Topics

- Clustering
  - semi-supervised clustering
  - clustering by dynamics: Markov models

- Structured probability models
  - Hidden Markov models
Semi-supervised clustering

- Let’s assume we have some additional *relevance* information about the examples to be clustered

  - $x_i$: Training example (e.g., a text document)
  - $y$: Relevance variable (e.g., a word)
  - $P(y|x_i)$: Relevance information (e.g., word distribution)

  where $i = 1, \ldots, n$.

- We wish to cluster documents into larger groups without losing information about words contained in the documents (documents with similar word frequencies should be merged into a single cluster).
Semi-supervised clustering cont’d

- We cluster the examples \( \{x_i\} \) on the basis of \( \{P(y|x_i)\} \), the predictive distributions.

- For any cluster \( C \) we define the predictive word distribution based on randomly picking a document in the cluster:

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

\[
\hat{P}(C) = \frac{|C|}{n}
\]
Semi-supervised clustering cont’d

- The distance between any two clusters measures how much information we lose about the words if the clusters are merged

\[
d(C_l, C_k) = (\hat{P}(C_l) + \hat{P}(C_k)) \cdot I(y; \text{cluster identity})
\]

\[
P(y|C_3) \quad P(y|C_1) \quad P(y|C_2)
\]

\[
P(y|x_1) \quad P(y|x_2) \quad P(y|x_3) \quad P(y|x_4)
\]
Semi-supervised clustering cont’d

- The distance between the clusters measures how much information we lose about the words if the clusters are merged

\[ d(C_l, C_k) = \left( \hat{P}(C_l) + \hat{P}(C_k) \right) \cdot I(y; \text{cluster identity}) \]

where

\[ I(y; \text{cluster identity}) = \]

\[
\frac{1}{\hat{P}(C_l) + \hat{P}(C_k)} \left[ \hat{P}(C_l) \sum_{j=1}^{m} \hat{P}(y = j|C_l) \log \frac{\hat{P}(y = j|C_l)}{\hat{P}(y = j|C_l \cup C_k)} + \hat{P}(C_k) \sum_{j=1}^{m} \hat{P}(y = j|C_k) \log \frac{\hat{P}(y = j|C_k)}{\hat{P}(y = j|C_l \cup C_k)} \right]
\]
Semi-supervised clustering: example

- Suppose we have a set of labeled examples $(x_1, y_1), \ldots, (x_n, y_n)$

- We can take the label as the relevance variable.

\[ P(y|x_i) = 1, \text{ if } y = y_i \text{ and zero otherwise} \]
Topics

- Clustering
  - semi-supervised clustering
  - clustering by dynamics: Markov models

- Structured probability models
  - Hidden Markov models
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.

1. 0010011001000101000001000011101101010100...
2. 01011111110100110101000001000000101011001...
3. 110101100000110111010001101111101011101...
4. 1101010111101011110111101101101101000101...

- We will use Markov models to capture the dynamics and a model selection criterion to induce an appropriate similarity measure
Modeling time course signals

- Full probability model

\[ P(s_1, \ldots, s_t) = P(s_1)P(s_2|s_1)P(s_3|s_2, s_1)P(s_4|s_3, s_2, s_1) \cdots \]

- 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

![Diagram of Markov model](image)

- 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
Discrete Markov models: properties

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

  ![Diagram of Markov model][1]

  - $P(\cdot)$:
    - $P(\cdot)$: 0.5
    - $P(\cdot)$: 0.5

  - $P_1(\cdot|\cdot)$:
    - $P_1(\cdot|\cdot)$: 0.5 0.5
    - $P_1(\cdot|\cdot)$: 0.0 1.0

- 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

  ![Language model diagram][2]

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

Is a homogeneous Markov model appropriate in this case?
Discrete Markov models: properties

- The values of each \( s_t \) are known as states

- 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 \textit{homogeneous}

- If after \( k \) transitions we can get from any state \( i \) to any other state \( j \), the Markov chain is \textit{ergodic}.

In other words, the \( k \)-step transition probabilities must satisfy \( P_k(s_{k+1} = j|s_1 = i) > 0 \) for all \( i, j \).
Discrete Markov models: ML estimation

1. 0010011001000101000001000011101101010100...
2. 0101111110100110101000001000000101011001...

