Probabilistic graphical models for Information Retrieval

Guillaume Obozinski

INRIA - Ecole Normale Supérieure - Paris

RussIR 2012

Yaroslavl, 6-10 August 2012

Dealing with structured data

IR is like many other fields confronted with

complex, structured, high-dimensional data.

Dealing with structured data

IR is like many other fields confronted with

complex, structured, high-dimensional data.

To build algorithms, need to:

- extract efficiently the relevant information in the data

Dealing with structured data

IR is like many other fields confronted with

complex, structured, high-dimensional data.

To build algorithms, need to:

- extract efficiently the relevant information in the data

$\rightarrow$ necessary to take into account in a precise and quantitative way the **structure** of the data

Structures in IR

- **Structure of languages:**
  - natural languages: synonymy, polysemy, word structure (stemming), syntax, topics.
  - artificial languages (html, xml)

Structures in IR

- **Structure of languages:**
  - natural languages: synonymy, polysemy, word structure (stemming), syntax, topics.
  - artificial languages (html, xml)
- **Structure of documents:**
  - header, title, dates, fields, picture legends, links, plain text, etc.

Structures in IR

- **Structure of languages:**
  - natural languages: synonymy, polysemy, word structure (stemming), syntax, topics.
  - artificial languages (html, xml)
- **Structure of documents:**
  - header, title, dates, fields, picture legends, links, plain text, etc.
- **Cross-lingual structure:**
  - alignment between sentences in two languages
  - context dependent translation of words

Structures in IR

- **Structure of languages:**
  - natural languages: synonymy, polysemy, word structure (stemming), syntax, topics.
  - artificial languages (html, xml)
- **Structure of documents:**
  - header, title, dates, fields, picture legends, links, plain text, etc.
- **Cross-lingual structure:**
  - alignment between sentences in two languages
  - context dependent translation of words
- **“User structure”:**
  - location, gender, interests, query style from browsing history

Structures in IR

- **Structure of languages:**
  - natural languages: synonymy, polysemy, word structure (stemming), syntax, topics.
  - artificial languages (html, xml)
- **Structure of documents:**
  - header, title, dates, fields, picture legends, links, plain text, etc.
- **Cross-lingual structure:**
  - alignment between sentences in two languages
  - context dependent translation of words
- **"User structure":**
  - location, gender, interests, query style from browsing history
- **Search structure:**
  - browsing behavior/search behavior
    (eg: `clicked on 3rd link` → `came back` → `2nd link` → `came back` → `7th link` → `came back` → `next page` → `new query`)

Goals of the course

- introduce graphical model formalism

Goals of the course

- introduce graphical model formalism
- Present some of the prominent probabilistic models for text corpora
    - Unigram mixture
    - LSI, pLSI
    - Latent Dirichlet Allocation

Goals of the course

- introduce graphical model formalism
- Present some of the prominent probabilistic models for text corpora
  - Unigram mixture
  - LSI, pLSI
  - Latent Dirichlet Allocation
- Derive some algorithms

Goals of the course

- introduce graphical model formalism
- Present some of the prominent probabilistic models for text corpora
  - Unigram mixture
  - LSI, pLSI
  - Latent Dirichlet Allocation
- Derive some algorithms
- Discuss the relevance of these models

Goals of the course

- introduce graphical model formalism
- Present some of the prominent probabilistic models for text corpora
  - Unigram mixture
  - LSI, pLSI
  - Latent Dirichlet Allocation
- Derive some algorithms
- Discuss the relevance of these models
- Some of:
  - Dictionary learning for topic models
  - Graphical models for alignments in translation models
  - Time varying models

A reference

A short paper covering material very similar to the course:

Parameter estimation for text analysis

Gregor Heinrich: *Technical report*

Fraunhofer IGD (2004)

Machine Learning references:

Pattern Recognition and Machine Learning

Chris M. Bishop, Springer Verlag, 2006

Bayesian reasoning and machine learning

David Barber, 2011 - *Cambridge University Press*.

`http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf`

Why probabilistic graphical models ?

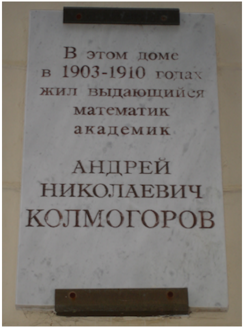
- Why a model?
  - To construct principled algorithm
  - To gain some understanding

Why probabilistic graphical models ?

