Machine Learning

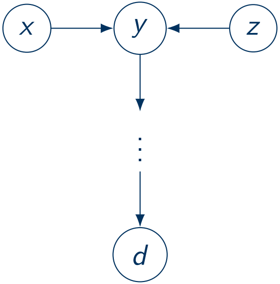
Lecture 26: Statistical dependencies 2

Hao Tang

November 17, 2022

Descendants of v-structure

v-structure



The variables x and z are not independent if any descendant of y is given.

Separation

- The basic structure $x \rightarrow y \rightarrow z$ is blocked given y.
- The basic structure $x \leftarrow y \rightarrow z$ is blocked given y.
- The basic structure $x \rightarrow y \leftarrow z$ is blocked if y and its descendants are not given.
- A path is blocked if any basic structure along the path is blocked.
- Two variables are separated if all paths connecting the two variables are blocked.
- Two sets of variables X and Y are independent given a third set Z if all pairs of in $X \times Y$ are separated given Z.

Independencies in the two objects

- Separation on graph implies independence in the distribution that factorizes according to the graph.

- Technically, separation does not necessarily include all independencies in the distribution that factorizes according to the graph.

Naive Bayes

- Our task is to predict y given d features $x[1], x[2], \ldots, x[d]$.
- Suppose $x[1], \ldots, x[d]$ are mutually independent given y.

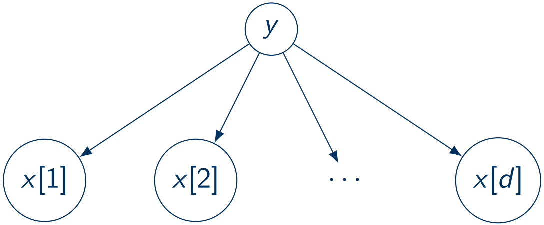
$$p(x[1], x[2], \ldots, x[d], y) = p(y)p(x[1], \ldots, x[d]|y) = p(y)\prod_{i=1}^{d} p(x[i]|y) \quad (1)$$

- The conditional probability

$$p(y|x[1], \ldots, x[d]) = \frac{p(x[1], \ldots, x[d], y)}{p(x[1], \ldots, x[d])} = \frac{p(x[1], \ldots, x[d], y)}{\sum_{y'} p(x[1], \ldots, x[d], y')} \quad (2)$$

$$= \frac{p(y)\prod_{i=1}^{d} p(x[i]|y)}{\sum_{y'} p(y')\prod_{i=1}^{d} p(x[i]|y')} \quad (3)$$

Naive Bayes



Naive Bayes

- When we have a data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we train the naive Bayes classifier with the log loss

$$L = -\log \prod_{i=1}^{n} p(y_i | x_i[1], x_i[2], \ldots, x_i[d]). \tag{4}$$

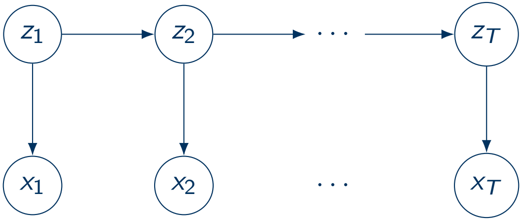
Hidden Markov Models

- For a sequence of observation $x_1, x_2, \ldots, x_T$, we assume there is a hidden sequence $z_1, z_2, \ldots, z_T$.

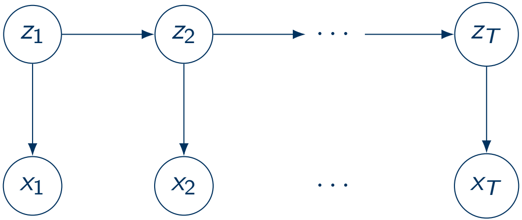
- The first assumption is that x_t is independent of everything else given z_t.

- The second assumption is that z_t is independent of $z_1, z_2, \ldots, z_{t-2}$ given z_t.

Hidden Markov Models



Hidden Markov Models



$$p(x_1, \ldots, x_T, z_1, \ldots, z_T) = p(z_1)p(x_1|z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})p(x_t|z_t) \tag{5}$$

An undirected graph representation

- Each vertex is a variable.
- Each edge signals a dependency.
- The graph is undirected.

