Machine learning

Graphical models recall

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

Prof. Giorgio Gambosi

a.a. 2021-2022

Conditional independence

- Two events α, β are (conditionally) independent, given a third event γ iff

$$p(\alpha, \beta|\gamma) = p(\alpha|\gamma)p(\beta|\gamma)$$

- Two random variables X, Y are (conditionally) independent, given a third random variable Z (denoted ad $X \perp\!\!\!\perp Y \mid Z$) iff for all $x \in V(X)$, $y \in V(Y)$, $z \in V(Z)$

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z)p(Y = y|Z = z)$$

- If $p(Y = y|Z = z) > 0$, this is equivalent to

$$p(X = x|Z = z, Y = y) = p(X = x|Z = z)$$

The same holds for sets of random variables $\mathbf{X}, \mathbf{Y}, \boldsymbol{Z}$:

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \boldsymbol{Z} \implies x_i \perp\!\!\!\perp y_j \mid z_1, \dots, z_k$$

for all $x_i \in X, y_j \in Y$, where $\boldsymbol{Z} = \{z_1, \dots, z_k\}$

If $\boldsymbol{Z} = \emptyset$, we have the usual independence $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \emptyset = \mathbf{X} \perp\!\!\!\perp \mathbf{Y}$

Conditional independence: properties

- $X \perp\!\!\!\perp Y \mid Z \implies p(X, Y|Z) = p(X|Z)p(Y|Z)$
- $X \perp\!\!\!\perp Y \mid Z \implies p(X|Y, Z) = p(X|Z)$, assuming $p(Y|Z) > 0$
- $X \perp\!\!\!\perp Y \mid Z \implies Y \perp\!\!\!\perp X \mid Z$, hence $p(Y|X, Z) = p(Y|Z)$ (symmetry)
- $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z$ (decomposition)
- $X \perp\!\!\!\perp Y, W \mid Z \implies X \perp\!\!\!\perp Y \mid Z, W$ (weak union)
- $(X \perp\!\!\!\perp W \mid Y, Z) \wedge (X \perp\!\!\!\perp Y \mid Z) \implies X \perp\!\!\!\perp Y, W \mid Z$ (contraction)

Conditional independence: tasks

Assume a set $\mathbf{X}$ of random variables is defined.

Probability queries

Given two subsets of random variables $\mathbf{Y} \subseteq \mathbf{X}$, $\boldsymbol{Z} \subseteq \mathbf{X}$ and a value assignment z on $\boldsymbol{Z}$, compute $p(\mathbf{Y}|\boldsymbol{Z} = z)$: the probability distribution of $\mathbf{Y}$, conditioned on $\boldsymbol{Z} = z$

MAP queries

Given a subset of random variables $\boldsymbol{Z} \subset \mathbf{X}$ and a value assignment z on $\boldsymbol{Z}$, compute the assignment on all the remaining random variables $\mathbf{W} = \mathbf{X} - \boldsymbol{Z}$ which maximizes $p(\mathbf{W}|\boldsymbol{Z} = z)$

Marginal MAP queries

Given two subsets of random variables $\boldsymbol{Z}, \mathbf{Y} \subset \mathbf{X}$ and a value assignment z on $\boldsymbol{Z}$, compute the assignment on random variables $\mathbf{Y}$ which maximizes $p(\mathbf{Y}|\boldsymbol{Z} = z)$.
If $\mathbf{W} = \mathbf{X} - \mathbf{Y} - \boldsymbol{Z}$ is the subset of the remaining random variables, the marginal MAP corresponds to maximizing $\sum_{\mathbf{W}} P(\mathbf{Y}, \mathbf{W}|\boldsymbol{Z} = z)$

Conditional independence

- In general, the chain rule holds

$$p(X_1, \dots, X_n) = p(X_1)p(X_2|X_1)\dots p(X_n|X_1, \dots, X_{n-1})$$

In the binary case (that is $|V(X_i)| = 2$): $2^0 + 2^1 + \dots + 2^{n-1} = 2^n - 1$ parameters to be specified:

1. $p(X_1 = 1)$
2. $p(X_2 = 1|X_1 = 0)$, $p(X_2 = 1|X_1 = 1)$
3. $p(X_3 = 1|X_1 = 0, X_2 = 0)$, $p(X_3 = 1|X_1 = 1, X_2 = 0)$, , $p(X_3 = 1|X_1 = 0, X_2 = 01)$, $p(X_3 = 1|X_1 = 1, X_2 = 1)$
4. $\cdots$

