6.867 Machine learning and neural networks

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

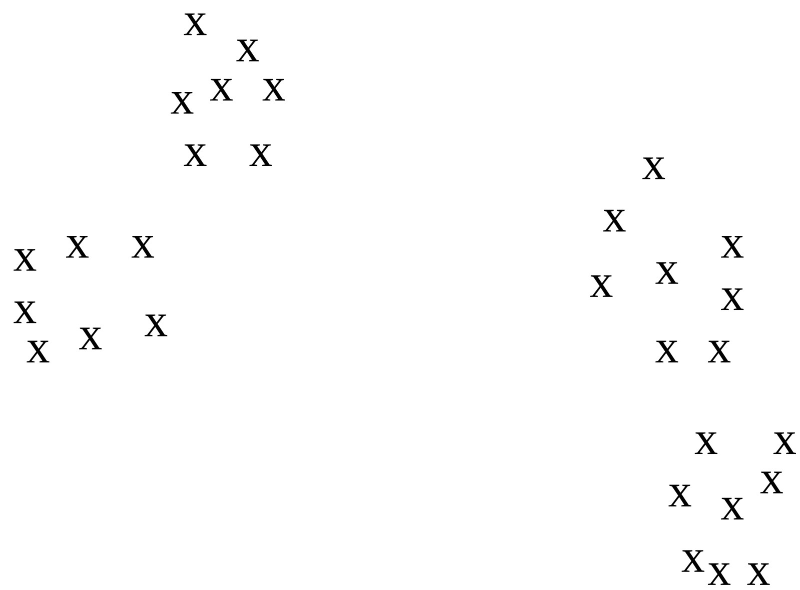
Lecture 15: clustering, markov models

Topics

- Finding structure in the data: clustering
  - flat clustering
  - hierarchical (top-down, bottom-up)
  - semi-supervised
- Markov models
  - motivation, definition
  - prediction, estimation

Finding structure in the data: clustering

- The definion of “ground truth” often missing ...
  - need external or internal validation



- There are various “metrics” for clustering: position in “space”, input/output relation, dynamics, ...

