_1, \ldots, a_k\}$ by the following loss-function:

$$W(A, P) = \int \inf_{1 \le i \le k} (x - a_i)^2 P(dx).$$

DEF 1.2. The loss-function for X can be written as

$$W(A, X) = E \inf_{1 \le i \le k} (X - a_i)^2.$$

P – probability distribution on $\mathbb{R}$

X – random variable, $X \sim P$

DEF 1.1. We measure the quality of approximation of the distribution P by a set $A = \{a_1, \ldots, a_k\}$ by the following loss-function:

$$W(A,P) = \int \inf_{1 \le i \le k} (x - a_i)^2 P(dx).$$

DEF 1.2. The loss-function for X can be written as

$$W(A,X) = E \inf_{1 \le i \le k} (X - a_i)^2.$$

DEF 1.3. Any A^* satisfying $W(A^* P) = \inf_{A \subset \mathbb{R}, |A|=k} W(A,P)$ is called k-mean of P (or X).

DEF 1.4. Let $S = \{S_1, \ldots, S_k\}$ be a partition of the support of P. S is Voronoi partition for A, if

$$S_i = \{x : |x - a_i| < |x - a_j|\} \bigcup \{x : |x - a_i| = |x - a_j|, i < j\}.$$

DEF 1.4. Let $S = \{S_1, \ldots, S_k\}$ be a partition of the support of P. S is Voronoi partition for A, if

$$S_i = \{x : |x - a_i| < |x - a_j|\} \bigcup \{x : |x - a_i| = |x - a_j|, i < j\}.$$

Example 1.1. Expectation EX is 1-mean of X. Calculate

$$\begin{aligned} W(a, X) &= E(X-a)^2 \\ &= E((X - EX) + (EX - a))^2 \\ &= E(X - EX)^2 + 2E((X - EX)(EX - a)) + E(EX - a)^2 \\ &= VarX + (EX - a)^2. \end{aligned}$$

2. Lloyd's algorithm

1. Choose inital points $C^0 = \{a_1^0, \ldots, a_k^0\}$
2. Given points C^m, find the Voronoi partition S^m
3. Calculate the centres of clusters of partition S^m, obtain points C^{m+1}
4. Calculate the loss $W(C^{m+1}, P)$
   Go to step 2 unless $W(C^m, P) - W(C^{m+1}, P)$ is small enough

3. Formulation of the problem

3. Formulation of the problem

Problem: high computational cost

3. Formulation of the problem

Problem: high computational cost

Idea: calculate the k-means for grouped data

3. Formulation of the problem

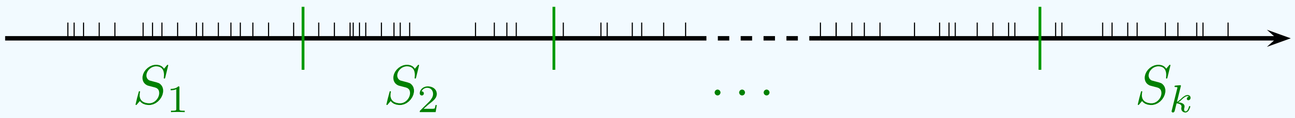
Problem: high computational cost

Idea: calculate the k-means for grouped data

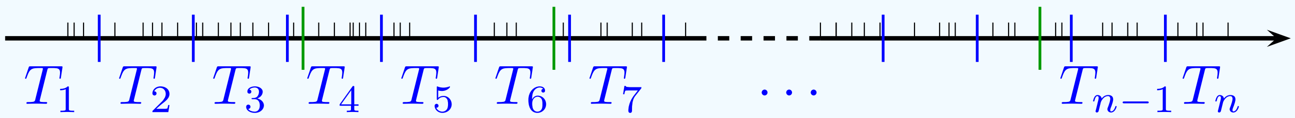
Question: how much information do we lose?

Graph 1. Original distribution P

Graph 2. Voronoi Partition $\mathcal{S}$ based on optimal k-mean for P (A^*)

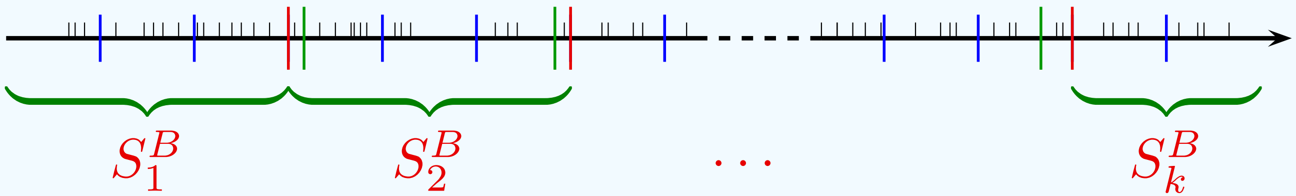


Graph 3. Grouping the original data (partition $\mathcal{T}$)



Graph 4. "Problematic" regions

Graph 5. "Closest" partion to $\mathcal{S}$ that follows $\mathcal{T}$: $\mathcal{S}^B$



4. Main question

P – original distribution

P^n – grouped data (based on partition $\mathcal{T}^n = \{T_1^n, \dots, T_n^n\}$)

