Review of Convexity-based clustering criteria: theory, algorithms, and applications in statistics

Hans-Hermann Bock

Hanchao Zhang

October 20, 2021

1 Traditional K-means

A classical approach of Kmeans that minimize sum of squares (SSQ)

$$
\begin{aligned}
g_n(\mathcal{C}) &= \frac{1}{n}\sum_{i=1}^{m}\sum_{k\in\mathcal{C}_i} ||x_k - \bar{x}_{\mathcal{C}_i}||^2 && (1)\\
&= \frac{1}{n}\sum_{i=1}^{m}\sum_{k\in\mathcal{C}_i} ||x_k||^2 + \frac{1}{n}\sum_{i=1}^{m}\sum_{k\in\mathcal{C}_i} ||\bar{x}_{\mathcal{C}_i}||^2 - \frac{2}{n}\sum_{i=1}^{m}\sum_{k\in\mathcal{C}_i} x_k\bar{x}_{\mathcal{C}_i} && (2)\\
&= \frac{1}{n}\sum_{k=1}^{n} ||x_k||^2 + \sum_{i=1}^{m}\frac{|\mathcal{C}_i|}{n}||\bar{x}_{\mathcal{C}_i}||^2 - \sum_{i=1}^{m}\frac{2|\mathcal{C}_i|}{n}||\bar{x}_{\mathcal{C}_i}||^2 \qquad \text{since } \bar{x}_{\mathcal{C}_i} = \frac{1}{|\mathcal{C}_i|}\sum_{k\in\mathcal{C}_i} x_k && (3)\\
&= \frac{1}{n}\sum_{k=1}^{n} ||x_k||^2 - \sum_{i=1}^{m}\frac{|\mathcal{C}_i|}{n}\cdot||\bar{x}_{\mathcal{C}_i}||^2 && (4)
\end{aligned}
$$

the minimization problem of equation 4 is equivalent to the maximization problem

$$
\max_{\mathcal{C}} \quad \tilde{h}_n(\mathcal{C}) = \sum_{i=m}\frac{|\mathcal{C}_i|}{n}\cdot||\bar{x}_{\mathcal{C}_i}||^2 \tag{5}
$$

We can converge the equation 5 to a two-parameters minimization problem.

$$
\min_{\mathcal{C},\mathcal{Z}} \quad g_n(\mathcal{C},\mathcal{Z}) = \frac{1}{n}\sum_{i=1}^{m}\sum_{k\in\mathcal{C}_i} ||x_k - z_i||^2 \tag{6}
$$

Where the one-parameter minimization problem becomes a two-parameters minimization with respect to all systems $\mathcal{Z} = (z_1, \ldots, z_m)$ and $\mathcal{C} = (C_1, \ldots C_m)$. Now, the algorithm can be done by recursively minimizing $g_n(\mathcal{C},\mathcal{Z})$.

2 Convexity Based Clustering

The equation 5 involves a convex function $\phi(\cdot) = ||\cdot||^2$. The main idea of this paper is to substitute the quadratic function $||\cdot||^2$ by any arbitrary convex function ϕ.

$$\max_{\mathcal{C}} \quad \tilde{h}_n(\mathcal{C}) = \sum_{i=m} \frac{|\mathcal{C}_i|}{n} \cdot \phi(\bar{x}_{\mathcal{C}_i}) \tag{7}$$

To generalize the equation 7, we define a the approach in a continuous format. We consider a random variable X in. $\mathbb{R}^p$ with a known probability distribution P, and look for m partitions $\mathcal{B} = (B_1, B_2, \ldots, B_m)$ of the entire space $\mathbb{R}^p$

$$H(\mathcal{B}) := \sum_{i=1}^{m} P(\mathcal{B}_i) \cdot \phi(E[X|X \in \mathcal{B}_i]) \tag{8}$$

Definition: Support Hyperplanes Defined by Derivatives

Define a support hyperplane $t(x; z, a)$ of the convex function $\phi : \mathbb{R}^p \to \mathbb{R}$ at the support point $z \in \mathbb{R}^p$ is any linear function $t(x; z, a) := a'(x - z) + \phi(z)$ of $x \in \mathbb{R}^p$ that fulfills

$$\phi(x) \geq t(x; z, a) \qquad \text{for all } x \in \mathbb{R}^p \tag{9}$$

Definition: Support Hyperplanes Defined by Conjugate Convex Function

For any convex function ϕ, we can define the conjugate convex function ϕ^* by

$$\phi^*(a) := \sup_{x \in \mathbb{R}^p} \left\{ a'x - \phi(x) \right\} \qquad \text{for } a \in \mathbb{R}^p \tag{10}$$