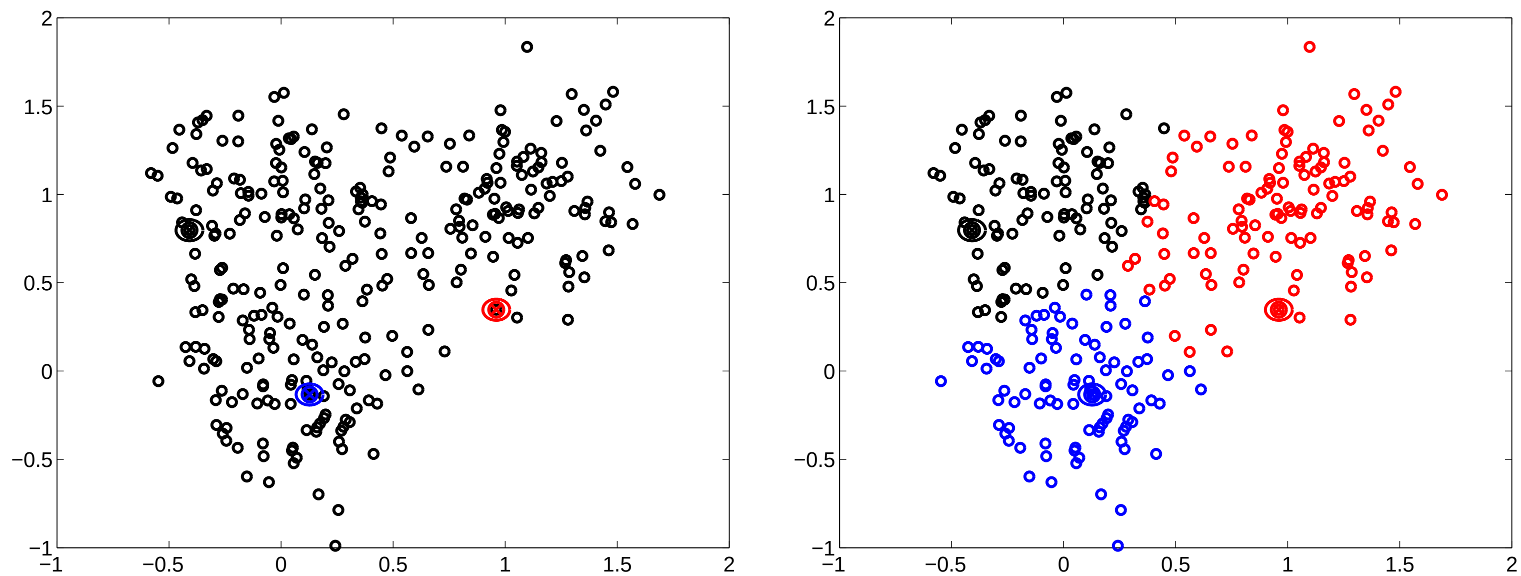
Basic clustering methods

- Flat clustering methods
  - e.g., mixture models, k-means clustering
- Hierarchical clustering methods:
  1. Top-down (splitting)
     - e.g., hierarchical mixture models
  2. Bottom-up (merging)
     - e.g., hierarchical agglomerative clustering
- Other clustering methods: spectral clustering, semi-supervised clustering, etc.