Conditional independence

- Assume $X_3 \perp\!\!\!\perp X_1 \mid X_2$. Then,

$$p(X_1, X_2, X_3) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2) = p(X_1)p(X_2|X_1)p(X_3|X_2)$$

- In general, assume $X_{i+1} \dots X_n \perp\!\!\!\perp X_1 \dots X_{i-1} \mid X_i$ for all i. Then,

$$p(X_1, \dots, X_n) = p(X_1)p(X_2|X_1)p(X_3|X_2) \dots p(X_n|X_{n-1})$$

  Binary case: $1 + 2 + \dots + 2 = 2n - 1 << 2^{n-1}$ parameters

Properties

- Simmetry

$$X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$$

- Decomposition

$$X \perp\!\!\!\perp Y, Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$$

- Contraction

$$(X \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp W \mid Y, Z) \Rightarrow X \perp\!\!\!\perp Y, W \mid Z$$

- Weak union

$$X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z, W$$

- Intersection (if probability not null)

$$(X \perp\!\!\!\perp Y \mid Z, W) \wedge (X \perp\!\!\!\perp W \mid Y, Z) \Rightarrow X \perp\!\!\!\perp Y, W \mid Z$$

Topics

Rappresentation Specify distributions satisfying predefined independence properties

Inference Exploit independence properties for efficient computations

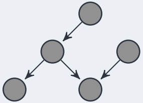
Learning Identify independence properties from data

Graphical models

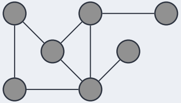
Formal description, by means of graph structures, of properties of families of distributions

Different types of graphical models:

- Bayesian networks (directed graphs)



- Markov random fields (undirected graphs)



Directed graphical models

- Also known as Bayesian networks: family of probability distributions that admit a compact parametrization that can be naturally described using a directed graph.
- General idea: by the chain rule, we can write any probability p as:

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1}, \dots, x_2, x_1)$$

- In a compact Bayesian network each factor on the right hand side depends only on a small number of ancestor variables x_{A_i}:

$$p(x_i \mid x_{i-1}, \dots, x_1) = p(x_i \mid x_{A_i})$$

Directed graphical models in the discrete case

In this framework, we may describe the factors $p(x_i \mid x_{A_i})$ as probability tables, where:

- rows correspond to assignments to x_{A_i}
- columns correspond to values of x_i
- entries contain the actual probabilities $p(x_i \mid x_{A_i})$

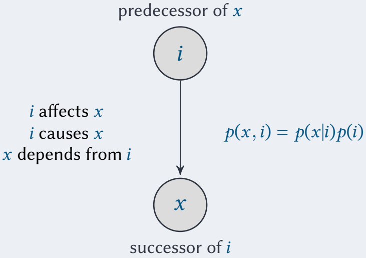
If each variable takes d values and has at most k ancestors, the table will contain at most $O(d^{k+1})$ entries. Since we have one table per variable, the entire probability distribution can be compactly described with only $O(nd^{k+1})$ parameters ($O(d^n)$ with a naive approach).

Directed graphical models in the discrete case

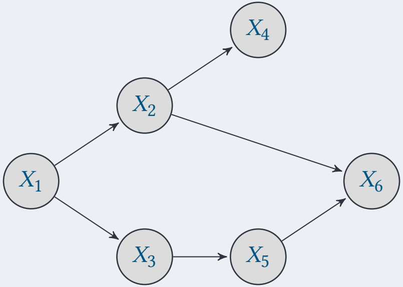
A bayesian network is modeled as a DAG (Directed Acyclic Graph) where

- nodes correspond to random variables
- arcs specify dependencies between variables: the variable corresponding to the destination node is not independent from the one associated to the source

The set of dependencies specified by the arcs of the graph make it possible to factor the joint distribution in a concise way.



Example



$A_2 = A_3 = \{1\}$, $A_4 = \{2\}$, $A_5 = \{3\}$ e $A_6 = \{2, 5\}$.

Observe that if we assume all variables are binary, $2^6 = 64$ probability values (63 indeed, since they must sum to 1) are necessary to describe the joint distribution $p(X_1, \dots, X_6)$

Predecessors and dependencies

Let us define for any r.v. X_i

- π_i the set of immediate predecessors: $X_j \in \pi_i$ iff there exist an arc $\langle X_j, X_i \rangle$
- ν_i the set of predecessors: $X_j \in \nu_i$ iff there exist a directed path $\langle X_j, \dots, X_i \rangle$

The set π_i specify the set of random variables from which X_i directly depends. That is, $p(X_i|\nu_i) = p(X_i|\pi_i)$ or equivalently $X_i \perp\!\!\!\perp \nu_i \mid \pi_i$.