The domain of the conjugate convex function ϕ^* is the gradient of the function ϕ, and denoted by $K(\phi) := \left\{ a \in \mathbb{R}^p \middle| \phi^*(a) < \infty \right\}$. The hyperplane can be defined using the conjugate function as follow

$$t(x, z(a), a) = a'x - \phi^*(a) \qquad \text{for } x \in \mathbb{R}^p \tag{11}$$

Thus, a $H(\mathcal{B})$ problem can me reformatted to a continuous problem using the support hyperplanes. We denote the new problem as minimum-volume problem $G(\mathcal{B}, \mathcal{Z})$, where $\mathcal{Z} = (z_1, \ldots, z_m)$ is the system that generate the support hyperplanes

$$G(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^{m} \int_{\mathcal{B}_i} \left[\phi(x) - t(x; z_i) \right] dP(x) = E[\phi(X)] - E[p(X; \mathcal{B}, \mathcal{Z})] \tag{12}$$

$G(\mathcal{B}, \mathcal{Z})$ represent a weighted volumn (weight by the probability density function of X) between the surface $\phi(x)$ and $p(x; \mathcal{B}, \mathcal{Z})$.

Theorem: Choose $\mathcal{Z}$ as the Centroid System Minimize the $G(\mathcal{B}, \mathcal{Z})$

For any fixed m-partition $\mathcal{B} = (B_1, \ldots, B_m)$ of $\mathbb{R}^p$, let $z_i^* := E[X|X \in B_i]$ be the class centroid of B_i. Define by $\mathcal{Z}(\mathcal{B}) := \mathcal{Z}^* = (z_1^*, \ldots, z_m^*)$ the system centroids of partition $\mathcal{B}$. Then the system $\mathcal{Z}(\mathcal{B})$ minimize the equation 12.

$$G(\mathcal{B}, \mathcal{Z}) \geq G(\mathcal{B}, \mathcal{Z}(\mathcal{B})) \equiv G(\mathcal{B}, \mathcal{Z}^*) := G(\mathcal{B}) \tag{13}$$

Proof of Theorem

$$G(\mathcal{B}, \mathcal{Z}) - G(\mathcal{B}, \mathcal{Z}^*) = \sum_{i=1}^{m} \int_{\mathcal{B}_i} \Big[\big(\phi(x) - t(x; z_i)\big) - \big(\phi(x) - t(x; z_i^*)\big)\Big] dP(x) \tag{14}$$

$$= \sum_{i=1}^{m} \int_{\mathcal{B}_i} \Big[t(x; z_i^*) - t(x; z_i)\Big] dP(x) \tag{15}$$

$$= \sum_{i=1}^{m} \int_{\mathcal{B}_i} \Big[\big(a_i^{*\prime}(x - z_i^*) + \phi(z_i^*)\big) - \big(a_i' * (x - z_i) + \phi(z_i)\big)\Big] dP(x) \tag{16}$$

$$= \sum_{i=1}^{m} \Bigg[\underbrace{a_i^* \int_{\mathcal{B}_i} [x_i - z_i^*] dP(x)}_{=0 \text{ by def. of } z_i^*} + \int_{\mathcal{B}_i} [\phi(z_i^*) - a_i'(x - z_i) - \phi(z_i)] dP(x)\Bigg] \tag{17}$$

$$= \sum_{i=1}^{m} P(B_i) \big[\phi(z_i^*) - a_i'(x - z_i) - \phi(z_i)\big] \tag{18}$$

$$= \sum_{i=1}^{m} P(B_i) \underbrace{\big[\phi(z_i^*) - t(z_i^*; z_i\big]}_{\geq 0 \text{ by equation 9}} \geq 0 \tag{19}$$

Corollary: The One-parameter Maximization Problem Alternative

The one-parameter problem formed in equation 7 and the two-parameter in equation 12 are equivalent.