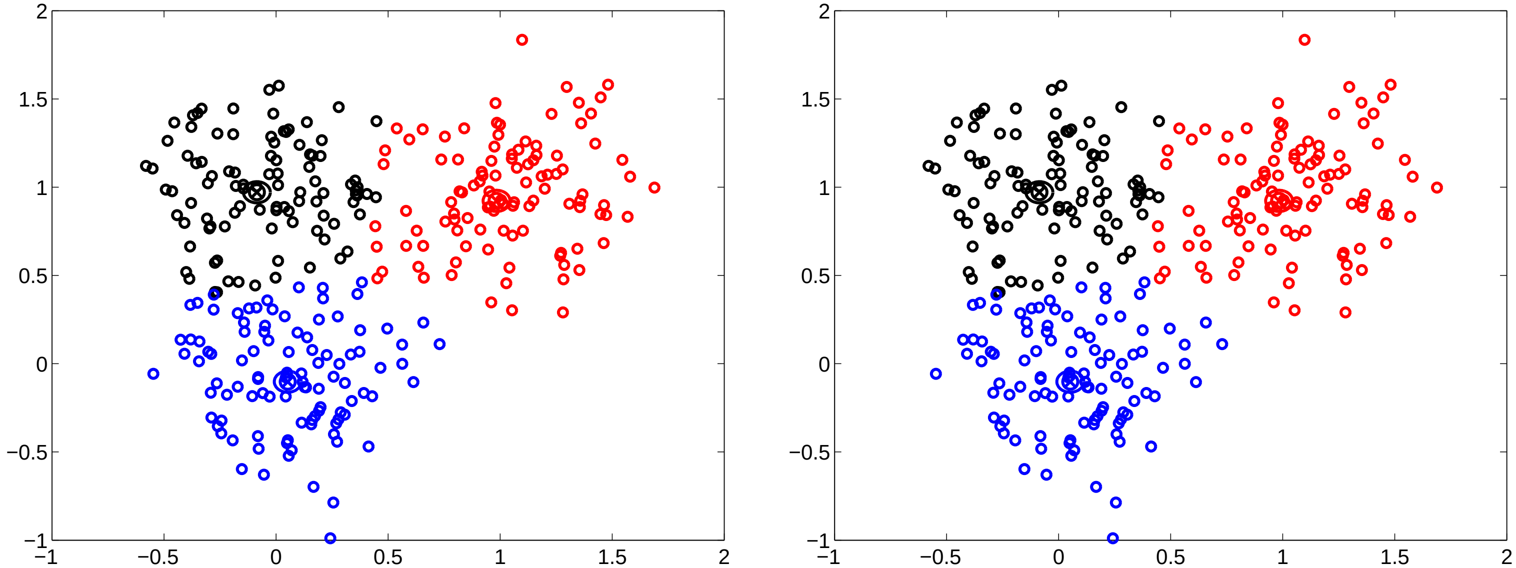
K-means clustering

- The procedure:
  1. Pick k arbitrary centroids (cluster means)
  2. Assign each example to its "closest" centroid (**E-step**)
  3. Adjust the centroids to be the means of the examples assigned to them (**M-step**)
  4. Goto step 2 (until no change)

- The K-means algorithm is guaranteed to converge in a finite number of iterations (different initialization $\Rightarrow$ possibly different result)



K-means clustering cont'd



- K-means clustering corresponds to a Gaussian mixture model estimation with EM whenever the covariance matrices of the Gaussian components are set to $\Sigma_j = \sigma^2 I$, for all j and some fixed small σ^2

Hierarchical (bottom-up) clustering

- Hierarchical agglomerative clustering: we sequentially merge the pair of "closest" points/clusters

- The procedure
  1. Find two closest points (clusters) and merge them
  2. Proceed until we have a single cluster (all the points)

- Two prerequisites:
  1. distance measure $d(\mathbf{x}_i, \mathbf{x}_j)$ between two points
  2. distance measure between clusters (cluster linkage)

Hierarchical (bottom-up) clustering

- A *linkage* method: we have to be able to measure distances between clusters of examples C_k and C_l

  a) Single linkage:

$$d_{kl} = \min_{i \in C_k, j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j)$$

  b) Average linkage:

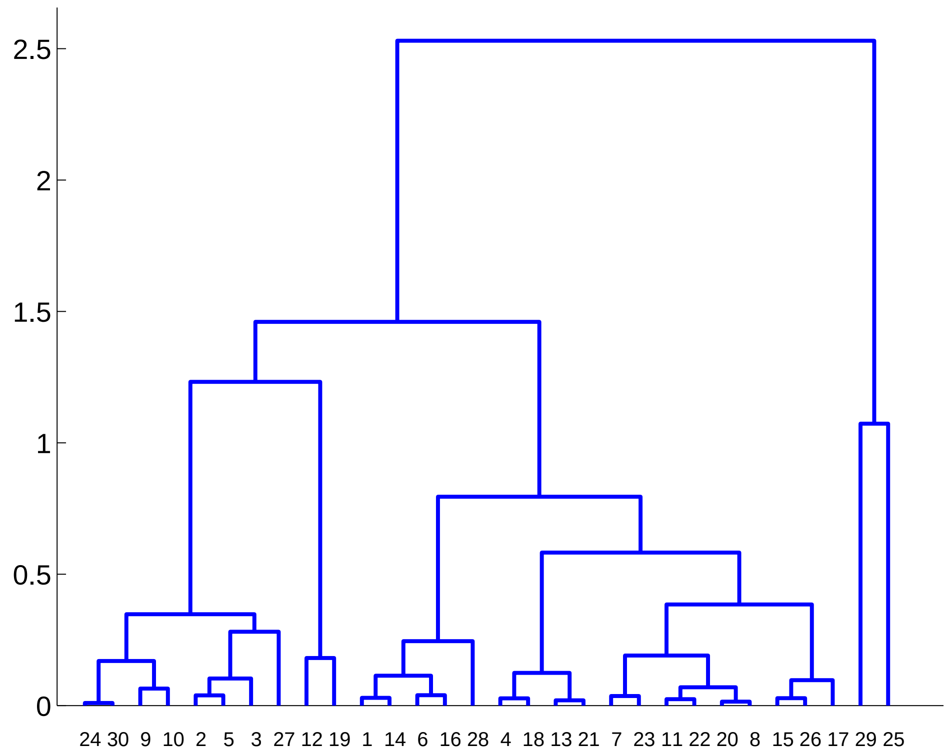
$$d_{kl} = \frac{1}{|C_l|\,|C_k|} \sum_{i \in C_k, j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j)$$