- Why a model?
  - To construct principled algorithm
  - To gain some understanding
- Why probabilistic?
  - Account for:
    - uncertainty
    - randomness
    - noise
  - Use statistical methodology to estimate/learn a **quantitative model** directly from the data

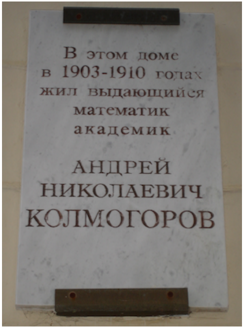
Why probabilistic graphical models ?

- Why a model?
  - To construct principled algorithm
  - To gain some understanding
- Why probabilistic?
  - Account for:
    - uncertainty
    - randomness
    - noise
  - Use statistical methodology to estimate/learn a **quantitative model** directly from the data

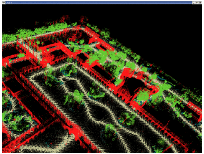


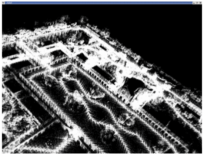
Why probabilistic graphical models ?

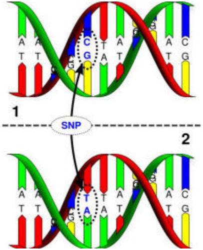
- Why a model?
  - To construct principled algorithm
  - To gain some understanding
- Why probabilistic?
  - Account for:
    - uncertainty
    - randomness
    - noise
  - Use statistical methodology to estimate/learn a **quantitative model** directly from the data
- Why *graphical*?

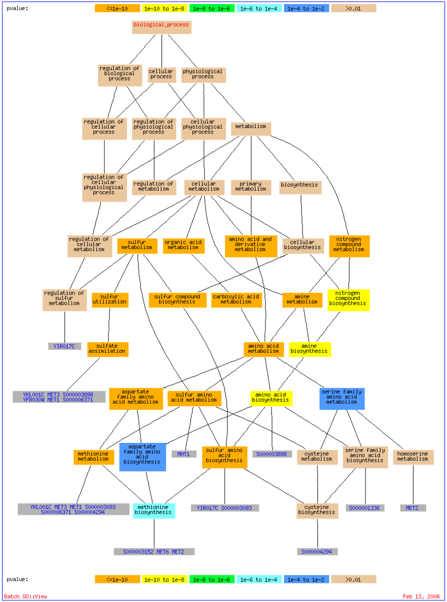


Structured problems in high-dimensions









Curse of dimensionality

Exponential growth of the "volume" with dimension
$\Rightarrow$ the number of parameters grows exponentially.

Curse of dimensionality

Exponential growth of the "volume" with dimension
$\Rightarrow$ the number of parameters grows exponentially.

Example: Histograms

Construct the histogram of $X \in [0,1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct the histogram of $X \in [0,1]^{10}$

$\rightarrow$ size and number of bins ?

$\rightarrow$ a priori impossible with 100 or even 10^6 observations !

Curse of dimensionality

Exponential growth of the "volume" with dimension
$\Rightarrow$ the number of parameters grows exponentially.

Example: Histograms

Construct the histogram of $X \in [0, 1]$ with 10 bins

$\rightarrow$ possible with 100 observations

Construct the histogram of $X \in [0, 1]^{10}$

$\rightarrow$ size and number of bins ?

$\rightarrow$ a priori impossible with 100 or even 10^6 observations !

Model for SNPs

SNP: Single-Nucleotide Polymorphism

- Correspond to 90% of human genetic variations
- Number of loci $k > 10^5$
- Number of configurations $> 2^{10^5}$...

Some concepts from probability and statistics...

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression
- Markov chain

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression
- Markov chain
- Maximum Likelihood estimator

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression
- Markov chain
- Maximum Likelihood estimator
- A priori distribution / a posteriori distribution

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression
- Markov chain
- Maximum Likelihood estimator
- A priori distribution / a posteriori distribution
- Kullback-Leibler divergence

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression
- Markov chain
- Maximum Likelihood estimator
- A priori distribution / a posteriori distribution
- Kullback-Leibler divergence
- Principal Component Analysis

Some concepts from probability and statistics...