Proof of Corollary

$$G(\mathcal{B}) = G(\mathcal{B}, \mathcal{Z}(\mathcal{B})) = G(\mathcal{B}, \mathcal{Z}^*) = \sum_{i=1}^{m} \int_{\mathcal{B}_i} [\phi(x) - t(x; z_i^*)] dP(x) \tag{20}$$

$$= E[\phi(X)] - \sum_{i=1}^{m} \int_{\mathcal{B}_i} [a_i'(x - z_i^*) + \phi(z_i^*)] dP(x) \tag{21}$$

$$= E[\phi(X)] - \sum_{i=1}^{m} a_i' \underbrace{\int_{\mathcal{B}_i} [(x - z_i^*)] dP(x))}_{=0 \text{ by def. of } z_i^*} + \int_{\mathcal{B}_i} [\phi(z_i^*)] dP(x) \tag{22}$$

$$= E[\phi(X)] - \sum_{i=1}^{m} \phi(E[X|X \in \mathcal{B}_i]) \tag{23}$$

$$= E[\phi(X)] - H(\mathcal{B}) \tag{24}$$

Definition: Maximum Support Plane (MSP)

For a fixed but arbitrary system $\mathcal{Z} = (z_1, z_2, \dots, z_m)$, consider an arbitrary system of m support-planes $t(\cdot; z_1, a_1), \dots, t(\cdot; z_m, a_m)$. A partition $\mathcal{B} = (B_1, \dots, B_m)$ is called a maximum-support-plane partition genereated by $\mathcal{Z}$ and is denoted by $\mathcal{B}(\mathcal{Z})$ if

$$x \in B_i \qquad \Rightarrow \qquad t(x; z_i, a_i) = \max_{j=1,\dots,m} t(x; z_j, a_j) \tag{25}$$

Esesentially, $\mathcal{B}(\mathcal{Z})$ has the classes

$$B_i^* := \left\{x \in \mathbb{R}^p \middle| t(x; z_i, a_i) = \max_{j=1,\dots,m} t(x; z_j, a_j)\right\} \tag{26}$$

Theorem: B(Z) Maximize G(B, Z) Comparing to All Possible Partition B

Any MSP partition $\mathcal{B}^* = \mathcal{B}(\mathcal{Z})$ generated by $\mathcal{Z}$ minimizes the two-parameters criterion equation 12

$$G(\mathcal{B}, \mathcal{Z}) \geq G(\mathcal{B}(\mathcal{Z}), \mathcal{Z}) \equiv G(\mathcal{B}^*, \mathcal{Z}) := \Gamma(\mathcal{Z}) \tag{27}$$

Proof of Theorem

$$G(\mathcal{B}, \mathcal{Z}) = \sum_{i=1}^{m} \int_{B_i} \big[\phi(x) - t(x; z_i, a_i)\big] dP(x) \overset{\text{by def.}}{\geq} \sum_{i=1}^{m} \int_{B_i} \big[\phi(x) - \gamma(x; z_i, a_i)\big] dP(x) \tag{28}$$

$$= \int_{\mathbb{R}^p} \big[\phi(x) - \gamma(x; z_i, a_i)\big] dP(x) = \sum_{i=1}^{m} \int_{B_i^*} \big[\phi(x) - \gamma(x; z_i, a_i)\big] dP(x) \tag{29}$$

$$\overset{\text{by def.}}{=} \sum_{i=1}^{m} \int_{B_i^*} \big[\phi(x) - t(x; z_i, a_i)\big] dP(x) \tag{30}$$

$$= G(\mathcal{B}^*, \mathcal{Z}) \tag{31}$$

From all the previous inequalities, we shows that the algorithm contains following 5 equivalent optimization problems. The solution of any one of these problems yields the solution of the other four.

$$\min_{\mathcal{B}, \mathcal{Z}} G(\mathcal{B}, \mathcal{Z}) \tag{32}$$

$$\min_{\mathcal{B}} G(\mathcal{B}, \mathcal{Z}(\mathcal{B})) = \min_{\mathcal{B}} G(\mathcal{B}) \tag{33}$$

$$\min_{\mathcal{Z}} \Gamma(\mathcal{Z}) = \min_{\mathcal{Z}} G(\mathcal{B}(\mathcal{Z}), \mathcal{Z}) = \min_{\mathcal{Z}} E[\phi(X)] - E[\max_{j=1,\dots,m} t(X; z_j)] \tag{34}$$

$$\max_{\mathcal{Z}} E[\max_{j=1,\dots,m} t(X; z_j)] \tag{35}$$

$$\max_{\mathcal{B}} H(\mathcal{B}) = \max_{\mathcal{B}} \sum_{i=1}^{m} P(B_i) \cdot \phi(E[X | X \in B_i]) \tag{36}$$

