Machine learning: lecture 15

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Topics

- Clustering
  - Markov random walk and spectral clustering
  - semi-supervised clustering
- Structured probability models
  - Markov models
  - Hidden markov models (next lecture)

Spectral clustering: review

- The spectral clustering method we define relies on a random walk representation over the points. We construct this in three steps

   1. a nearest neighbor graph

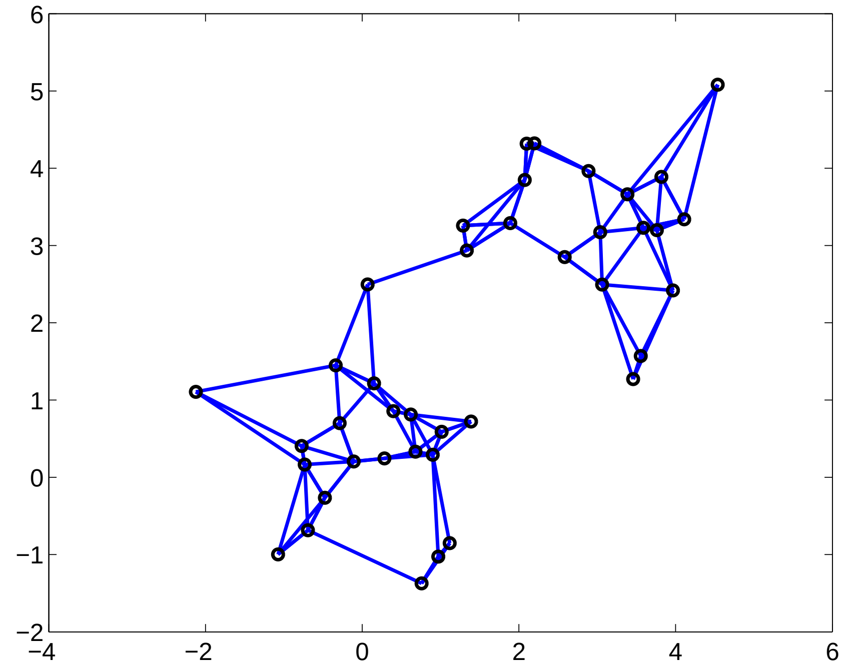
   2. similarity weights on the edges:

$$W_{ij} = \exp\{-\beta \|\mathbf{x}_i - \mathbf{x}_j\|\}$$

      where $W_{ii} = 1$ and the weight is zero for non-edges.

   3. transition probability matrix

$$P_{ij} = W_{ij} / \sum_{j'} W_{ij'}$$



Properties of the random walk

- If we start from i_0, the distribution of points i_t that we end up in after t steps is given by

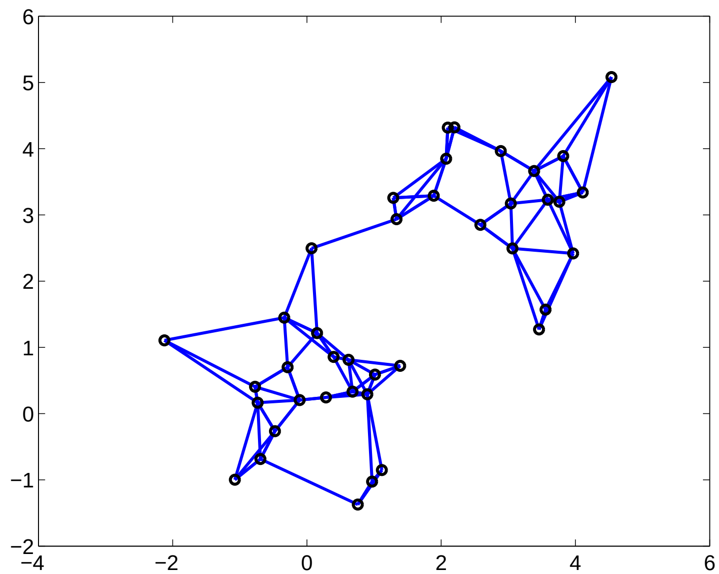
$$\begin{aligned}
i_1 &\sim P_{i_0\,i_1}, \\
i_2 &\sim \sum_{i_1} P_{i_0,i_1} P_{i_1\,i_2} = [P^2]_{i_0\,i_2}, \\
i_3 &\sim \sum_{i_1} \sum_{i_2} P_{i_0,i_1} P_{i_1\,i_2} P_{i_2\,i_3} = [P^3]_{i_0\,i_3}, \\
&\ldots \\
i_t &\sim [P^t]_{i_0\,i_t}
\end{aligned}$$

where $P^t = PP\ldots P$ (t matrix products) and $[\cdot]_{ij}$ denotes the i,j component of the matrix.

