Machine learning: lecture 17
Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Topics
• Semi-supervised clustering cont’d
  – clustering by dynamics
  – semi-supervised clustering

Overview of clustering methods
• Flat clustering methods
  – e.g., mixture models, k-means clustering
• Hierarchical clustering methods:
  – Top-down (splitting)
    + e.g., hierarchical mixture models
  – Bottom-up (merging)
    + e.g., hierarchical agglomerative clustering
• Spectral clustering
• Semi-supervised clustering
• Clustering by dynamics
  Etc.

Semi-supervised clustering
• Let’s assume we have some additional relevance information about the examples and we’d like clusters to preserve this information as much as possible.

For example, by merging together documents we do not wish to lose information about the words they contain (word distributions).

Semi-supervised clustering cont’d
• We cluster documents \( \{ x_i \} \) on the basis of their word distributions \( \{ P(y|x_i) \} \)
• The word distribution for a cluster is the average word distribution of documents in the cluster

\[
P(y|C) = \frac{1}{|C|} \sum_{i \in C} P(y|x_i)
\]

• When merging two clusters we need to take into account their sizes: for example, if \( C_2 = C_1 \cup x_3 \) then

\[
P(y|C_2) = \frac{1}{2 + 1} \left( 2 \cdot P(y|C_1) + 1 \cdot P(y|x_3) \right)
\]
**Semi-supervised clustering cont’d**
- We still need to specify a distance metric to determine which clusters to merge and in what order.
- The distance should reflect how much relevance information we lose by merging the clusters
  \[ d(C_k, C_l) = (|C_k| + |C_l|) \cdot I(y; \text{cluster identity}) \]
  \[ = |C_k| \sum_y P(y|C_k) \log \frac{P(y|C_k)}{P(y|C_k \cup C_l)} + |C_l| \sum_y P(y|C_l) \log \frac{P(y|C_l)}{P(y|C_k \cup C_l)} \]

**Clustering by dynamics**
- We may wish to cluster time course signals not by direct comparison but in terms of dynamics that governs the signals
  - system behavior monitoring (anomaly detection)
  - biosequences, processes etc.
- We will use Markov models to capture the dynamics. The distance metric for clustering is based on similarity of models.

**Modeling time course signals**
- Full probability model
  \[ P(s_1, \ldots, s_t) = P(s_1)P(s_2|s_1)P(s_3|s_2)P(s_4|s_3) \cdots \]
- First order Markov model
  \[ P(s_1, \ldots, s_t) = P(s_1)P(s_2|s_1)P(s_3|s_2)P(s_4|s_3) \cdots \]

**Discrete Markov models**
- Representation in terms of variables and dependencies (a graphical model):
  \[ P(s_1, \ldots, s_t) = P(s_1)P(s_2|s_1)P(s_3|s_2)P(s_4|s_3) \cdots \]
- Representation in terms of state transitions (transition diagram)

**Discrete Markov models: properties**
- The values of each \( s_t \) are known as states
  \[ \begin{align*}
  & P(\cdot): & 0.5 & 0.5 \\
  & P(\cdot|s_1): & 0.5 & 0.5 \\
  & P(\cdot|s_2): & 0.0 & 1.0 \\
  \end{align*} \]
- When successive state transitions are governed by the same (one-step) transition probability matrix \( P(s_t|s_{t-1}) \), the Markov model is homogeneous
Discrete Markov models: properties

- The values of each $s_t$ are known as states

$$P(\cdot|\cdot)_{v/ljl}$$

- When successive state transitions are governed by the same (one-step) transition probability matrix $P_1(s_t|s_{t-1})$, the Markov model is homogeneous

- Example: a language model

This $\rightarrow$ is $\rightarrow$ a $\rightarrow$ boring $\rightarrow$ . . .

Is a homogeneous Markov model appropriate in this case?