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

\[ \hat{n}(i) = \# \text{ of times } s_1 = i \]
\[ \hat{n}(i, j) = \# \text{ number of times } i \rightarrow j \]
\[ \hat{P}(i) = \frac{\hat{n}(i)}{\sum_{i'} \hat{n}(i')} \]
\[ \hat{P}_1(j|i) = \frac{\hat{n}(i, j)}{\sum_{j'} \hat{n}(i, j')} \]
Simple clustering example cont’d

- Four binary sequences of length 50:
  1. 001001100100010100000100001110110101010100…
  2. 010111111010011010100001000000101011001…
  3. 1101011000000110110010001101111101011101…
  4. 1101010111101011110111101101101100101…

- We still need to derive the clustering metric based on the transition probabilities (dynamics)

- We can turn the clustering problem into a model selection problem: which sequences should be modeled with the same transition probabilities?
Cluster criterion

- To determine whether two arbitrary sequences

\[ S^{(1)} = \{ s_{1}^{(1)}, \ldots, s_{n_1}^{(1)} \} \quad \text{and} \quad S^{(2)} = \{ s_{1}^{(2)}, \ldots, s_{n_2}^{(2)} \} \]

should be in the same cluster, we compare (approximate) description lengths of either encoding the sequences separately or jointly

\[ DL^{(1)} + DL^{(2)} \geq DL^{(1+2)} \]

where \( DL^{(1+2)} \) uses the same Markov model for both sequences whereas \( DL^{(1)} \) and \( DL^{(2)} \) use models specific to each sequence.
Cluster criterion cont’d

- Approximate description lengths:

\[
DL^{(1)} + DL^{(2)} = - \log P(S^{(1)}|\hat{\theta}_1) + \frac{d}{2}\log(n_1) \\
- \log P(S^{(2)}|\hat{\theta}_2) + \frac{d}{2}\log(n_2)
\]

\[
DL^{(1+2)} = - \log P(S^{(1)}|\hat{\theta}) - \log P(S^{(2)}|\hat{\theta}) \\
+ \frac{d}{2}\log(n_1 + n_2)
\]

where the maximum likelihood parameter estimates \(\hat{\theta}_1, \hat{\theta}_2,\) and \(\hat{\theta}\) include the initial state distribution and the transition probabilities; \(d = 3\) for binary sequences.

- We are essentially testing here whether the two sequences have the same first order Markov dynamics.
Simple example cont’d

- Four binary sequences of length 50:
  1. 001001100100010100001000011101101010100...
  2. 0101111110101010100001000000101011001...
  3. 110101100000110110010001101111101011101...
  4. 1101010111101011110111101101101101000101...

Evaluations:

\[
\begin{align*}
    DL^{(1)} + DL^{(2)} - DL^{(1+2)} &= 6.6 \text{ bits} \\
    DL^{(1+2)} + DL^{(3+4)} - DL^{(1+2+3+4)} &= -0.9 \text{ bits}
\end{align*}
\]

Agglomerative hierarchical clustering with Euclidean distance would give (((2, 3), 4), 1)
Topics

- Clustering
  - semi-supervised clustering
  - clustering by dynamics: Markov models

- Structured probability models
  - Hidden Markov models
Beyond Markov models

- How can we model

  1. 0101010101010101010101010101010101010101...
Beyond Markov models

- How can we model
  1. 0101010101010101010101010101010101010101010101010101010101010101010101...
  2. 0010010010010010010010010010010010010010010010010010010010010010010010010...
Beyond Markov models

- How can we model

1. 0101010101010101010101010101010101010101...
2. 0010010010010010010010010010010010010010...
3. 0100100010000101001000100001010010001000...

Tommi Jaakkola, MIT CSAIL
Beyond Markov models

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

- What about
Hidden Markov models (HMMs)

- HMMs are Markov models with observations: the state variables $s_t$ are not observed directly, only via associated observations $x_t$

$$P(s_1, x_1, s_2, x_2, \ldots) = P(s_1)P(x_1|s_1)P(s_2|s_1)P(x_2|s_2)\cdots$$
Hidden Markov models (HMMs)

\begin{align*}
P(x_1 | s_1 = i) &= N(x_1; \mu_i, \Sigma_i) 
\end{align*}