- Independence: "X_1 and X_2 are independent random variables"
- Conditional independence: "X_1 and X_2 are independent given Z"
- i.i.d. data
- Linear regression
- Markov chain
- Maximum Likelihood estimator
- A priori distribution / a posteriori distribution
- Kullback-Leibler divergence
- Principal Component Analysis
- Regularization / Ridge regression

Outline

Notations, formulas, definitions

- Joint distribution of X_A et X_B: $p(x_A, x_B)$
- Marginal distribution : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Conditional distribution: $p\left(x_A|x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$

Notations, formulas, definitions

- Joint distribution of X_A et X_B: $p(x_A, x_B)$
- Marginal distribution : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Conditional distribution: $p\left(x_A | x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$

Bayes formula

$$p\left(x_A | x_B\right) = \frac{p\left(x_B | x_A\right) p\left(x_A\right)}{p\left(x_B\right)}$$

Notations, formulas, definitions

- Joint distribution of X_A et X_B: $p(x_A, x_B)$
- Marginal distribution : $p(x_A) = \sum_{x_{A^c}} p(x_A, x_{A^c})$
- Conditional distribution: $p(x_A|x_B) = \frac{p(x_A,x_B)}{p(x_B)}$ si $p(x_B) \neq 0$

Bayes formula

$$p(x_A|x_B) = \frac{p(x_B|x_A)\, p(x_A)}{p(x_B)}$$

$\rightarrow$ Bayes formula is not a "Bayesian formula".

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Covariance :

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Covariance :

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

- Conditional expectation of X given Y :

$$\mathbb{E}[X|Y] = \sum_x x \cdot p(x|y)$$

Expectation and Variances

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Covariance :

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

- Conditional expectation of X given Y :

$$\mathbb{E}[X|Y] = \sum_x x \cdot p(x|y)$$

Entropy and Kullback-Leibler divergence

Entropie

$$H(p) = -\sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

$\rightarrow$ Expectation of the negative log-likelihood

Entropy and Kullback-Leibler divergence

Entropie

$$H(p) = -\sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

$\rightarrow$ Expectation of the negative log-likelihood

Kullback-Leibler divergence

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p\Big[\log \frac{p(X)}{q(X)}\Big]$$

$\rightarrow$ Expectation of the log-likelihood ratio

Entropy and Kullback-Leibler divergence

Entropie

$$H(p) = -\sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

$\rightarrow$ Expectation of the negative log-likelihood

Kullback-Leibler divergence

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p\Big[\log \frac{p(X)}{q(X)}\Big]$$

$\rightarrow$ Expectation of the log-likelihood ratio

$\rightarrow$ Property: $KL(p\|q) \geq 0$

Independence concepts

Independence: $X \perp\!\!\!\perp Y$

We will say that X and Y are independent and write $X \perp\!\!\!\perp Y$ iff:

$$\forall x, y, \qquad P(X = x, Y = y) = P(X = x)\, P(Y = y)$$

Independence concepts

Independence: $X \perp\!\!\!\perp Y$

We will say that X and Y are independent and write $X \perp\!\!\!\perp Y$ iff:

$$\forall x, y, \qquad P(X = x, Y = y) = P(X = x)\, P(Y = y)$$

Conditional Independence: $X \perp\!\!\!\perp Y \mid Z$

- We will say that X and Y are independent conditionally on Z and
- write $X \perp\!\!\!\perp Y \mid Z$ ssi:

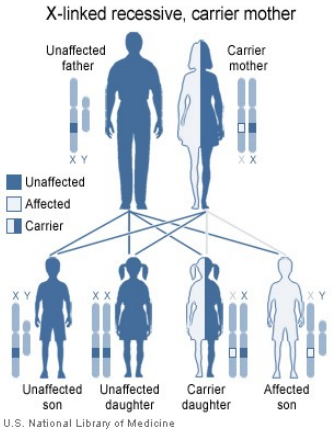
$\forall x, y, z,$

$$P(X = x, Y = y \mid Z = z) = P(X = x|Z = z)\, P(Y = y|Z = z)$$

Conditional Independence exemple

Example of
"X-linked recessive inheritance":

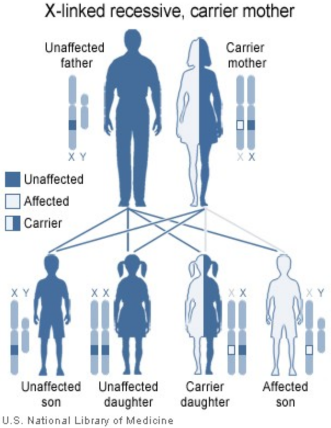
Transmission of the gene
responsible for hemophilia



Conditional Independence exemple

Example of
“X-linked recessive inheritance”:

Transmission of the gene
responsible for hemophilia



Risk for sons from an unaffected father:

- dependance between the situation of the two brothers.
- conditionally independent given that the mother is a carrier of the gene or not.

