Penalized K-Means Algorithms for Finding the Number of Clusters

Behzad Kamgar-Parsi
Office of Naval Research
Arlington, VA, USA

Behrooz Kamgar-Parsi
Naval Research Laboratory
Washington, DC, USA

Penalized k-means

- K-means error $E_k = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \| \mathbf{x}_i - \mathbf{c}_j \|^2, \quad \mathbf{c}_j = \frac{1}{N_j} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i$

  # of clusters   Data points   Cluster centroids

- Increasing k reduces error monotonically, hence k-means algorithm cannot find the correct number of clusters.

- K-means error with additive penalty $E_k^{(a)} = E_k + \lambda\, k$

- **Problem:** no principled method to determine a good value for λ

K-means clusters & Ideal clusters

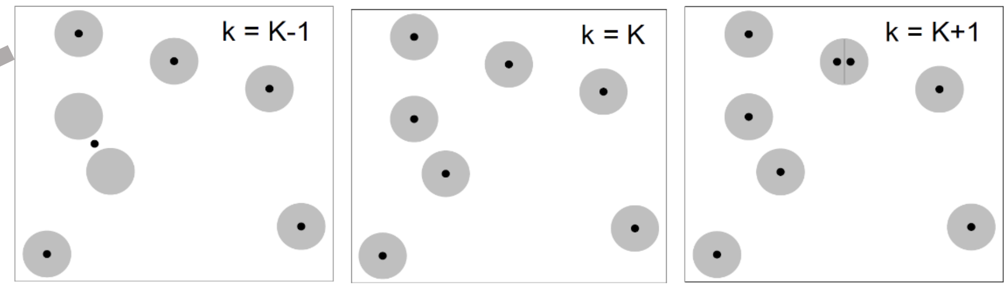
- K-means algorithm cannot guarantee optimal solution

- **Consider ideal clusters [**has provably optimal clustering algorithms**]**

- Slight differences in the underlying assumptions
  - K-means clusters
    - Spherically symmetric (with normal distributions)
    - Same size
    - Sufficiently separated
    - No background noise
  - Ideal clusters
    - Spheres
    - Same size
    - Sufficiently separated
    - No background noise
    - Full (for computational convenience only: replace sums with integrals)

Optimal clusters and clustering errors

- *K* correct number of clusters
- Optimal ideal clusters are
  - *k = K-1*: *K-2* spheres + 1 dumbbell
  - *k = K*: *K* spheres
  - *k = K+1*: *K-1* spheres + 2 half-spheres



- These yield clustering errors:

$$
\begin{aligned}
E_{K-1} &= (K-2)E_s + E_d = K\,V R^2 \alpha + 2V L^2 \\
E_K &= K E_s = K\,V R^2 \alpha \\
E_{K+1} &= (K-1)E_s + 2E_h = (K-1)V R^2 \alpha + 2V R^2 \beta
\end{aligned}
$$

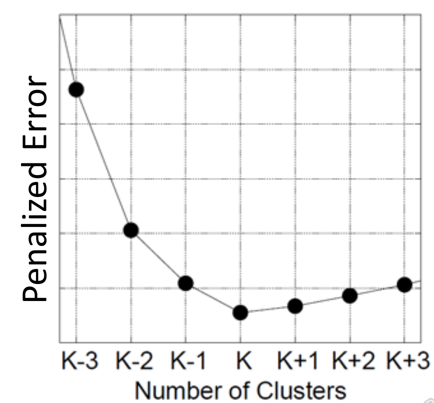
- Errors for single clusters: sphere, half-sphere, dumbbell

$$
\begin{aligned}
E_s &= V R^2 \alpha \\
E_h &= V R^2 \beta \\
E_d &= 2E_s + 2V L^2 = 2(V R^2 \alpha + V L^2)
\end{aligned}
$$

Volume of *d*-dim sphere $V = \dfrac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d+2}{2})} R^d$

and $\alpha = \dfrac{d}{d+2}, \quad \beta = \dfrac{1}{2}(\alpha - \gamma^2), \quad \gamma = \dfrac{\Gamma(\frac{d+2}{2})}{\sqrt{\pi}\,\Gamma(\frac{d+3}{2})}$

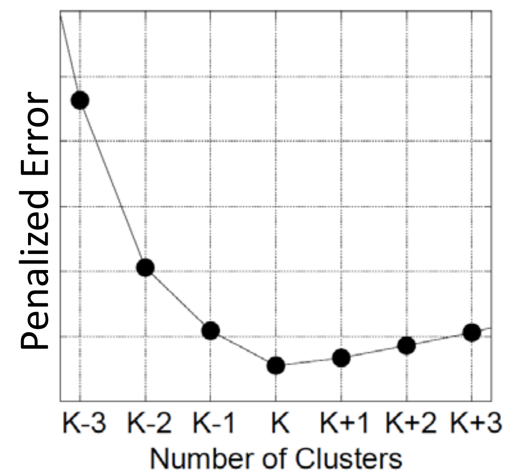
- Then impose conditions for *k = K* to be a minimum.