As a consequence,

$$p(\mathbf{X}) = p(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} p(X_i|\pi_i)$$

Example factorization

The joint distribution $p(X_1, X_2, X_3, X_4, X_5, X_6)$ in the example can be factorized by the conditional distributions $p(X_i|\pi_i), i = 1, \dots, 6$ since

$$\begin{aligned}
p(X_1, \dots, X_6) &= p(X_1, \dots, X_5) p(X_6|X_1, \dots, X_5) = p(X_1, \dots, X_5) p(X_6|X_2, X_5) \\
p(X_1, \dots, X_5) &= p(X_1, \dots, X_4) p(X_5|X_1, \dots, X_4) = p(X_1, \dots, X_4) p(X_5|X_3) \\
p(X_1, \dots, X_4) &= p(X_1, \dots, X_3) p(X_4|X_1, \dots, X_3) = p(X_1, \dots, X_3) p(X_4|X_2) \\
p(X_1, \dots, X_3) &= p(X_1, X_2) p(X_3|X_1, X_2) = p(X_1, X_2) p(X_3|X_1) \\
p(X_1, X_2) &= p(X_1) p(X_2|X_1)
\end{aligned}$$

The resulting factorization is

$$p(X_1, \dots, X_6) = p(X_1) p(X_2|X_1) p(X_3|X_1) p(X_4|X_2) p(X_5|X_3) p(X_6|X_2, X_5)$$

Example factorization

The resulting factorization

$$p(X_1, \dots, X_6) = p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2)p(X_5|X_3)p(X_6|X_2, X_5)$$

make it possible to define the joint distribution by means of the conditional ones, where

- $p(X_1)$ can be described by means of 1 value, for example $p(X_1 = 0)$
- $p(X_2|X_1)$, $p(X_3|X_1,)p(X_4|X_2)$, $p(X_5|X_3)$ can be described by means of 2 values each, for example $p(X_2 = 0|X_1 = 0)$ and $p(X_2 = 0|X_1 = 1)$
- $p(X_6|X_2, X_5)$ can be described by means of 4 values, for example $p(X_6 = 0|X_2 = 0, X_5 = 0)$, $p(X_6 = 0|X_2 = 0, X_5 = 1)$, $p(X_6 = 0|X_2 = 1, X_5 = 0)$, $p(X_6 = 0|X_2 = 1, X_5 = 1)$

Hence, the factorized distribution can be described by means of $1 + 4 * 2 + 4 = 13$ values

Probability evaluation

Given $p(\mathbf{X})$, it is possible to derive any (marginal or conditional distribution) defined on $\mathbf{X}$, that is any distribution $p(\mathbf{Y}|\boldsymbol{Z} = \mathbf{z})$ with $\mathbf{Y} \subseteq \mathbf{X}$, $\boldsymbol{Z} \subseteq \mathbf{X} - \mathbf{Y}$, and $\mathbf{z}$ set of values assigned to variables in $\boldsymbol{Z}$. This can be done as follows:

- select all assignments to the set of variables $\mathbf{X}$ where $\boldsymbol{Z} = \mathbf{z}$, $(\mathbf{X} - \boldsymbol{Z}, \boldsymbol{Z} = \mathbf{z})$
- normalize the corresponding probabilities to sum 1
- for any assignment $\mathbf{y}$ on $\mathbf{Y}$, sum the probabilities of all assignments $(\mathbf{X} - \mathbf{Y} - \boldsymbol{Z}, \mathbf{Y} = \mathbf{y}, \boldsymbol{Z} = \mathbf{z})$

Probability evaluation example

Assume we are interested, in the above example, to $p(X_5|X_1 = 0, X_2 = 1)$. This can be done by

- considering the (8) probabilities $p(X_1 = 0, X_2 = 1, X_3 = x_3, X_4 = x_4, X_5 = x_5)$ with $x_3, x_4, x_5 \in \{0, 1\}$
- normalize them dividing each of them by

$$\sum_{x_3 \in \{0,1\}} \sum_{x_4 \in \{0,1\}} \sum_{x_5 \in \{0,1\}} p(X_1 = 0, X_2 = 1, X_3 = x_3, X_4 = x_4, X_5 = x_5)$$

  let $n(X_1 = 0, X_2 = 1, X_3 = x_3, X_4 = x_4, X_5 = x_5)$ be the normalized values
- to compute $p(X_5 = 0|X_1 = 0, X_2 = 1)$ sum the set of normalized values where $X_5 = 0$, that is

$$\sum_{x_3 \in \{0,1\}} \sum_{x_4 \in \{0,1\}} p(X_1 = 0, X_2 = 1, X_3 = x_3, X_4 = x_4, X_5 = 0)$$