Random walk and clustering

- The distributions of points we end up in after t steps converge as t increases. If the graph is connected, the resulting distribution is independent of the starting point

  Even for large t, the transition probabilities $[P^t]_{ij}$ have a slightly higher probability of transitioning within "clusters" than across; we want to recover this effect from eigenvalues/vectors



Eigenvalues/vectors and spectral clustering

- Let W be the matrix with components W_{ij} and D a diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. Then

$$P = D^{-1}W$$

- To find out how P^t behaves for large t it is useful to examine the eigen-decomposition of the following symmetric matrix

$$D^{-\frac{1}{2}}WD^{-\frac{1}{2}} = \lambda_1\mathbf{z}_1\mathbf{z}_1^T + \lambda_2\mathbf{z}_2\mathbf{z}_2^T + \ldots + \lambda_n\mathbf{z}_n\mathbf{z}_n^T$$

where the ordering is such that $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$.

Eigenvalues/vectors cont'd

- The symmetric matrix is related to P^t since

$$(D^{-\frac{1}{2}} W D^{-\frac{1}{2}}) \cdots (D^{-\frac{1}{2}} W D^{-\frac{1}{2}}) = D^{\frac{1}{2}} (P \cdots P) D^{-\frac{1}{2}}$$

This allows us to write the t step transition probability matrix in terms of the eigenvalues/vectors of the symmetric matrix

$$\begin{aligned} P^t &= D^{-\frac{1}{2}} \left(D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \right)^t D^{\frac{1}{2}} \\ &= D^{-\frac{1}{2}} \left(\lambda_1^t \mathbf{z}_1 \mathbf{z}_1^T + \lambda_2^t \mathbf{z}_2 \mathbf{z}_2^T + \ldots + \lambda_n^t \mathbf{z}_n \mathbf{z}_n^T \right) D^{\frac{1}{2}} \end{aligned}$$

where $\lambda_1 = 1$ and

$$P^\infty = D^{-\frac{1}{2}} \left(\mathbf{z}_1 \mathbf{z}_1^T \right) D^{\frac{1}{2}}$$

Eigenvalues/vectors and spectral clustering

- We are interested in the largest correction to the asymptotic limit

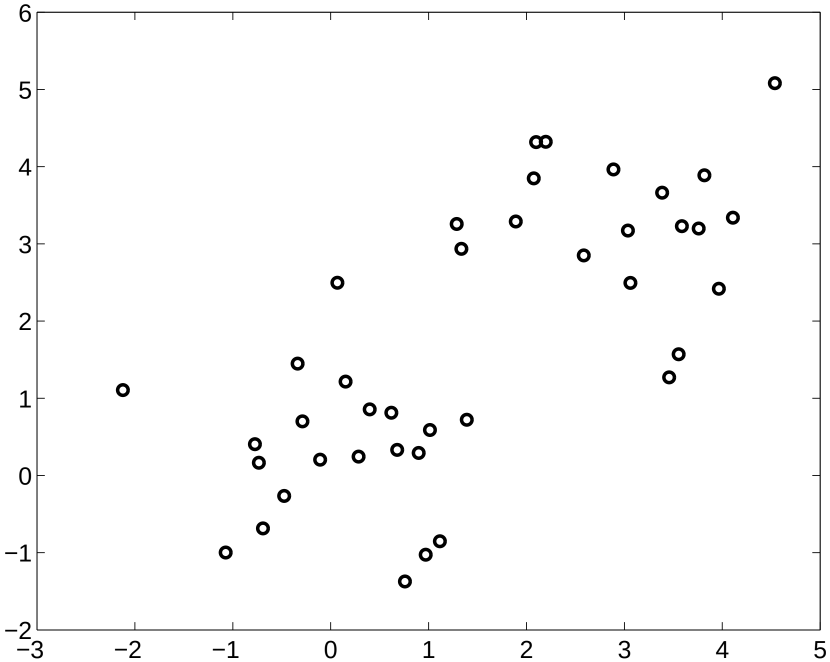
$$P^t \approx P^\infty + D^{-\frac{1}{2}} \left(\lambda_2^t \, \mathbf{z}_2 \mathbf{z}_2^T \right) D^{\frac{1}{2}}$$

