ADVANCED MACHINE LEARNING

Clustering Evaluation Metrics

Lecture: Prof. Aude Billard (aude.billard@epfl.ch)

Teaching Assistants:
Guillaume de Chambrier, Nadia Figueroa, Denys Lamotte
Evaluation of Clustering Methods

*Clustering methods rely on hyper parameters*

- Number of clusters
- Distance metric
- Type of Kernel
- ...

→ Need to determine the goodness of these choices

*Clustering is unsupervised classification*

→ Do not know the real number of clusters and the data labels.
→ Difficult to evaluate these choices without *ground truth.*
Evaluation of Clustering Methods

Two types of measures: Internal versus external measures

External measures rely on ground truth (class labels):
- Given a (sub)-set of known class labels compute similarity of clusters to class labels.
- In real-world data, it’s hard/infeasible to gather ground truth.

Internal measures rely on measures of similarity:
- (low) intra-cluster distance versus (high) inter-cluster distances

Residual Sum of Square
Sum of squared distances of each point $x$ to its assigned centroid $\mu^k$:

$$\text{RSS} = \sum_{k=1}^{K} \sum_{x \in C_k} |x - \mu^k|^2$$

→ Internal measures are problematic as the metric of similarity is often already optimized by clustering algorithm (as in K-means).
RSS for K-Means

Goal of K-means is to find cluster centers \( \mu^k \) which minimize distortion.

\[
\text{RSS} = \sum_{k=1}^{K} \sum_{x \in C_k} |x - \mu^k|^2
\]

Measure of Distortion

By \( \uparrow K \) we \( \downarrow \text{RSS} \), what is the optimal \( K \) such that \( \text{RSS} \rightarrow 0 \)?

\( \Rightarrow \) \( \text{RSS} = 0 \) when \( K = M \). One has as many clusters as datapoints!

\( M: 100 \) datapoints
\( D: 2 \) dimensions

However, it can still be used to determine an ‘optimal’ \( k \) by monitoring the slope of the decrease of the measure as \( k \) increases.
K-means Clustering: Examples

**Procedure:** Run K-means – increase monotonically number of clusters – run K-means with several initialization and take best run;

use RSS measure to measure improvement in clustering $\rightarrow$ determine a plateau

$M$: 100 datapoints
$D$: 2 dimensions

$k$: 4 clusters
**K-means Clustering: Examples**

The ‘elbow’ or ‘plateau’ method for choosing the optimal $k$ from the RSS curve can be unreliable for certain datasets:

$M$: 100 datapoints  
$D$: 3 dimensions

Which one is the ‘optimal’ $k$?

$k$: 3  
$k$: 11

We don’t know! We need an additional penalty or criterion!
Evaluation of Clustering Methods

K-Means, soft-K-Means and GMM have several hyper-parameters:
(Fixed number of clusters, beta, number of Gaussian functions)

→ Measure to determine how well the choice of hyperparameters fit the dataset (maximum-likelihood measure)

\[ M : \text{number of datapoints}; \ B : \text{number of free parameters} \]

- Aikaike Information Criterion: \( \text{AIC} = -2 \ln L + 2B \)
- Bayesian Information Criterion: \( \text{BIC} = -2 \ln L + B \ln (M) \)

L: maximum likelihood of the model given B parameters

Choosing AIC versus BIC depends on the application:

→ Is the purpose of the analysis to make predictions, or to decide which model best represents reality?
AIC may have better predictive ability than BIC, but BIC finds a computationally more efficient solution.
K-means Clustering: Examples

**Procedure:** Run K-means – increase monotonically number of clusters – run K-means with several initialization and take best run;

- use AIC/BIC curves to find the optimal \( k \), which is \( \min(AIC) \) or \( \min(BIC) \)

\( M: 100 \) datapoints
\( D: 3 \) dimensions

\( k: 2 \) clusters
AIC for K-Means

For the particular case of K-means, we do not have a maximum likelihood estimate of the model:

$$AIC = -2 \ln(L) + 2B$$

$L$: likelihood of model
$B$: number of free parameters

However, we can formulate a metric based on the RSS that penalizes for model complexity (# K-clusters), conceptually following AIC:

$$AIC_{RSS} = RSS + \lambda B$$

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?
AIC for K-Means

For such an ideal dataset, we can see that any value of $\lambda$ yields the optimal cluster number, yet which $\lambda$ yields the ‘best’ AIC curve?