Probability evaluation example

The information provided by the Bayesian network make it possible to simplify such computation, by observing that

- X_5 is not affected by the value of X_2
- X_5 is indirectly affected by the value of X_1 through X_3
- this implies that

$$p(X_1 = 0, X_2 = 1, X_3 = x_3, X_4 = x_4, X_5) = p(X_5|X_3)p(X_3|X_1 = 0)$$

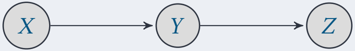
As a consequence, we have

$$p(X_5 = 0|X_1 = 0) = p(X_5 = 0|X_3 = 0)p(X_3 = 0|X_1 = 0) + p(X_5 = 0|X_3 = 1)p(X_3 = 1|X_1 = 0)$$

Independence relations and structures in BN

Cascade

In general, a variable is independent from all predecessors, given the immediate ones.



If Y is observed, then, again $X \perp\!\!\!\perp Z \mid Y$. However, if Y is unobserved, then $X \not\!\perp\!\!\!\perp Z$. Here, the intuition is again that Y holds all the information that determines the outcome of Z; thus, it does not matter what value X takes.

$$Z \perp\!\!\!\perp X \mid Y \qquad \text{or} \qquad p(Z|X,Y) = p(Z|Y) \qquad \text{that is} \qquad p(X,Y,Z) = p(Z|Y)p(Y|X)p(X)$$

Independence relations and structures in BN

For example, in the BN above,

$$X_6 \perp\!\!\!\perp \{X_1, X_3\} \mid \{X_2, X_5\}$$

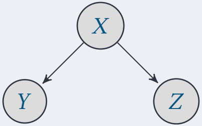
- X: ‘engine breaks”
- Y: “car stops”
- Z: “I shall be late”

Knowing that car stopped increases the probability of being late, independently from knowing the car stopped by an engine break

Independence relations and structures in BN

Common cause

Knowing the cause makes the knowledge of any other consequence irrelevant



If X is observed, then $Y \perp\!\!\!\perp Z \mid X$. However, if X is unobserved, then $Y \not\perp\!\!\!\perp Z$. Intuitively this stems from the fact that Z contains all the information that determines the outcomes of Y and Z; once it is observed, there is nothing else that affects these variables' outcomes.

$Y \perp\!\!\!\perp Z \mid X$ or $p(Y, Z|X) = p(Y|X)p(Z|X)$ that is $p(X, Y, Z) = p(Y|X)p(Z|X)p(X)$

Independence relations and structures in BN

- X: “You passed ML with a 30 score”
- Y: “You meet your friends to celebrate”
- Z: “You ask for a thesis to the professor”

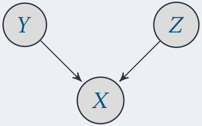
If a know you scored 30 in ML, knowing you went to celebrate does not increase the probability that you may ask for a thesis.

Observe that, if your score is not known, knowing you went to celebrate does increase the probability that you may ask for a thesis

Independence relations and structures in BN

Explaining away

Knowing an effect and one of its possible causes affects the probability of any other possible cause



$$Z \perp\!\!\!\perp Y \quad \text{or} \quad p(Y,Z) = p(Y)p(Z) \quad \text{that is} \quad p(X,Y,Z) = p(X|Y,Z)p(Y)p(Z)$$

$$Z \not\perp\!\!\!\perp Y \mid X \quad \text{or} \quad p(Y,Z|X) \neq p(Y|X)p(Z|X) \quad \text{that is} \quad p(X,Y,Z) \neq p(Y|X)p(Z|X)p(X)$$

Knowing X couples Y and Z. In other words, $Y \perp\!\!\!\perp Z$ if X is unobserved, but $Y \not\perp\!\!\!\perp Z \mid X$ if X is observed.

Independence relations and structures in BN

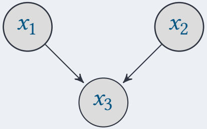
- X: “I have been late”
- Y: “I didn’t have a watch”
- Z: ‘There was traffic”

If I was late, knowing that there was traffic (a possible cause) decreases the probability that it happened because I had no watch (possible alternative cause)

An example

Coin toss

$p(x_1)$: 0.5 0.5



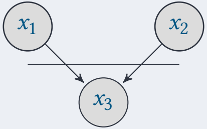
$p(x_2)$: 0.5 0.5

x_3 = same side?

