Unsupervised Learning: Model Selection and Evaluation

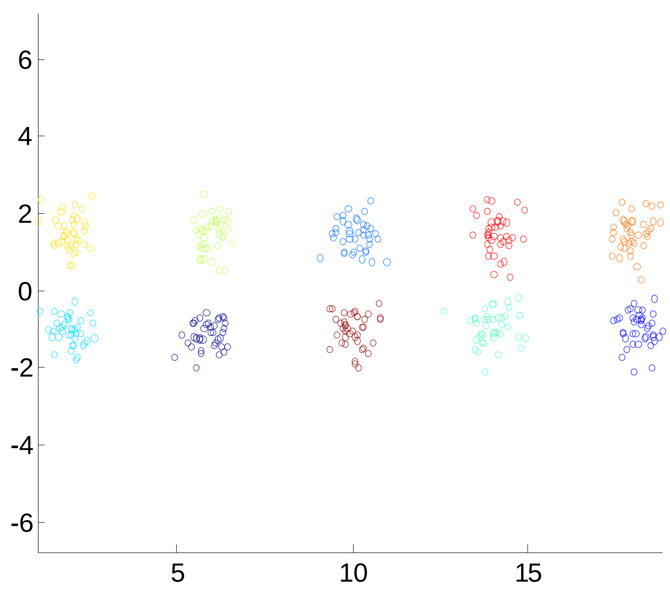
CS-534

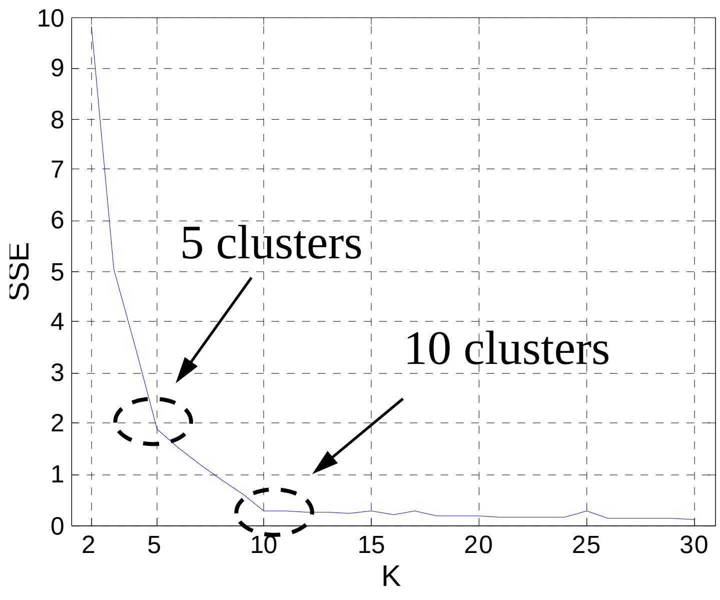
Selecting k: A Model Selection Problem

- Each choice of k corresponds to a different statistical model for the data
- Model selection searches for a model (a choice of k) that gives us the best fit of the training data
  - Penalty method
  - Cross-validation method
  - Model selection methods can also be used to make other model decisions such as choosing among different ways of constraining Σ

Selecting k: heuristic approaches

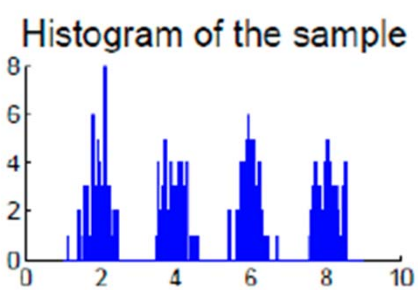
- For kmeans, plot the sum of squared error for different k values
  - SSE will monotonically decrease as we increase k
  - The knee points on the curve suggest good candidates for k

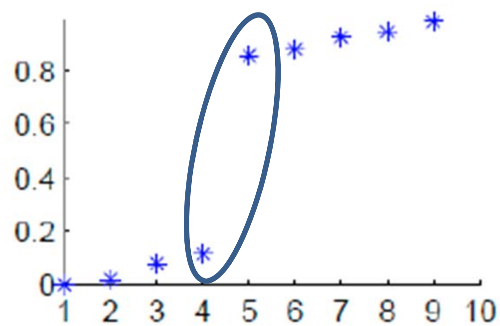




Selecting k: heuristic approaches

- For kmeans, plot the sum of squared error for different k values
  - SSE will monotonically decrease as we increase k
  - The knee points on the curve suggest good candidates for k
- Spectral clustering
  - Find the k that maximizes the eigen-gap





Eigen-values of $D^{-1}W$

Penalty Method: Bayesian Information Criterion

- Based on Bayesian Model Selection
  - Determine the range of k values to consider $1 \leq k \leq K_{max}$
  - Apply EM to learn a maximum likelihood fitting of the Gaussian mixture model for each possible value of *k*
  - Choose *k* that maximizes BIC

# of data points

$$\underbrace{2l_{\mathcal{M}}(x, \hat{\theta})}_{\text{Loglikelihood of the resulting Gaussian Mixture Model}} - m_{\mathcal{M}} \log(n) \equiv \text{BIC}$$

# of parameters to be estimated in *M*

  - Given two estimated models, the model with higher BIC is preferred
  - Larger k increases the likelihood, but will also cause the second term to increase
  - Often observed to be biased toward less complex model
  - Similar method: $\text{AIC} = 2l_m - 2m_M$, which penalize complex model less severely

Cross-validation Likelihood
(Smyth 1998)

- The likelihood of the training data will always increase as we increase k – more flexibility leads to better fitting of the data
- So we cannot use training data likelihood to perform model selection
- How about the likelihood on unseen data?
  - For each possible choice of k
  - Randomly split the data into training and test set
  - Learn the GMM model using the training data and compute the log-likelihood on test data
  - Repeat this multiple times to get a stable estimate of the test log-likelihood
  - Select k that maximizes the test log-likelihood

Stability based method

- Stability: repeatedly produce similar clusterings on data originating from the same source.
- High level of agreement among a set of clusterings $\Rightarrow$ the clustering model (k) is appropriate for the data
- Evaluate multiple models, and select the model resulting in the highest level of stability.