Bounds for λ

- Penalized error $E_k^{(a)}$ to have a minimum at K must have:

$$\Delta^{(a)}_{K-1,K} = E_{K-1} - E_K = 2VL^2 - \lambda > 0 \quad \text{for } d \geq 1 \text{ and } K > 1,$$
$$\Delta^{(a)}_{K,K+1} = E_K - E_{K+1} = VR^2(\alpha - 2\beta) - \lambda < 0 \quad \text{for } d \geq 1 \text{ and } K \geq 1.$$



- Or $\boxed{\frac{N\rho^2}{K} < \lambda < \frac{2NL^2}{K}}$

- For tests, choose mid-point of the range $\boxed{\lambda = \frac{N(\rho^2 + 2L^2)}{2K} \approx \frac{NL^2}{K}}$
  - N number of data points
  - L smallest inter-centroid distance
  - ρ distance of half-sphere centroid to its equator

Multiplicative penalty

- Additive penalty often gives multiple solutions: ambiguous
- Use multiplicative penalty, $E_k^{(m)} = \lambda\, E_k$, to confirm the correct solution

$$\begin{aligned}
\Delta^{(m)}_{K-1,K} &= (K-1)E_{K-1} - KE_K \\
&= 2(K-1)VL^2 - KVR^2\alpha > 0 \\
&\quad \text{for } d \geq 1 \text{ and } K \geq 2, \\
\Delta^{(m)}_{K,K+1} &= KE_K - (K+1)E_{K+1} \\
&= VR^2[\alpha - 2(K+1)\beta] < 0 \\
&\quad \text{for } d \geq 2 \text{ and } K \geq 2.
\end{aligned}$$

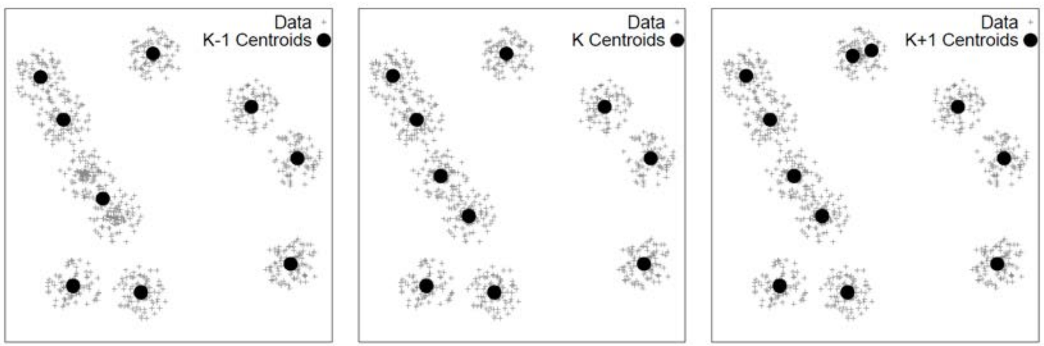
- Both inequalities are automatically satisfied

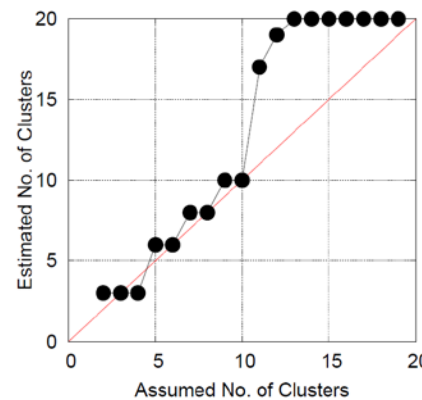
Experiments

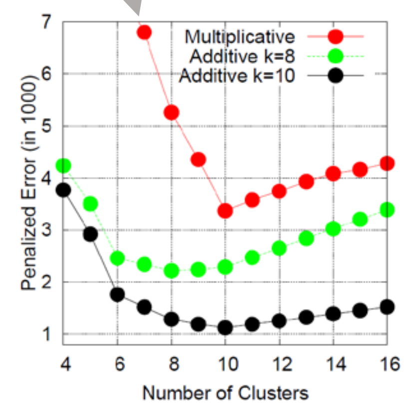
- Additive penalty
  - Assume a (small) k, and run k-means
  - If Estimated k is equal to the Assumed k => a candidate solution
  - Then increment k, and repeat

- Multiplicative penalty
  - Assume a (small) k, run k-menas.
  - Increment k, and repeat
  - Minimum of $E_k^{(m)}$ vs k is the solution

An example with $K = 10$ clusters







- Additive solutions: $k = 3, 6, 8, 10$
- Multiplicative solution: $k = 10$
- Combined solution: $k = 10$

For derivations and more tests, please see the paper.

Thank you!