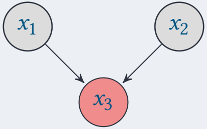
$p(x_3 \vert x_1, x_2)$:		hh	ht	th	tt
	y	1	0	0	1
	n	0	1	1	0

Example

What does the graph structure say?



x_1 e x_2 (marginally) independent



x_1 and x_2 become dependent if x_3 is known

Knowing that coin tosses had the same side makes the knowledge of the side of money 1 affect the knowledge about the side of money 2.

Another example

Student:

Intelligence: $I = \{H, L\}$ (H=high, L=low)

Score: $V = \{h, l\}$ (h=high, l=low)

Joint distribution

I	V	$p(I, V)$
L	l	0.665
L	h	0.035
H	l	0.06
H	L	0.24

Another example

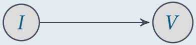
Student:

Intelligence: $I = \{H, L\}$ (H=high, L=low)

Score: $V = \{h, l\}$ (h=high, l=low)

Alternative representation

$$p(I, V) = p(I)p(V|I)$$



I	$p(I)$
L	0.7
H	0.3

$p(V\|I)$	l	h
L	0.95	0.05
H	0.2	0.8

Another example cont.d

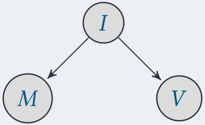
Student:

Intelligence: $I = \{H, L\}$ (H=high, L=low)

Score: $V = \{h, l\}$ (h=high, l=low)

Mean: $M = \{A, B, C\}$

Hypothesis: $V \perp\!\!\!\perp M \mid I$



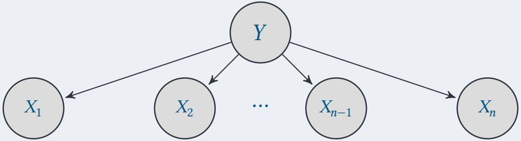
$$p(I, V, M) = p(V, M|I)p(I) = p(V|I)p(M|I)P(I)$$

$p(M\|I)$	A	B	C
L	0.2	0.34	0.46
H	0.74	0.17	0.09

$$p(H, h, B) = 0.3 \cdot 0.8 \cdot 0.17 = 0.0408$$

Naive Bayes

- Class variable Y
- Feature variables $X_1, \dots, X_n$
- Hypothesis: $X_A \perp\!\!\!\perp X_B \mid Y$ for all subsets X_A, X_B of $\{X_1, \dots, X_n\}$



Parameters: $p(Y)$, $p(X_i|Y)$ $i = 1, \dots, n$

Joint distribution

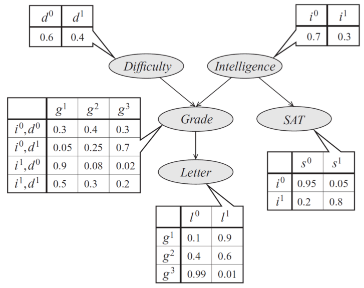
$$p(X_1, \dots, X_n, Y) = p(Y) \prod_{i=1}^{n} p(X_i|Y)$$

A more complex example

Model of a student's grade g on an exam. The grade depends on the exam's difficulty d and the student's intelligence i; it also affects the quality l of the reference letter from the professor who taught the course. The student's intelligence i affects the SAT score s as well. Each variable is binary, except for g, which takes 3 possible values.

The joint probability distribution over the 5 variables naturally factorizes as:

$$p(l, g, i, d, s) = p(l \mid g), p(g \mid i, d), p(i), p(d), p(s \mid i)$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Generative interpretation

- Interpretation in terms of stories for how the data was generated.

Example

To determine the quality of the reference letter

- sample an intelligence level
- sample an exam difficulty

Then,

- a student's grade is sampled given these parameters
- the recommendation letter is generated based on the grade

Formal definition

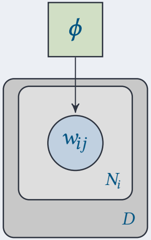
Directed graph $G = (V, E)$ with:

- a random variable x_i for each node $i \in V$
- one conditional distribution $p(x_i \mid x_{A_i})$ per node, specifying the probability of x_i conditioned on its parents' values.

- A Bayesian network defines an overall probability distribution p.
- Conversely, a probability p factorizes over a DAG G if it can be decomposed into a product of factors, as specified by G.