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \mathrm{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \mathrm{Ber}(\theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Binomial model: $X \sim \mathrm{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

Statistical Model

Parametric model – Definition:

Set of distributions parametrized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$$

Bernoulli model: $X \sim \mathrm{Ber}(\theta) \qquad \Theta = [0,1]$

$$p(x|\theta) = \theta^x (1-\theta)^{(1-x)}$$

Binomial model: $X \sim \mathrm{Bin}(n,\theta) \qquad \Theta = [0,1]$

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

Multinomial model: $X \sim \mathcal{M}(n, \pi_1, \pi_2, \ldots, \pi_K) \qquad \Theta = [0,1]^K$

$$p(x|\theta) = \binom{n}{x_1, \ldots, x_k} {\pi_1}^{x_1} \ \ldots \ {\pi_k}^{x_k}$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}(\mu, \sigma^2)$

X real valued r.v., and $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Gaussian model

Scalar Gaussian model : $X \sim \mathcal{N}(\mu, \sigma^2)$

X real valued r.v., and $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

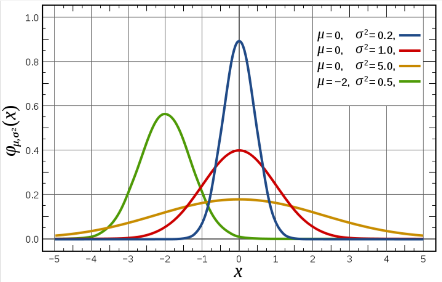
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Multivariate Gaussian model: $X \sim \mathcal{N}(\mu, \Sigma)$

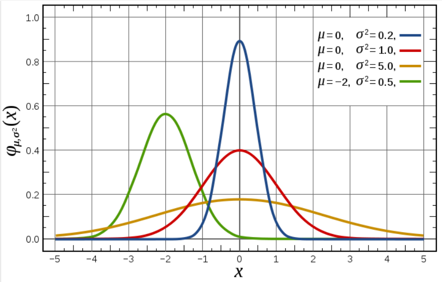
X r.v. taking values in $\mathbb{R}^d$. If $\mathcal{K}_n$ is the set of positive definite matrices of size $n \times n$, and $\theta = (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \mathcal{K}_n$.

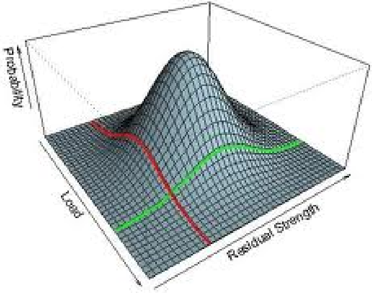
$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Gaussian densities



Gaussian densities





Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P.

Sample/Training set

The data used to learn or estimate a model typically consists of a collection of observation which can be thought of as instantiations of random variables.

$$X^{(1)}, \dots, X^{(n)}$$

A common assumption is that the variables are **i.i.d.**

- **independent**
- **identically distributed**, i.e. have the same distribution P.

This collection of observations is called

- the *sample* or the *observations* in statistics
- the *sample***s** in engineering
- the *training set* in machine learning

Outline

Maximum likelihood Principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Maximum likelihood Principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Maximum likelihood Principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Likelihood:

$$\begin{aligned} \mathcal{L} : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta \in \Theta} p(x|\theta)$$

Sir Ronald Fisher
(1890-1962)

Maximum likelihood Principle

- Let $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$ be a given model
- Let x be an observation

Sir Ronald Fisher (1890-1962)

Likelihood:

$$\begin{aligned}\mathcal{L} : \Theta &\to \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta)\end{aligned}$$

Maximum likelihood estimator:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(x|\theta)$$

Case of i.i.d data

If $(x_i)_{1 \le i \le n}$ is an i.i.d. sample of size n:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log\ p(x_i|\theta)$$

Examples of computation of the MLE

- Bernoulli model
- Multinomial model
- Gaussian model

Outline

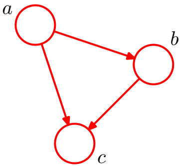
Notations

- $G = (V, E)$ is a graph.
- A random variable X_i is associated to each node $i \in E$.
- We will write its values x_i.
- If $A \subset E$ is a set of nodes, we will write $X_A = (X_i)_{i \in A}$ et $x_A = (x_i)_{i \in A}$.