Assessing Stability

- Based on resampling (Levine & Domany, 2001)
- For each k
  1. Generate clusterings on random samplings of the original data
  2. Compute pairwise similarity between each pair of clusterings
  3. Stability(k) = mean pairwise similarity.
- Select k that maximize stability

- Based on prediction accuracy (Tibshirani et al., 2001)
- For each k
  1. Randomly split data into training and testing
  2. For each split
     - cluster the training data using k
     - Predict assignment for test set and compare it to the clustering result on test set
  3. Stability(k) = mean Prediction strength
- Select k that maximize stability

How to Evaluate Clustering?

- By user interpretation
  - does a document cluster seem to correspond to a specific topic?
- Internal criterion – a good clustering will produce high quality clusters:
  - high intra-cluster similarity
  - low inter-cluster similarity

  - The measured quality of a clustering depends on both the object representation and the similarity measure used

External indexes

If true class labels (*ground truth*) are known, the validity of a clustering can be verified by comparing the class labels and clustering labels.

$$\begin{array}{c|c} \boxed{N} & \cdot \\ \hline \cdot & n_{\cdot\cdot} \end{array} = \begin{array}{cccc|c} n_{11} & n_{12} & \cdots & n_{1l} & n_{1\cdot} \\ n_{21} & n_{22} & \cdots & n_{2l} & n_{2\cdot} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{k1} & n_{k2} & \cdots & n_{kl} & n_{k\cdot} \\ \hline n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot l} & n_{\cdot\cdot} \end{array}$$

n_{ij} = number of objects in class *i* and cluster *j*

Rand Index and Normalized Rand Index

- Given partition (*P*) and ground truth (*G*), measure the number of vector pairs that are:

  a: in the same class <u>both</u> in *P* and *G*.
  b: in the same class in *P*, but different classes in *G*.
  c: in different classes in *P*, but in the same class in *G*.
  d: in different classes <u>both</u> in *P* and *G*.

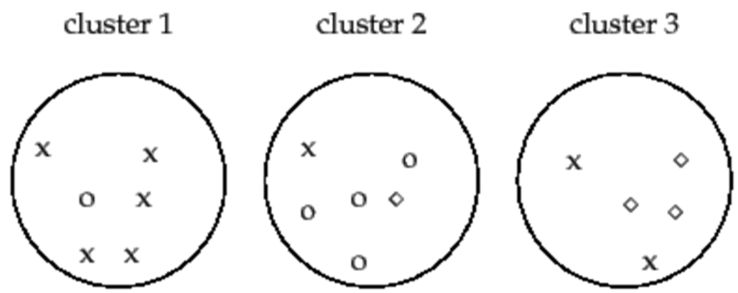
$$R = \frac{a + d}{a + b + c + d}$$

- Adjusted rand index: corrected-for-chance version of rand index
  - Compare to the expectation of the index assuming a random partition of the same cluster sizes

$$ARI = \frac{Index - ExpectedR}{MaxIndex - ExpectedR} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}$$

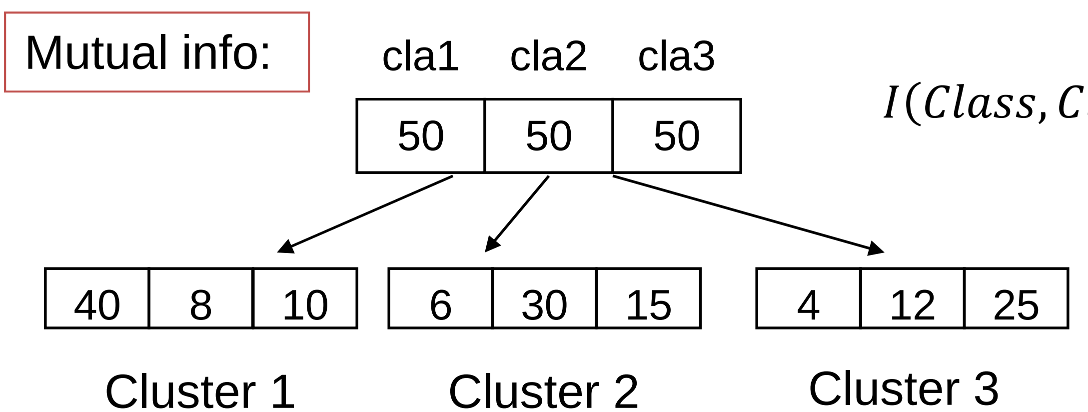
Purity and Normalized Mutual Information

- Purity



▶ **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

- Normalized Mutual Information



$$I(Class, Clust) = H(Class) - H(Class|Clust)$$

$$NMI = \frac{2I(Class, Clust)}{H(Clust) + H(Class)}$$