Unigram model

Terms, in all documents, are instances of i.i.d. random variables, distributed according to a same distribution (language model) on the elements of dictionary $\mathbf{V}$: for example a multinomial (with parameter ϕ)



$w_{ij} \sim \text{Mult}(t|\phi)$

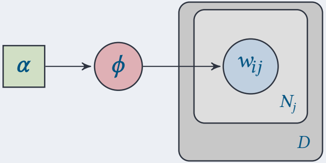
N_i is the length document d_i.

$$p(\mathbf{W}|\phi) = \prod_{j=1}^{D}\prod_{i=1}^{N_i} p(w_{ij}|\phi) = \prod_{k=1}^{V} p(t_k|\phi)^{n_k} = \prod_{k=1}^{V} \text{Mult}(t_k|\phi)^{n_k}$$

where $V = |\mathbf{V}|$ and n_k is the number of occurrences of term $t_k \in \mathbf{V}$ into $\mathbf{W}$

Unigram model

We may also assume a hierarchical model, where the multinomial distribution is instance of another random variable with its own distribution, for example a Dirichlet with parameter $\boldsymbol{\alpha}$



$$\phi \sim \mathrm{Dir}(\phi|\alpha)$$
$$w_{ij} \sim \mathrm{Mult}(w|\phi)$$

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\phi}} \prod_{j=1}^{D} \prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi}) p(\boldsymbol{\phi}|\boldsymbol{\alpha}) d\boldsymbol{\phi} = \int_{\boldsymbol{\phi}} p(\boldsymbol{\phi}|\boldsymbol{\alpha}) \prod_{j=1}^{D} \prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi}) d\boldsymbol{\phi}$$
$$= \int_{\boldsymbol{\phi}} p(\boldsymbol{\phi}|\boldsymbol{\alpha}) \prod_{k=1}^{V} p(t_k|\boldsymbol{\phi})^{n_k} d\boldsymbol{\phi} = \int_{\boldsymbol{\phi}} \mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}) \prod_{k=1}^{V} \mathrm{Mult}(t_k|\boldsymbol{\phi})^{n_k} d\boldsymbol{\phi}$$

Bayesian inference in the bag of words case

By Bayes theorem,

$$p(\phi|\mathbf{W},\alpha) = \frac{p(\mathbf{W}|\phi,\alpha)p(\phi|\alpha)}{p(\mathbf{W}|\alpha)}$$

by the bayesian network structure, we have that $\mathbf{W}$ is conditionally independent from α, given ϕ, that is

$$p(\mathbf{W},\alpha|\phi) = p(\mathbf{W}|\phi)p(\alpha|\phi)$$

then

$$p(\mathbf{W}|\alpha,\phi) = \frac{p(\mathbf{W},\alpha|\phi)}{p(\alpha|\phi)} = \frac{p(\mathbf{W}|\phi)p(\alpha|\phi)}{p(\alpha|\phi)} = p(\mathbf{W}|\phi)$$

and

$$\begin{aligned} p(\phi|\mathbf{W},\alpha) &= \frac{p(\mathbf{W}|\phi)p(\phi|\alpha)}{p(\mathbf{W}|\alpha)} = \frac{\prod_{j=1}^{D}\prod_{i=1}^{N_j} p(w_{ij}|\phi)p(\phi|\alpha)}{p(\mathbf{W}|\alpha)} \\ &= \frac{\prod_{j=1}^{D}\prod_{i=1}^{N_j} \text{Mult}(w_{ij}|\phi)\text{Dir}(\phi|\alpha)}{p(\mathbf{W}|\alpha)} = \frac{\prod_{k=1}^{V} \text{Mult}(t_k|\phi)^{n_k}\text{Dir}(\phi|\alpha)}{p(\mathbf{W}|\alpha)} \end{aligned}$$

Bayesian inference in the bag of words case

Let $Z = p(\mathbf{W}|\boldsymbol{\alpha})$, which is constant with respect to $\boldsymbol{\phi}$, and let $\Delta(\boldsymbol{\alpha})$ be the inverse of the normalization constant of the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})$

$$\Delta(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{V} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{V} \alpha_k)}$$

By the definition of Multinomial and Dirichlet distributions, we then have

$$p(\boldsymbol{\phi}|\mathbf{W}, \boldsymbol{\alpha}) = \frac{1}{Z} \prod_{k=1}^{V} \phi_k^{n_k} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{V} \phi_k^{\alpha_k - 1} = \frac{1}{Z} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{V} \phi_k^{\alpha_k + n_k - 1}$$

which, apart from the normalizing constant, is

$$\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha} + \mathbf{n}) = \frac{1}{\Delta(\boldsymbol{\alpha} + \mathbf{n})} \prod_{k=1}^{V} \phi_k^{\alpha_k + n_k - 1}$$

where $\mathbf{n} = (n_1, \dots, n_V)$.