  c) Centroid linkage:

$$d_{kl} = d(\bar{\mathbf{x}}_k, \bar{\mathbf{x}}_l), \quad \bar{\mathbf{x}}_l = \frac{1}{|C_l|} \sum_{i \in C_l} \mathbf{x}_i$$

Hierarchical (bottom-up) clustering

- A dendrogram representation of hierarchical clustering



(the linear ordering of examples is chosen for clarity of representation)

Semi-supervised clustering

- Let's assume we have identified the *relevant* information for clustering the examples

  For each $i = 1, \ldots, n$:

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

- We wish to cluster the examples into larger groups without loosing the relevance information (in this case word frequency)
- Documents with similar word frequencies should be put into the same cluster

Semi-supervised clustering cont'd

- We derive a metric for clusters of examples (documents) $\{\mathbf{x}_i\}$ based on the relevance information $\{P(y|\mathbf{x}_i)\}$ (word frequency)

  The word frequencies for a cluster C come from randomly picking a document in the cluster

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

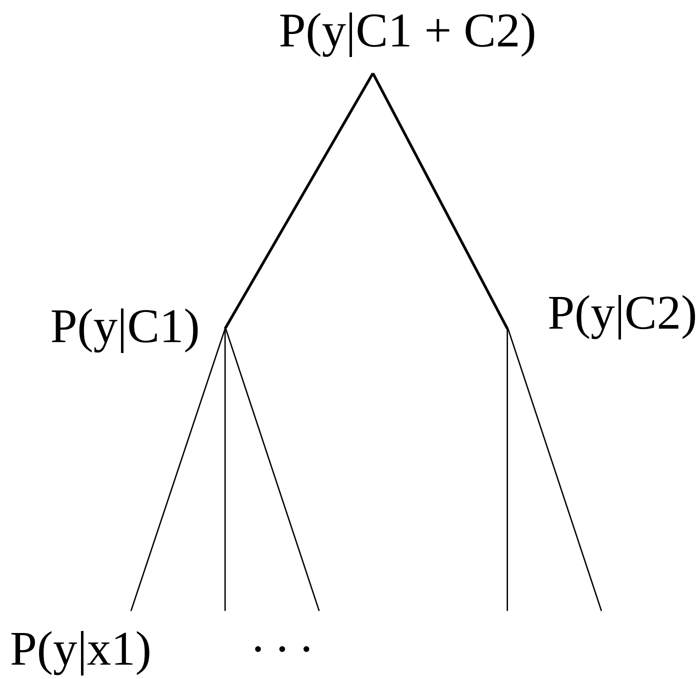
  The distance between the clusters measures how much information we loose about the words if the clusters are merged

