AN MDL FRAMEWORK FOR DATA CLUSTERING

Petri Myllymäki

Complex Systems Computation Group (CoSCo)
Helsinki Institute for Information Technology (HIIT), Finland

P.Kontkanen, P.Myllymäki, W.Buntine, J.Rissanen, H.Tirri, An MDL Framework for Data Clustering. In Advances in Minimum Description Length: Theory and Applications, edited by P. Grünwald, I.J. Myung and M. Pitt. The MIT Press, 2005.

HELSINKI
INSTITUTE FOR
INFORMATION
TECHNOLOGY

CoSCo

Defining the Problem

☞ Given a set of data vectors, define clustering as a *data partitioning* problem

- Clustering is "hard" as opposed to "soft" clustering offered by the model estimation approaches (e.g., mixture modelling)

☞ A data assignment (partitioning) can be represented as a vector of cluster labels

- Given a set of n vectors $\mathbf{x}^n$, find the best *clustering vector* y^n
- Clustering is flat, there is no hierarchy between the clusters

☞ The number of clusters (possible labels) is unknown, determining it is part of the problem

- Need to be able to compare clusterings with different number of cluster labels

☞ Distinguish between

- selection criterion (a function determining the goodness of a clustering), and
- search (a procedure for finding good clusterings)

Information-Theoretic Clustering

- Intuitive idea: assign together those data vectors that compress well together
- Why?
  - In order to compress several data vectors together in an optimal manner, you need to capture all the common regularities found in the data
  - Hence, the more the data vectors in a cluster are "similar" (the more they are governed by the same regularities), the better you can compress the cluster
  - The total code length (sum of all the compressed clusters) is a global criterion forming a dependence between the clusters
  - Code length offers a "universal scale", making it possible to compare clusterings of different complexity, i.e., with different number of cluster labels ("Occam's razor")
- *P.Kontkanen, P.Myllymäki, W.Buntine, J.Rissanen, H.Tirri, An MDL Framework for Data Clustering*
  - focus on comparing different clustering criteria.
- *P.Kontkanen, P.Myllymäki, An Empirical Comparison of NML Clustering Algorithms. (Pascal Workshop on Statistics and Optimization of Clustering, July 2005.)*
  - use one clustering criterion, focus on comparing different search algorithms.

Stochastic Complexity

- Central concept in the *Minimum Description Length (MDL)* framework for statistical modeling.
- Stochastic complexity = the shortest description length of a given data set relative to a model class $\mathcal{M}$.
- Model class: a set of parametric distributions indexed by elements of $\Theta \in \mathbb{R}^d$:

$$\mathcal{M} = \{P(\cdot|\theta) : \theta \in \Theta\}.$$

- Old formalizations of SC: BIC, marginal likelihood.
- Modern formalization: **Stochastic complexity = Normalized Maximum Likelihood (NML)**.

Normalized Maximum Likelihood

- The maximum likelihood model $\hat{\theta}(D)$ (in model class $\mathcal{M}$) with respect to data D is

$$\hat{\theta}(D) = \arg\max_{\theta \in \Theta} \{P(D \mid \theta, \mathcal{M})\}.$$

- Define stochastic complexity as the result of the following minmax optimization problem:

$$P_{NML}(\cdot) = \arg\min_{Q} \max_{D'} (\log P(D' \mid \hat{\theta}(D'), \mathcal{M}) - \log Q(D'))$$

- *Solution (the NML distribution/code)*:

$$P_{NML}(D) = \frac{P(D \mid \hat{\theta}(D), \mathcal{M})}{\sum_{D'} P(D' \mid \hat{\theta}(D'), \mathcal{M})}$$

An Example Model Class

☞ Assume the observed data $\mathbf{x}^n$ to consist of values of m discrete variables $X_1, \ldots, X_m$.

☞ The cluster labels y^n are interpreted as missing data concerning a discrete variable Y.

☞ The "goodness" of a clustering y^n is defined as the NML code length for the complete data $D = (\mathbf{x}^n, y^n)$ with respect to chosen model class.

☞ The local independence model class: variables $X_1, \ldots, X_m$ are conditionally independent given the value of Y (the "Naive Bayes" model).

- Computational reasons
- Easily interpretable clusters

The NML Clustering Criterion

- The optimal clustering y^n is the one that leads to shortest code length when the clustering is compressed together with the observed data $\mathbf{x}^n$ with respect to the NML distribution

$$P_{NML}(\mathbf{x}^n, y^n) = \frac{P(\mathbf{x}^n, y^n \mid \hat{\theta}(\mathbf{x}^n, y^n))}{\sum_{\mathbf{x}'^n, y'^n} P(\mathbf{x}'^n, y'^n \mid \hat{\theta}(\mathbf{x}'^n, y'^n))}.$$

- In principle, computing the denominator (”regret”) takes exponential time, but:
- With respect to the local independence model class discussed earlier, can be computed *exactly* in time $O(n + K)$ and approximated in time $O(K)$ (where n is the size of the observed data set $\mathbf{x}^n$ and K is the number of cluster labels found in y^n).

Summary

☞ Formulated clustering as a missing data estimation problem

☞ Presented an information-theoretic NML criterion for choosing between alternative clusterings

☞ For an interesting, practically useful model class, the criterion can be computed efficiently

☞ Empirical results are very encouraging

☞ Clustering search space still exponential, clever heuristics are necessary