Bayesian inference in the bag of words case: evidence

Assigning to the normalizing constant Z a value suitable to obtain a probability distribution makes it possible to derive the evidence

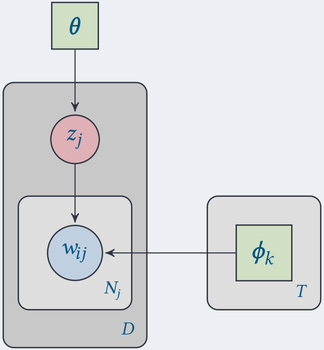
$$Z = p(\mathbf{W}|\boldsymbol{\alpha}) = \frac{\Delta(\boldsymbol{\alpha} + \mathbf{n})}{\Delta(\boldsymbol{\alpha})} = \frac{\Gamma(\sum_{j=1}^{V} \alpha_j) \cdot \prod_{k=1}^{V} \Gamma(\alpha_k + n_k)}{\Gamma(\sum_{j=1}^{V}(\alpha_j + n_j)) \cdot \prod_{k=1}^{V} \Gamma(\alpha_k)}$$

A point estimate for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_V)$ can be derived, in case, by considering the expected value of the posterior distribution

$$\hat{\boldsymbol{\phi}} = E[\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha} + \mathbf{n})] = \frac{\boldsymbol{\alpha} + \mathbf{n}}{\sum_{k=1}^{V}(\alpha_k + n_k)}$$

Mixture of unigrams model

- Latent variable model: one variable z with domain $1, \dots, T$
- For each document, a multinomial ϕ_j (corresponding to a topic) is sampled from a set of T predefined topics
- Each term occurrence in a document is sampled from the multinomial associated to that document



$w_{ij} \sim \text{Mult}(w|\phi)$

Mixture of unigrams model

Distribution

$$
\begin{aligned}
p(\mathbf{W}|\boldsymbol{\theta}, \boldsymbol{\phi}_{1:T}) &= \prod_{j=1}^{D}\left(\sum_{k=1}^{T} p(z_j = k|\boldsymbol{\theta}) \prod_{i=1}^{N_j} p(w_{ij}|z_j = k, \boldsymbol{\phi}_{1:T})\right) \\
&= \prod_{j=1}^{D}\left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{N_j} p(w_{ij}|\boldsymbol{\phi}_k)\right) = \prod_{j=1}^{D}\left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{N_j} \text{Mult}(w_{ij}|\boldsymbol{\phi}_k)\right) \\
&= \prod_{j=1}^{D}\left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{V} \text{Mult}(t_i|\boldsymbol{\phi}_k)^{n_i}\right) = \prod_{j=1}^{D}\left(\sum_{k=1}^{T} \pi_k \prod_{i=1}^{V} \phi_{ki}^{n_i}\right)
\end{aligned}
$$

with $\pi_k = p(z = k|\boldsymbol{\theta})$ and $\phi_{ki} = p(t_i|\boldsymbol{\phi}_k)$

The characterization in terms of mixture can be used also for classification: a mixture component = a class

Mixture of unigrams model

Parameter estimate

The evidence corresponds to the probability distribution for the observed dataset in a mixture model

$$p(\mathbf{X}|\psi) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k q_k(x_i|\theta_k)$$

whose maximal likelihood is evaluated by means of suitable techniques (Expectation Maximization)

In this case, we have that $N = D$ (dataset items correspond to documents) and the class-conditioned distributions $q_k(x_i|\theta_k)$ correspond to the probabilities of a document under the multinomial distribution associated to a topic

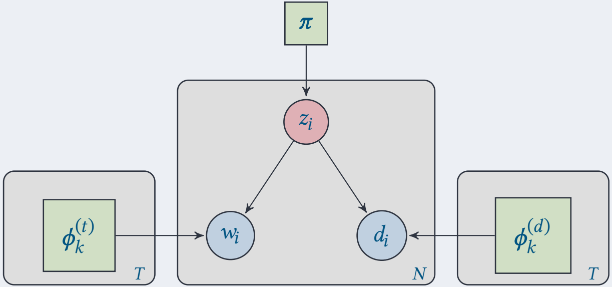
$$q_k(d_i|\boldsymbol{\phi}_k) = \prod_{j=1}^{V} \phi_{kj}^{n_{ij}}$$

where n_{ij} is the number of occurrences of term t_j in document d_i.