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

Semi-supervised clustering cont'd

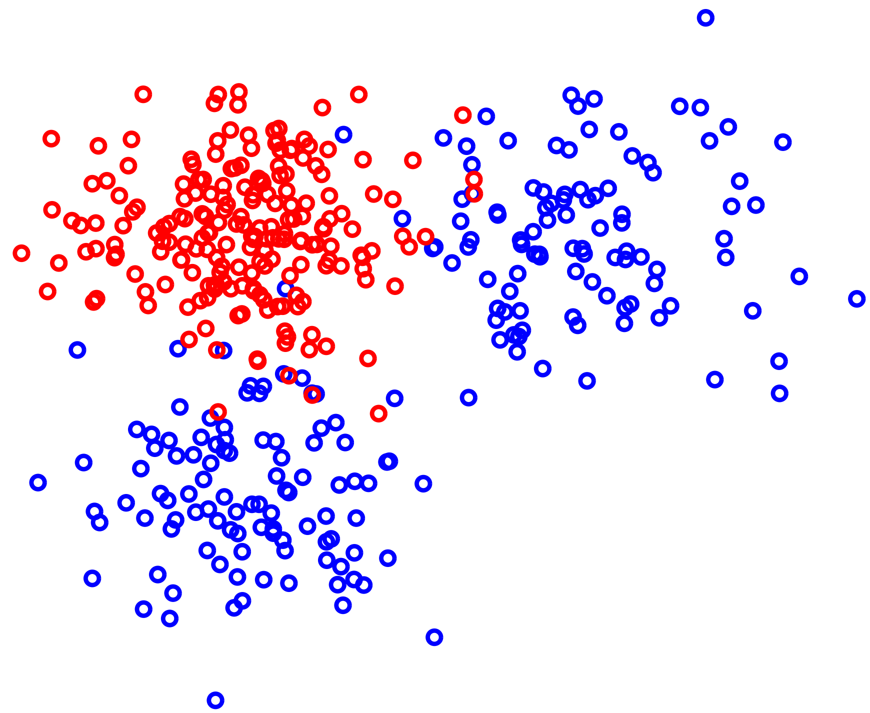
The distance between the clusters measures how much information we loose about the words if the clusters are merged

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



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

- Markov models
  - motivation, definition
  - prediction, estimation

Markov models

- Often we want to capture or model dynamical systems, not just static distributions

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

- Uncertainty captured by a probabilistic dynamical system
- 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:**

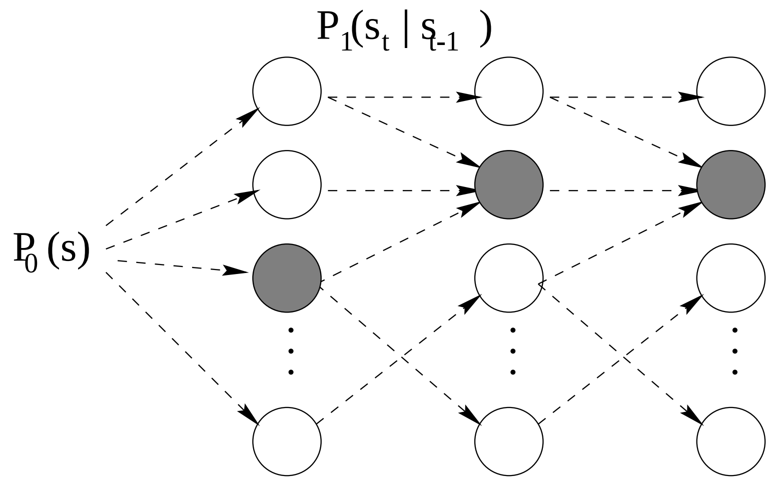
$$P(s_{t+1}|s_t, \underbrace{s_{t-1}, \ldots, s_0}_{\text{past}}) = P_1(s_{t+1}|s_t)$$

- The resulting stochastic system generates a sequence of states

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

where s_0 is drawn from $P_0(s_0)$ and s_{t+1} from $P_1(s_{t+1}|s_t)$ for all t

Markov chain: state diagram



- The initial state s_0 is drawn form $P_0(s_0)$.
- There are a set of possible transtions from each state. These are marked with dashed arrows and correspond to transitions for which $P_1(s'|s_t) > 0$.
- Given s_{t-1} we draw a new state s_t from $P_1(s_t|s_{t-1})$

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

- This is a *homogeneous Markov chain* where the transition probability does not change with time t

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

$$s_0 \longrightarrow s_1 \longrightarrow s_2 \longrightarrow s_3 \longrightarrow \ldots$$

  might be

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

- Is a Markov chain an appropriate model in this context?