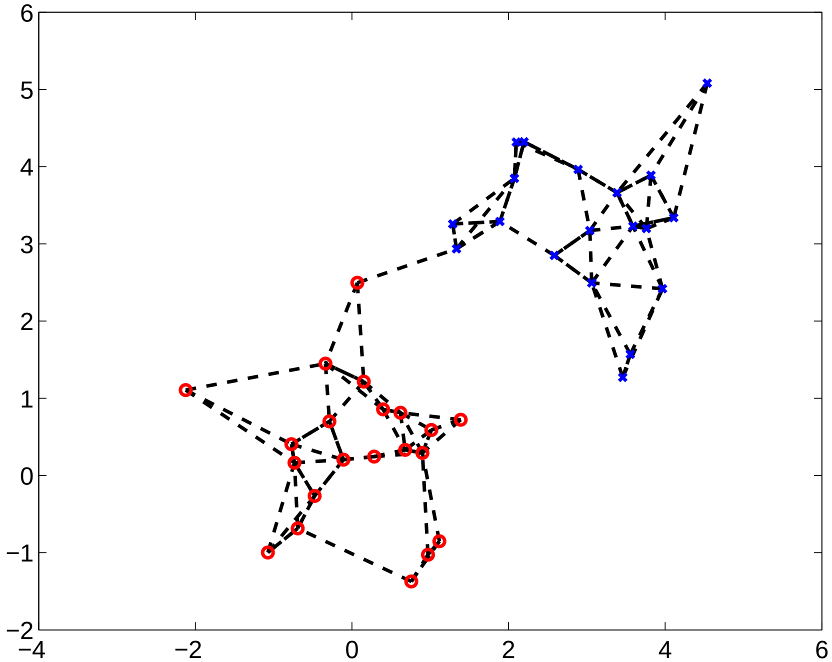
  Note: $[\mathbf{z}_2 \mathbf{z}_2^T]_{ij} = z_{2i} z_{2j}$ and thus the largest correction term increases the probability of transitions between points that share the same sign of z_{2i} and decreases transitions across points with different signs

- Binary spectral clustering: we divide the points into clusters based on the sign of the elements of $\mathbf{z}_2$

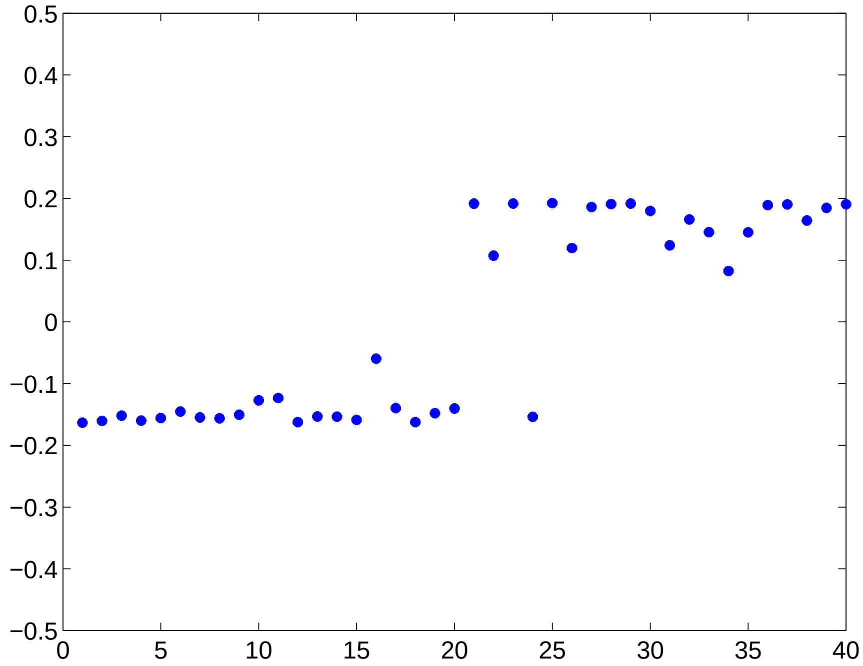
$$z_{2j} > 0 \Rightarrow \text{ cluster 1, otherwise cluster 0}$$

Spectral clustering: example





Spectral clustering: example cont'd



Components of the eigenvector corresponding to the second largest eigenvalue

Semi-supervised clustering

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

$\mathbf{x}_i$	Training example (e.g., a text document)
y	Relevance variable (e.g., a word)
$P(y\|\mathbf{x}_i)$	Relevance information (e.g., word distribution)