3 The Maximum Support Plane Algorithm (MSP)

In the previous theorem, we have shown that we can partially minimize the two-parameter criterion $G(\mathcal{B}, \mathcal{Z})$ with respect to its first or to its second variable in an explicit way.

Maximum Support Plane Algorithm (MSP)

- $t = 0$ start with an inital system $\mathcal{Z}^{(0)} = (z_1^{(0)}, \ldots, z_m^{(0)})$, m distinct support points from $\mathbb{R}^p$
- $t \to t + 1$:
  - i) Determine a m-partition $\mathcal{B}^{(t+1)}$ that minimizes the $G(\mathcal{B}, \mathcal{Z}^{(t)})$. $\mathcal{B}^{(t+1)}$ can be chosen to be any MSP partition generated by the center system $\mathcal{Z}^{(t)}$
  - ii) Determine a system of support points $\mathcal{Z}^{(t+1)}$ which minimizes the criterion $G(\mathcal{B}^{(t)}, \mathcal{Z})$. This system is given by the clas centroids $z_i^{(t+1)} := E[X|X \in B_i^{(t+1)}]$
- stopping criterion:
  Iterate until 'convergence', e.g. until the center systems $\mathcal{Z}^{(t)}$ attain an approximately stationary state

4 Generalized Convexity Based Criterion

The generalized idea based on MSP is presented with change of the $\lambda : \mathcal{X} \to \mathcal{R}^q$ function. We assume that the random variable X has a domain $\mathcal{X}$, and look for m-partition of $\mathcal{B}$ of that domain $\mathcal{X}$ that maximizes the generalized convexity based clustering criterion.

$$\max_{\mathcal{B}} H_\lambda(\mathcal{B}) := \max_{\mathcal{B}} \sum_{i=1}^{n} \int_{B_i} P(B_i) \cdot \phi(E[\lambda(X)|X \in B_i]) \tag{37}$$

Similarly to the previous way of defining a two-parameter problem, now, we represent $t(\lambda; a(w), w) = \phi(w) + a(w)'(\lambda - w)$ for the tangent hyperplane of ϕ in the support point $w \in \mathbb{R}^q$

$$H_\lambda(\mathcal{B}, \mathcal{W}) = \sum_{i=1}^{m} \int_{B_i} t(\lambda(x); a(w_i), w_i) dP(x) \quad \to \quad \max_{\mathcal{B}, \mathcal{W}} \tag{38}$$

where the maximization is over all m-partition $\mathcal{B} = (B_1, \ldots, B_m)$ of $\mathcal{X}$ denote and all systems $\mathcal{W} = (w_1, \ldots, w_m)$ of support points $w_i \in \mathbb{R}^q$.

Theorem: W(B) is the system that maximize G(B,W)

For a fixed partition $\mathcal{B} = (B_1, \ldots, B_m)$ of $\mathcal{X}$ denote by $w_i^* := E[\lambda(x)|X \in \mathcal{B}_i]$ the class-specific expectation of the random vector $Y := \lambda(X)$. Then $\mathcal{W}(\mathcal{B}) := \mathcal{W}^* = (w_1^*, \ldots, w_m^*)$ is an optimum system of support points for G_λ

$$G_\lambda(\mathcal{B}, \mathcal{W}) \leq G_\lambda(\mathcal{B}, \mathcal{W}(\mathcal{B})) = G_\lambda(\mathcal{B}, \mathcal{W}^*) = H_\lambda(\mathcal{B}) \qquad \text{for all } \mathcal{B}, \mathcal{W} \tag{39}$$