Oriented graphical model or Bayesian Network

Let G be a *directed acyclic graph (DAG)*. We say that a distribution factorizes according to the graph if it can be written as a product of conditional distributions involving exactly each variable and its parent variables in the graph.

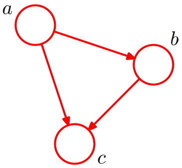
$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$



Oriented graphical model or Bayesian Network

Let G be a *directed acyclic graph (DAG)*. We say that a distribution factorizes according to the graph if it can be written as a product of conditional distributions involving exactly each variable and its parent variables in the graph.

$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$

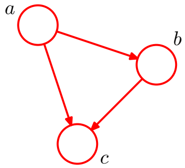


$$p(x_1, x_2) = p(x_1)p(x_2)$$

Oriented graphical model or Bayesian Network

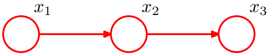
Let G be a *directed acyclic graph (DAG)*. We say that a distribution factorizes according to the graph if it can be written as a product of conditional distributions involving exactly each variable and its parent variables in the graph.

$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$



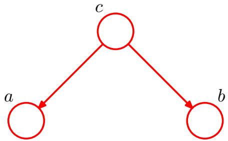
$$p(x_1, x_2) = p(x_1)p(x_2)$$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$$



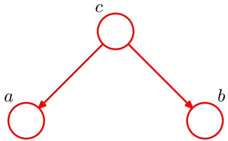
Oriented graphical model or Bayesian Network

$a \perp\!\!\!\perp b \mid c$

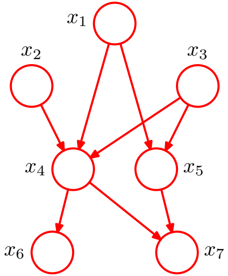


Oriented graphical model or Bayesian Network

$$a \perp\!\!\!\perp b \mid c$$

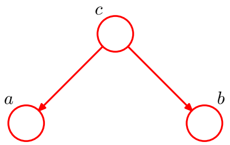


$$\prod_{j=1}^{p} p(x_j | x_{\Pi_j})$$

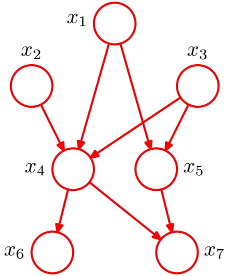


Oriented graphical model or Bayesian Network

$$a \perp\!\!\!\perp b \mid c$$



$$\prod_{j=1}^{p} p(x_j | x_{\Pi_j})$$



$$p(x_1) \prod_{j=2}^{M} p(x_j | x_{j-1})$$

Factorization and Independence

- A factorization induces conditional independence properties

$$\forall x,\ p(x) = \prod_{j=1}^{p} p(x_j|x_{\Pi_j}) \quad \Leftrightarrow \quad \forall j,\ X_j \perp\!\!\!\perp X_{\{1,\ldots,j-1\}\backslash\Pi_j} \mid X_{\Pi_j}$$

- Is it possible to read directly from the graph the conditional independence statements that are true given the factorization?

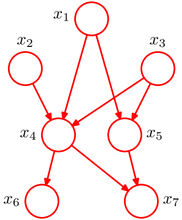
Factorization and Independence

- A factorization induces conditional independence properties

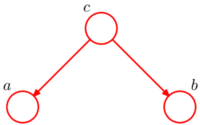
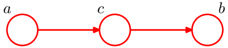
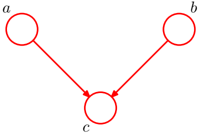
$$\forall x,\ p(x) = \prod_{j=1}^{p} p(x_j | x_{\Pi_j}) \quad \Leftrightarrow \quad \forall j,\ X_j \perp\!\!\!\perp X_{\{1,\dots,j-1\}\setminus\Pi_j} \mid X_{\Pi_j}$$

- Is it possible to read directly from the graph the conditional independence statements that are true given the factorization?