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

- We wish to cluster the documents into larger groups without loosing information about words contained in the documents

  documents with similar word frequencies should be merged into a single cluster

Semi-supervised clustering cont'd

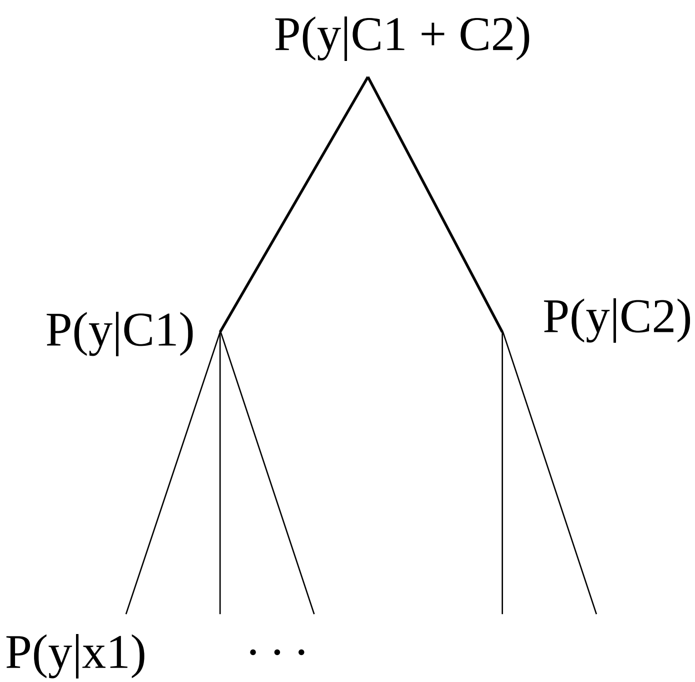
- We cluster the examples $\{\mathbf{x}_i\}$ on the basis of $\{P(y|\mathbf{x}_i)\}$, the predictive distributions
- For any cluster C we define the predictive word distribution based on randomly picking a document in the cluster

$$\begin{aligned}\hat{P}(y = j|C) &= \frac{1}{|C|}\sum_{i \in C} P(y = j|\mathbf{x}_i) \\ \hat{P}(C) &= \frac{|C|}{n}\end{aligned}$$

Semi-supervised clustering cont'd

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

$$d(C_l, C_k) = \frac{|C_l| + |C_k|}{n} \cdot I(\, y\, ;\, \text{cluster identity}\,)$$



Semi-supervised clustering cont'd

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

$$d(C_l, C_k) \;=\; \frac{|C_l| + |C_k|}{n} \cdot I(\,y\,;\,\text{cluster identity}\,)$$

where

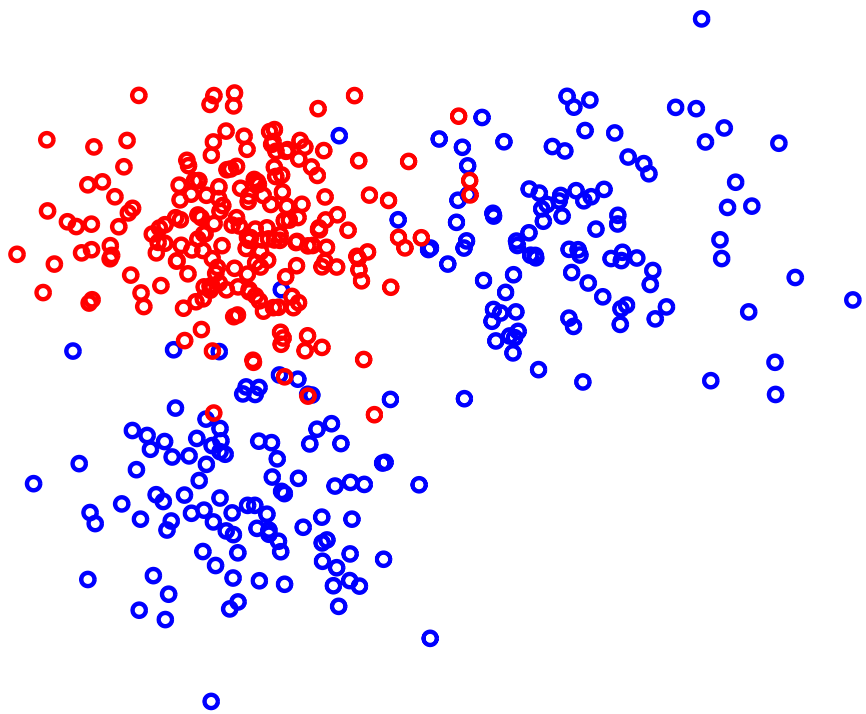
$$I(\,y\,;\,\text{cluster identity}\,) =$$

$$\frac{1}{\hat{P}(C_l) + \hat{P}(C_k)} \Bigg[\hat{P}(C_l) \sum_{j=1}^{m} \hat{P}(y=j|C_l) \log \frac{\hat{P}(y=j|C_l)}{\hat{P}(y=j|C_l \cup C_k)}$$

$$+ \hat{P}(C_k) \sum_{j=1}^{m} \hat{P}(y=j|C_k) \log \frac{\hat{P}(y=j|C_k)}{\hat{P}(y=j|C_l \cup C_k)} \Bigg]$$