Proof of Theorem

For any support system $\mathcal{W}$, we have

$$G_\lambda(\mathcal{B},\mathcal{W}) = \sum_{i=1}^m \int_{B_i} t(\lambda(x); a(w_i), w_i) dP(x) \tag{40}$$
$$= \sum_{i=1}^m \int_{B_i} \{\phi(w_i) + a(w_i)'(\lambda(x) - w_i)\} dP(x) \tag{41}$$
$$= \sum_{i=1}^m P(B_i)\{\phi(w_i) + a(w_i)'(E[\lambda(X)|X \in B_i] - w_i)\} \tag{42}$$
$$= \sum_{i=1}^m P(B_i) t(w_i^*; a(w_i), w_i) \overset{\text{by def.}}{\leq} \sum_{i=1}^m P(B_i)\phi(w_i^*) = H_\lambda(\mathcal{B}) \tag{43}$$
$$= \sum_{i=1}^m P(B_i) t(w_i^*; a(w_i^*), w_i^*) \tag{44}$$
$$= \sum_{i=1}^m P(B_i) t(E[\lambda(X)|X \in B_i]; a(w_i), w_i) \tag{45}$$
$$= \sum_{i=1}^m \int_{B_i} \{\phi(w_i^*) + a(w_i^*)'(\lambda(x) - w_i^*)\} dP(x) \tag{46}$$
$$= G_\lambda(\mathcal{B},\mathcal{W}^*) \tag{47}$$

Theorem: B(W) is the partition that maximize G(B,W)

Consider an arbitrary system $\mathcal{W} = (w_1, \ldots, w_m)$ of m distinct support vectors from $\mathbb{R}^q$. Denote by $\mathcal{D}(\mathcal{W}) = \mathcal{D}^*$ any MSP partition of $\mathbb{R}^q$ corresponding to $\mathcal{W}$ and define the partition $\mathcal{B}_\lambda(\mathcal{W}) = \mathcal{B}^* := (B_1^*, \ldots, B_m^*)$ of $\mathcal{X}$ by the classes $B_i^* := \lambda^{-1}(D_i^*) := \{x \in \mathcal{X} | \lambda(x) \in D_i^*\}$. Then $\mathcal{B}^*$ maximize the two-parameter criterion with respect to all m-paritions $\mathcal{B} = (B_1, \ldots, B_m)$ of $\mathcal{X}$:

$$G_\lambda(\mathcal{B},\mathcal{W}) \leq G_\lambda(\mathcal{B}_\lambda(\mathcal{W}),\mathcal{W}) =: \Gamma_\lambda(\mathcal{W}) \quad \text{for all } \mathcal{B},\mathcal{W} \tag{48}$$

Proof of Theorem

For a fixed $\mathcal{W}$, we have for all m-partition $\mathcal{B}$ of $\mathcal{X}$:

$$G_\lambda(\mathcal{B},\mathcal{W}) = \sum_{i=1}^m \int_{B_i} t\big(\lambda(x); a(w_i), w_i\big) dP(x) \tag{49}$$
$$\leq \sum_{i=1}^m \int_{B_i} \max_{j=1,\ldots,m} t\big(\lambda(x); a(w_i), w_i\big) dP(x) \tag{50}$$
$$= \int_{\mathbb{R}^p} \max_{j=1,\ldots,m} t\big(\lambda(x); a(w_j), w_j\big) dP(x) \tag{51}$$
$$\overset{\text{by def.}}{=} \int_{B_i^*} t\big(\lambda(x); a(w_j), w_j\big) \tag{52}$$
$$= G_\lambda(\mathcal{B}_\lambda(\mathcal{W}),\mathcal{W}) \tag{53}$$

The above two theorems is the analogue of the previous theorem

The Generalized Maximum Support Plane Algorithm (MSP)

- Given $\mathcal{B}^{(t)}$ calculate support points by $w_i^{(t)} = E[\lambda(X)|X \in B_i^{(t)}] \in \mathbb{R}^q$ for $i = 1, \ldots, m$ yielding $\mathcal{W}^{(t)}$

- Determine a MSP partition $\mathcal{D}(\mathcal{W}^{(t)})$ of $\mathbb{R}^q$ with classes

$$D_i^{(t)} = \{\lambda \in \mathbb{R}^q | t(\lambda; a(w_i^{(t)}), w_i^{(t)}) = \max_j t(\lambda; a(w_i^{(t)}), w_i^{(t)})\}$$

- Build the m-partition $\mathcal{B}^{(t_1)}$ of $\mathcal{X}$ with classes $B_i^{(t+1)} := \lambda^{-1}(D_i^{(t)})$