Aspect model

The model can be seen as a generative model on data, with the following behaviour:

- For $i = 1$ to N
  - Sample a value (class) z_i with probability $p(z_i)$
  - Sample a term w_i with probability $p(w_i|z_i)$
  - Sample a document d_i with probability $p(d_i|z_i)$



Joint probability

Probability

Joint probability of observations (evidence):

$$p(\mathbf{W}) = \prod_{i=1}^{N} p(w_i, d_i)$$

Joint probability of observations and latent variables:

$$p(\mathbf{W}, \boldsymbol{Z}) = \prod_{i=1}^{N} p(w_i, d_i, z_i) = \prod_{i=1}^{N} p(z_i) p(w_i|z_i) p(d_i|z_i)$$

Mixtures and aspect model

Mixture of distributions for the aspect model

$$p(w_i, d_i) = \sum_{z_i=1}^{T} p(w_i, d_i, z_i) = \sum_{z_i=1}^{T} p(w_i, d_i|z_i)p(z_i) = \sum_{z_i=1}^{T} p(w_i|z_i)p(d_i|z_i)p(z_i)$$

Effect on evidence

$$p(\mathbf{W}) = \prod_{i=1}^{N} \sum_{z_i=1}^{T} p(w_i|z_i)p(d_i|z_i)p(z_i)$$

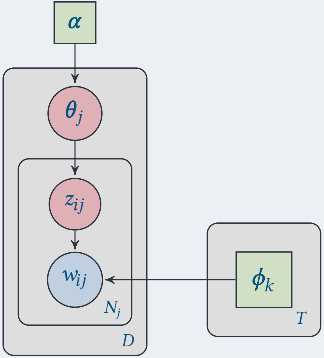
$$= \prod_{w\in\mathbf{V}} \prod_{d\in\mathbf{D}} \left(\sum_{z=1}^{T} p(w|z)p(d|z)p(z) \right)^{n(w,d)}$$

where

$$n(w, d) = |\{(w_i, d_i) \in \mathbf{W} \,:\, w_i = w \wedge d_i = d\}|$$

Latent Dirichlet Allocation

As for PLSA, a document is a mixture of latent classes (topics), defined as multinomial distributions of terms.



$$\theta_j \sim \mathrm{Dir}(\theta|\alpha)$$
$$z_{ij} \sim \mathrm{Mult}(z|\theta)$$
$$w_{ij} \sim \mathrm{Mult}(w|\phi)$$

Latent Dirichlet Allocation

Generative model

The model can be seen as a generative model, working as follows:

- For $j = 1$ to D
  - Sample a vector of classes proportions $\boldsymbol{\theta}_j \sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ for document d_j
  - For $i = 1$ to N_j
    - Sample a class $z_{ij} \sim \mathrm{Mult}(z|\boldsymbol{\theta}_j)$, $z_{ij} \in \{1, \dots, T\}$ for the i-th term in d_j
    - Sample the i-th term $w_{ij} \sim \mathrm{Mult}(w|\boldsymbol{\phi}_{z_{ij}})$ in d_j

Latent Dirichlet Allocation

Parameters

- $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_T)$, parameter of the Dirichlet distributions related to the proportions of classes in a document
- Φ, where $\boldsymbol{\phi}_{ij}$ is the proportion of term $w_j \in V$ in the i-th class

Random variables

A random variable with Dirichlet distribution, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_T)$. Its instances $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D)$ describe proportions of classes for each document d

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{T} \alpha_i)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} = \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1}$$

Latent Dirichlet Allocation

Notes

- The value T (number of classes) is assumed known “a priori”
- The probabilities ϕ_{ki} are unknown but given “a priori” (and to be learned)
- The length $N_i$of a document is assumed known “a priori” and independent with respect to all the other variable or parameters of the model

Document length

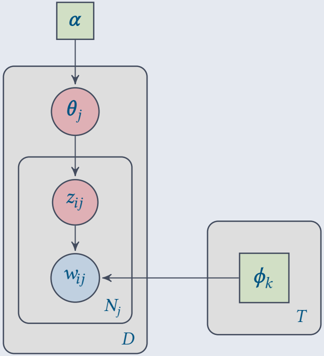
The length of each document can also be modeled as a random variable, distributed according to some predefined distribution. For example,

$$N_i \sim \text{Poisson}(N|\xi)$$

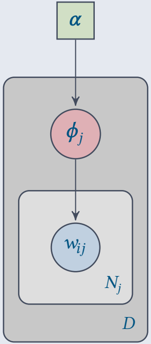
However, learning ξ is assumed to be an independent task with respect to learning other parameters and latent variable values

LDA and clustering

Different generative models



LDA



Clustering