Discrete Markov models: ML estimation

$$l(S^{(1)}, S^{(2)}) = \sum_{i=1,2} \log P(s^{(1)}_t) + \sum_{t=2}^{T} \log P_1(s^{(i)}_t|s^{(i)}_{t-1})$$

- ML estimates of the parameters (initial state and transition probabilities) are based on simple counts:

$$\hat{n}(s) = \# \text{ of times } s_1 = s$$

$$\hat{n}(r,s) = \# \text{ number of times } r \rightarrow s$$

$$\hat{P}(s) = \frac{\hat{n}(s)}{\sum_{s'} \hat{n}(s')}$$

$$\hat{P}_1(s|r) = \frac{\hat{n}(r,s)}{\sum_{s'} \hat{n}(r,s')}$$

Simple clustering example cont’d

- Four binary sequences of length 40:
  1. 0010011001000101000001000011101101010100
  2. ...
  3. 110101100000110110011000111011101111011001
  4. 110101011110101111011110111011011001

- We can estimate a Markov model based on any subset of the sequences
- We still need to derive the clustering metric based on the resulting transition probabilities (dynamics)

Clustering metric

- To determine a distance between two arbitrary sequences

$$S^{(1)} = \{s^{(1)}_1, \ldots, s^{(1)}_{n_1}\} \text{ and } S^{(2)} = \{s^{(2)}_1, \ldots, s^{(2)}_{n_2}\}$$

we measure how well a Markov model would capture the sequences separately or jointly.

$$l(S^{(1)}|\hat{\theta}_1) = \log P(S^{(1)}|\hat{\theta}_1)$$

$$l(S^{(2)}|\hat{\theta}_2) = \log P(S^{(2)}|\hat{\theta}_2)$$

$$l(S^{(1)}, S^{(2)}|\hat{\theta}) = \log P(S^{(2)}|\hat{\theta}) + \log P(S^{(2)}|\hat{\theta})$$

where the parameters are ML estimates. Our distance is now defined as

$$d_M(S^{(1)}, S^{(2)}) = l(S^{(1)}) + l(S^{(2)}) - l(S^{(1)}, S^{(2)})$$

Simple example cont’d

- Four binary sequences of length 40:
  1. 0010011001000101000001000011101101010100
  2. 01011111101011000100001000100010110000101
  3. 110101100000110110011000111011101111011001
  4. 110101011110101111011110111011011001

- The resulting pairwise distance matrix:

(dists need to be recomputed after merging)
Clustering metrics, equivalence

- We have seen two different ways of inducing a metric, one based on information theory, the other through defining a model. These are closely related ideas.

Suppose we transform each observed binary sequence 0010011001000... into a bag of pairs {00, 01, 10, ...}. Let’s use a variable $y$ to refer to the possible pairs (4 of them). For example, $y = 1$ would mean 00.

Let’s model each sequence with a distribution over pairs $P(y|S^{(1)})$ whose ML estimate is obtained by counting

$$\hat{P}(y|S^{(1)}) = \frac{n(y)}{n}$$

The maximum value of the log-likelihood is given by

$$l(S^{(1)}) = \sum_{i} \log \hat{P}(y|S^{(1)}) = \sum_{y} n(y) \log \hat{P}(y|S^{(1)})$$

$$= n \sum_{y} \frac{n(y)}{n} \log \hat{P}(y|S^{(1)})$$

$$= n \sum_{y} \hat{P}(y|S^{(1)}) \log \hat{P}(y|S^{(1)})$$

Similarly, if we merge two sequences

$$l(S^{(1)}, S^{(2)}) = n \sum_{y} \hat{P}(y|S^{(1)}) \log \hat{P}(y|S^{(1)} \cup S^{(2)})$$

$$+ n \sum_{y} \hat{P}(y|S^{(2)}) \log \hat{P}(y|S^{(1)} \cup S^{(2)})$$

The model based metric is now given by

$$d_{Mf}(S^{(1)}, S^{(2)}) = l(S^{(1)}) + l(S^{(2)}) - l(S^{(1)}, S^{(2)})$$

$$= n \sum_{y} \hat{P}(y|S^{(1)}) \log \frac{\hat{P}(y|S^{(1)})}{\hat{P}(y|S^{(1)} \cup S^{(2)})}$$

$$+ n \sum_{y} \hat{P}(y|S^{(2)}) \log \frac{\hat{P}(y|S^{(2)})}{\hat{P}(y|S^{(1)} \cup S^{(2)})}$$

$$= n \cdot I(y; \text{seq identity})$$
Beyond Markov models

- How can we model
  1. 0101010101010101010101010101010101010101...
  2. 0010010010010010010010010010010010010010...
  3. 0100100010000101001000100001010010001000...

- What about