Separation on an undirected graph

- There are no child-parent relationships.
- A path is blocked if any vertex on the path is given.
- Two variables are separated if all paths between the two variables are blocked.
- Two sets of variables X and Y are independent given a third set Z if X and Y are separated given Z.

Factorization

- A distribution is said to factorize according to an undirected graph if

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i=1}^{K} \phi_i(C_i), \tag{6}$$

where

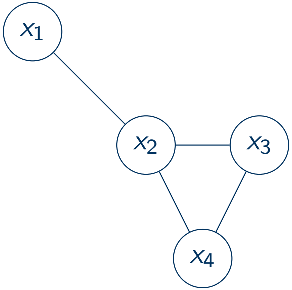
$$Z = \sum_{x_1,\ldots,x_n} \prod_{i=1}^{K} \phi_i(C_i). \tag{7}$$

- The value Z is called the partition function.
- Note that Z does not depend on any assignment of $x_1, \ldots, x_n$.

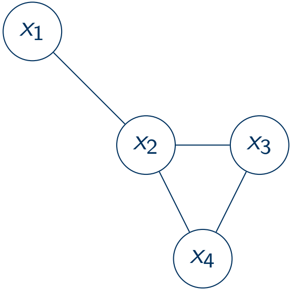
Factorization

- The set $C_i \subseteq \{x_1, \ldots, x_n\}$ is a maximal clique.
- A clique is a set of fully connected vertices.
- A clique is maximal if we cannot include another vertex to make a new clique.
- The function $\phi_i : \mathcal{C}_i \to \mathbb{R}$ is called a factor, where $\mathcal{C}_i$ is all the possible values that can be assigned to C_i.

Maximal clique



Maximal clique



$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4) \tag{8}$$

Independencies in the two objects

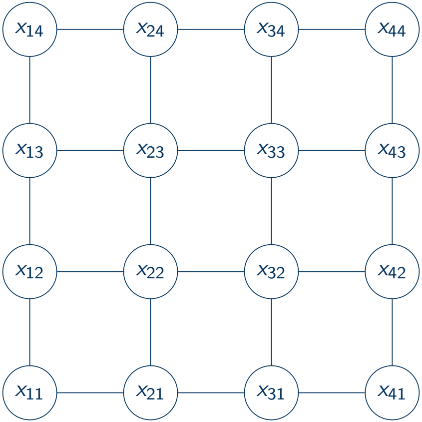
- Similar to the directed case, separation on undirected graph implies independence in the distribution that factorizes according to the graph.

- Technically, separation does not necessarily include all independencies in the distribution that factorizes according to the graph.

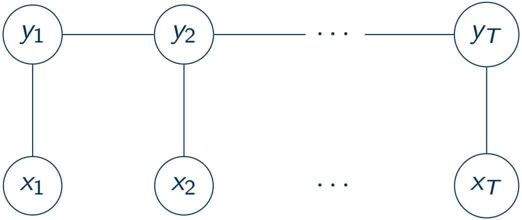
Names

- A directed graph and a distribution that factorizes according to the graph is called a Bayesian network.
- An undirected graph and a distribution that factorizes according to the graph is called a Markov random field.
- An undirected graph and a distribution that factorizes according to the graph is typically called a Markov random field (MRF) when modeling joint distributions, but is typically called a conditional random field (CRF) when modeling conditional distributions.

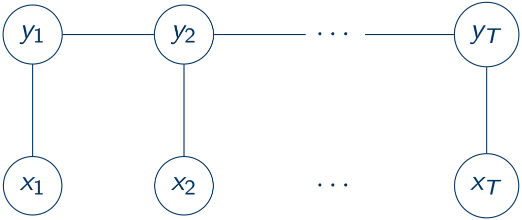
Ising model



Linear-chain conditional random field



Linear-chain conditional random field



$$p(y_1, \ldots, y_T | x_1, \ldots, x_T) = \frac{1}{Z(x_1, \ldots, x_T)} \phi(x_1, y_1) \prod_{t=2}^{T} \phi(y_{t-1}, y_t) \phi(x_t, y_t) \tag{9}$$

Independencies to factorization

- If a distribution matches all the independencies on a directed graph, then the distribution factorizes according to the graph.
- (Hammersley–Clifford) If a distribution matches all the independencies on an undirected graph and the distribution is strictly positive, then the distribution factorizes according to the graph.