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 $\Sigma$. 
Selecting $k$: heuristic approaches

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

\[
2l_{\mathcal{M}}(x, \hat{\theta}) - m_{\mathcal{M}} \log(n) \equiv \text{BIC}
\]

- 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: 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 clusters, more flexibility leads to better fitting of the data

• Use cross-validation
  – For each fold, learn the GMM model using the training data
  – Compute the log-likelihood of the learned model on the remaining fold as test data
Stability Based Methods

• Stability: repeatedly produce similar clusterings on data originating from the same source.

• High level of agreement among a set of clusterings ⇒ 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.

\[ n_{ij} = \text{number of objects in class } i \text{ and cluster } j \]
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 both 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 both 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

\[
ARI = \frac{\text{Index} - \text{ExpectedR}}{\text{MaxIndex} - \text{ExpectedR}} = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[ \sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[ \sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2} \right] - \left[ \sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2} \right] / \binom{n}{2}}
\]
Purity and Normalized Mutual Information

• Purity

![Figure 16.1](image)

> 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: $\nu$, 5 (cluster 1); $\omega$, 4 (cluster 2); and $\nu$, 3 (cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

• Normalized Mutual Information

\[ I(Class, Clust) = H(Class) - H(Class | Clust) \]

\[ NMI = \frac{2I(Class, Clust)}{H(Clust) + H(Class)} \]
References