A^* – k-means for P

A^{n*} – k-means for P^n

$\mathcal{S} = \{S_1, \dots, S_k\}$ – Voronoi partition corresponding to A^*

$\{T_1^*, \dots, T_{k-1}^*\} \subset \mathcal{T}^n$ – "problematic" regions:

$$T_i^* \cap S_i \neq \emptyset,\ T_i^* \cap S_{i+1} \neq \emptyset, \quad i = 1, \dots, k-1$$

$\{x_1^*, \dots, x_{k-1}^*\}$ – cluster means of $\{T_1^*, \dots, T_{k-1}^*\}$

4. Main question

P – original distribution

P^n – grouped data (based on partition $\mathcal{T}^n = \{T_1^n, \dots, T_n^n\}$)

A^* – k-means for P

A^{n*} – k-means for P^n

$\mathcal{S} = \{S_1, \dots, S_k\}$ – Voronoi partition corresponding to A^*

$\{T_1^*, \dots, T_{k-1}^*\} \subset \mathcal{T}^n$ – "problematic" regions:

$$T_i^* \cap S_i \neq \emptyset,\ T_i^* \cap S_{i+1} \neq \emptyset, \quad i = 1, \dots, k-1$$

$\{x_1^*, \dots, x_{k-1}^*\}$ – cluster means of $\{T_1^*, \dots, T_{k-1}^*\}$

How to estimate $W(A^{n*}, P) - W(A^*, P)$?

Some additional notations:

$\mathcal{S}^B = \{S_1^B, \ldots, S_k^B\}$ – partition defined by

$$S_i^B = \begin{cases} S_i \cup T_{i-1}^* \cup T_i^*, & \text{if} \quad x_{i-1}^* \in S_i, x_i^* \in S_i \\ S_i \cup T_{i-1}^* \backslash T_i^*, & \text{if} \quad x_{i-1}^* \in S_i, x_i^* \notin S_i \\ S_i \cup T_i^* \backslash T_{i-1}^*, & \text{if} \quad x_{i-1}^* \notin S_i, x_i^* \in S_i \\ S_i \backslash T_{i-1}^* \backslash T_i^*, & \text{if} \quad x_{i-1}^* \notin S_i, x_i^* \notin S_i. \end{cases}$$

$B = \{b_1, \ldots, b_k\}$ – means of clusters $\{S_1^B, \ldots, S_k^B\}$

DEF 4.1. For each $A = \{a_1, \ldots, a_k\}$ and partition $\mathcal{S} = \{S_1, \ldots, S_k\}$ denote $W(A, P|\mathcal{S}) := \sum_{i=1}^k \int_{S_i} \|x - a_i\|^2 dP$.

5. Results

Idea:

$$
\begin{aligned}
0 \le W(A^{n*},P) - W(A^{*},P) &= W(A^{n*},P) - W(A^{n*},P^{n}) \\
&+ W(A^{n*},P^{n}) - W(B,P^{n}|\mathcal{S}^{B}) \\
&+ W(B,P^{n}|\mathcal{S}^{B}) - W(B,P|\mathcal{S}^{B}) \\
&+ W(B,P|\mathcal{S}^{B}) - W(A^{*},P)
\end{aligned}
$$

5. Results

Idea:

$$\begin{aligned} 0 \le W(A^{n*},P) - W(A^*,P) &= W(A^{n*},P) - W(A^{n*},P^n) \\ &+ W(A^{n*},P^n) - W(B,P^n|\mathcal{S}^B) \\ &+ W(B,P^n|\mathcal{S}^B) - W(B,P|\mathcal{S}^B) \\ &+ W(B,P|\mathcal{S}^B) - W(A^*,P) \end{aligned}$$

General result:

$$W(A^{n*},P) - W(A^*,P) \le 2\sum_{i=1}^{k-1}(a_{i+1}-a_i)\Delta T_i^* \cdot P(T_i^*).$$

Example 5.1. Take $P(T_i^n) = \frac{1}{n}$, then

$$W(A^{n*}, P) - W(A^*, P) \leq \frac{2}{n}(a_k - a_1) \max_{1 \leq i < k} \Delta T_i^*.$$

Example 5.1. Take $P(T_i^n) = \frac{1}{n}$, then

$$W(A^{n*}, P) - W(A^*, P) \leq \frac{2}{n}(a_k - a_1) \max_{1 \leq i < k} \Delta T_i^*.$$

Example 5.2. Take $\Delta(T_i^n) \leq \Delta$, then

$$W(A^{n*}, P) - W(A^*, P) \leq 2\Delta(a_k - a_1) \max_{1 \leq i < k} P(T_i^*).$$

Example 5.1. Take $P(T_i^n) = \frac{1}{n}$, then

$$W(A^{n*}, P) - W(A^*, P) \leq \frac{2}{n}(a_k - a_1) \max_{1 \leq i < k} \Delta T_i^*.$$

Example 5.2. Take $\Delta(T_i^n) \leq \Delta$, then

$$W(A^{n*}, P) - W(A^*, P) \leq 2\Delta(a_k - a_1) \max_{1 \leq i < k} P(T_i^*).$$

Example 5.3. Take $P(T_i^n) \leq p$ ja $\Delta(T_i^n) \leq \Delta$, then

$$W(A^{n*}, P) - W(A^*, P) \leq 2p\Delta(a_k - a_1).$$

THANK YOU!