$$AIC_{RSS} = RSS + \lambda B$$

Number of free parameters $B = (K \times D)$

- $K$: # clusters
- $D$: # dimensions

$M$: 100 datapoints

$D$: 2 dimensions

$k$: 4 clusters
AIC for K-Means

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda B$$

- Idea: Find $\lambda$ which penalizes choosing $k=M$ with at least the same magnitude as $RSS(k=1)$
  
  $$AIC_{RSS}(k = M) \geq RSS(k = 1)$$

  $$RSS(k = M) + \lambda \cdot D \cdot M \geq RSS(k = 1)$$

  $$\lambda \geq RSS(k = 1)/D \cdot M$$

  $$\lambda \geq 5$$

Number of free parameters $B=(K\cdot D)$

$K$: # clusters

$D$: # dimensions

$M$: 100 datapoints

$D$: 2 dimensions

Flat curves: not penalizing for excessive parameters

$AIC_{RSS}(k = M) \geq RSS(k = 1)$ can be seen as a lower-bound, can we compute an upper-bound?
AIC for K-Means

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda B$$

Number of free parameters $B=(K*D)$
$K$: # clusters
$D$: # dimensions

$M$: 100 datapoints
$D$: 2 dimensions
$k$: 6 clusters

$\lambda \geq RSS(k = 1)/D \cdot M$
$\lambda \geq 5$

$\lambda = 20, k = 6$
$\lambda = 10, k = 9$
$\lambda = 5, k = 13$
**AIC for K-Means**

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda B$$

- Number of free parameters $B = (K \times D)$
  - $K$: # clusters
  - $D$: # dimensions

- $M$: 100 datapoints
- $D$: 2 dimensions
- $k$: 1 cluster

When to stop? How can we compute an upper-bound for $\lambda$?
**AIC for K-Means**

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda B$$

- Idea: Find $\lambda$ which generates AIC curves with values $< RSS(k=1)$ for at least the first $\alpha$ # of parameters.
  
  $$AIC_{RSS}(k = \alpha) \leq RSS(k = 1)$$

  $$RSS(k = \alpha) + \lambda \cdot D \cdot \alpha \leq RSS(k = 1)$$

  $$\lambda \geq \frac{RSS(k = 1) - RSS(k = \alpha)}{D \cdot \alpha}$$

  For $\alpha = \frac{M}{4}$, $\lambda \leq 20$

- Number of free parameters $B = (K \times D)$
  
  $K$: # clusters
  
  $D$: # dimensions

- $M$: 100 datapoints

- $D$: 2 dimensions
AIC for K-Means

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda B$$

There is no particular way of estimating an optimal $\lambda$, however we can estimate an ‘optimal range’ of $\lambda$’s which yield a good trade-off, given the characteristics of the dataset:

$$\frac{RSS(K=1)}{D \cdot M} \leq \lambda \leq \frac{RSS(K=1) - RSS(K=\alpha)}{D \cdot \alpha}, \alpha \in [2, M]$$
AIC for K-Means

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda (K \cdot D)$$

There is no particular way of estimating an optimal $\lambda$, however we can estimate an ‘optimal range’ of values which yield a good trade-off, given the characteristics of the dataset:

- $4 \leq \lambda \leq 20$
- $\alpha = \frac{M}{4}$

$M$: 100 datapoints
$D$: 2 dimensions
$k$: 13 clusters
AIC for K-Means

How do we choose $\lambda$ such that we can achieve an optimal trade-off between distortion ($RSS$) and model complexity ($\lambda B$)?

$$AIC_{RSS} = RSS + \lambda(K \cdot D)$$

There is no particular way of estimating an optimal $\lambda$, however we can estimate an ‘optimal range’ of values which yield a good trade-off, given the characteristics of the dataset:

$2 \leq \lambda \leq 11$

$\alpha = M/4$

$M$: 100 datapoints

$D$: 2 dimensions

$k$: 4 clusters
BIC for K-Means

For the particular case of K-means, we do not have a maximum likelihood estimate of the model:

$$BIC = -2\ln(L) + \ln(M)B$$

However, we can formulate a metric based on the $RSS$ that penalizes for model complexity (# K-clusters, # M-datapoints), conceptually following $BIC$:

$$BIC_{RSS} = RSS + \ln(M)B$$

Weighting factor penalizes wrt. # datapoints (i.e. computational complexity)

Number of free parameters $B = (K*D)$
- $K$: # clusters
- $D$: # dimensions
BIC for K-Means

\[ BIC_{RSS} = RSS + \ln(M) (K \cdot D) \]

- **M**: 100 datapoints
- **D**: 2 dimensions
- **K**: 14 clusters
BIC for K-Means

\[ BIC_{RSS} = RSS + \ln(M) (K \cdot D) \]

\( M: 100 \) datapoints
\( D: 2 \) dimensions
\( K: 4 \) clusters