Semi-supervised clustering: example

- Suppose we have a set of labeled examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$



- We can take the label as the relevance variable.

$$P(y|\mathbf{x}_i) = 1, \text{ if } y = y_i \text{ and zero otherwise}$$

Topics

- Structured probability models
  - Markov models
  - Hidden markov models (next lecture)

Markov models

- Often we want to model dynamical systems, not just individual examples

  1. Speech/language processing
  2. Human behavior (e.g., user modeling)
  3. Modeling physical/biological processes
  4. Stock market etc.

- We need to define a class of probability models that we can estimate from observed behavior of the dynamical system

Markov chain: definition

- We define here a finite state Markov chain (stochastic finite state machine)

  1. States: $s \in \{1, \ldots, m\}$, where m is finite.
  2. Starting state: The starting state s_0 may be fixed or drawn from some a priori distribution $P_0(s_0)$.
  3. Transitions (dynamics): we define how the system transitions from the current state s_t to the next state s_{t+1}

     The transitions satisfy the first order Markov property:

$$P(s_{t+1}|s_t, \underbrace{s_{t-1}, \ldots, s_0}_{\text{past}}) = P_1(s_{t+1}|s_t)$$

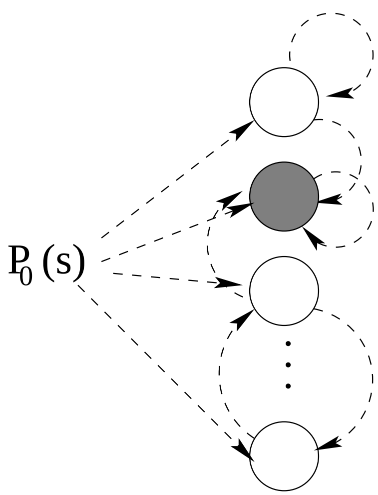
Markov chain cont'd

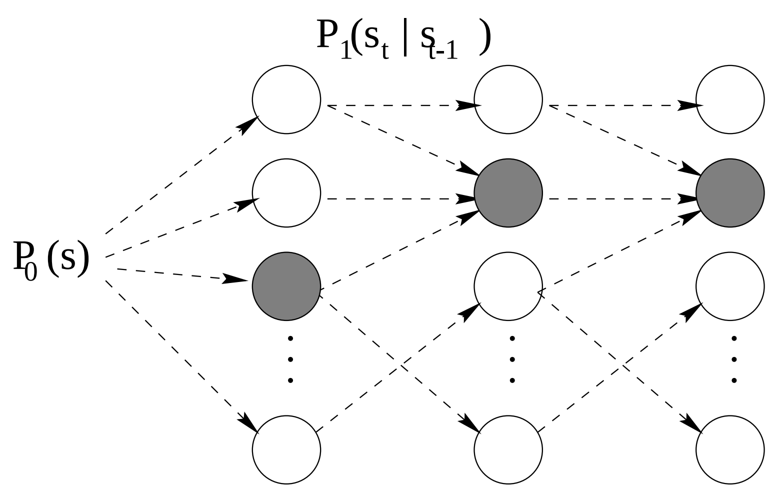
- The resulting stochastic system generates a sequence of states

$$s_0 \longrightarrow s_1 \longrightarrow s_2 \longrightarrow \ldots$$

where s_0 is drawn from $P_0(s_0)$ and each s_{t+1} from one step transition probabilities $P_1(s_{t+1}|s_t)$

- We can represent the Markov chain as a state transition diagram





Markov chain: example

- The states correspond to words in a sentence
- The transitions are defined in terms of successive words in a sentence

  Example: a particular realization of the state sequence

  $$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow \ldots$$

  might be

  $$\text{This} \rightarrow \text{is} \rightarrow \text{a} \rightarrow \text{boring} \rightarrow \ldots$$

- Is Markov chain an appropriate model?