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
  \[ E_k = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - c_j \|^2, \quad c_j = \frac{1}{N_j} \sum_{x_i \in C_j} x_i \]

  - 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 \( \lambda \)
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

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

- $E_s = V R^2 \alpha$
- $E_h = V R^2 \beta$
- $E_d = 2E_s + 2VL^2 = 2(VR^2\alpha + VL^2)$

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

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

These yield clustering errors:

- $E_{K-1} = (K-2)E_s + E_d = KVR^2\alpha + 2VL^2$
- $E_K = KE_s = KVR^2\alpha$
- $E_{K+1} = (K-1)E_s + 2E_h = (K-1)VR^2\alpha + 2V\beta$

Then impose conditions for $k = K$ to be a minimum.
Bounds for $\lambda$

- Penalized error $E_k^{(a)}$ to have a minimum at $K$ must have:
  \[
  \Delta_{K-1,K}^{(a)} = E_{K-1} - E_K = 2VL^2 - \lambda > 0 \\
  \text{for } d \geq 1 \text{ and } K > 1,
  \]
  \[
  \Delta_{K,K+1}^{(a)} = E_K - E_{K+1} = VR^2(\alpha - 2\beta) - \lambda < 0 \\
  \text{for } d \geq 1 \text{ and } K \geq 1.
  \]

- Or
  \[
  \frac{N\rho^2}{K} < \lambda < \frac{2NL^2}{K}
  \]

- For tests, choose mid-point of the range
  \[
  \lambda = \frac{N(\rho^2 + 2L^2)}{2K} \approx \frac{NL^2}{K}
  \]

- $N$ number of data points
- $L$ smallest inter-centroid distance
- $\rho$ 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

  \[
  \Delta_{K-1,K}^{(m)} = (K - 1)E_{K-1} - KE_K \\
  = 2(K - 1)VL^2 - KVR^2\alpha > 0 \\
  \text{for } d \geq 1 \text{ and } K \geq 2,
  \]

  \[
  \Delta_{K,K+1}^{(m)} = KE_K - (K + 1)E_{K+1} \\
  = VR^2[\alpha - 2(K + 1)\beta] < 0 \\
  \text{for } d \geq 2 \text{ and } K \geq 2.
  \]

• 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-means.
  - 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!