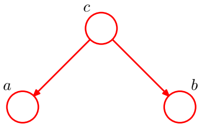
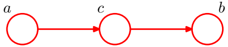
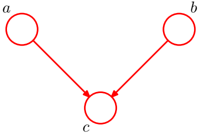
$$X_5 \overset{?}{\perp\!\!\!\perp} X_2 \mid X_4$$



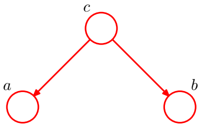
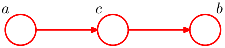
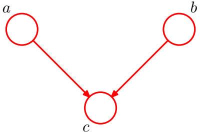
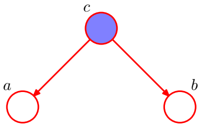
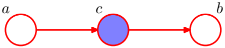
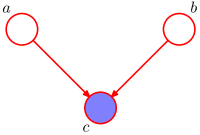
Blocking nodes

diverging edges	consecutive edges	"v"-structure
		

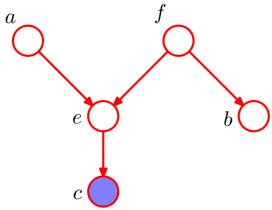
Blocking nodes

diverging edges	consecutive edges	"v"-structure
		
		$\not\leftrightarrow$
$a \not\perp\!\!\!\perp b$	$a \not\perp\!\!\!\perp b$	$a \perp\!\!\!\perp b$

Blocking nodes

diverging edges	consecutive edges	"v"-structure
 $a \not\perp\!\!\!\perp b$	 $a \not\perp\!\!\!\perp b$	 $\not\leftrightarrow$ $a \perp\!\!\!\perp b$
 $\not\leftrightarrow$ $a \perp\!\!\!\perp b \mid c$	 $\not\leftrightarrow$ $a \perp\!\!\!\perp b \mid c$	 $a \not\perp\!\!\!\perp b \mid c$

d-separation



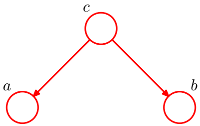
Theorem

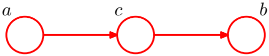
Let A, B and C three disjoint sets of nodes. The property $X_A \perp\!\!\!\perp X_B | X_C$ holds if and only if all paths connecting A to B are *blocked*, which means that they contain at least one blocking node. Node j is a *blocking node*

- if there is no "v-structure" in j and j is in C or
- if there is a "v-structure" in j and if neither j nor any of its descendants in the graph is in C.

Factorization and Independence II

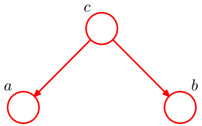
- Several graphs can induce the same set of conditional independence statements.

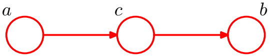




Factorization and Independence II

- Several graphs can induce the same set of conditional independence statements.

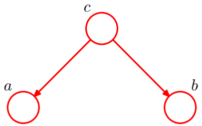


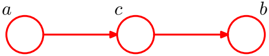


- Some combinations of conditional independence statements cannot be represented by a graphical model.

Factorization and Independence II

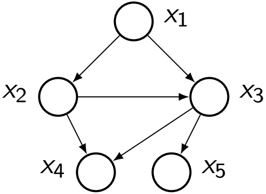
- Several graphs can induce the same set of conditional independence statements.





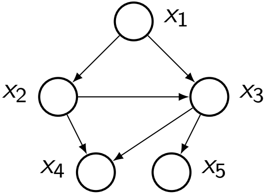
- Some combinations of conditional independence statements cannot be represented by a graphical model.

How to parameterize an Oriented graphical model?



$$p(\mathbf{x};\boldsymbol{\theta}) = p(x_1;\theta_1)\,p(x_2|x_1;\theta_2)\,p(x_3|x_2,x_1;\theta_3)\,p(x_4|x_3,x_2;\theta_4)\,p(x_5|x_3;\theta_5)$$

How to parameterize an Oriented graphical model?

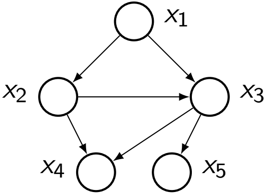


Conditional Probability tables

- $x_1 \in \{0, 1\}$
- $x_2 \in \{0, 1, 2\}$
- $x_3 \in \{0, 1, 2\}$

$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \theta_1)\, p(x_2|x_1; \theta_2)\, p(x_3|x_2, x_1; \theta_3)\, p(x_4|x_3, x_2; \theta_4)\, p(x_5|x_3; \theta_5)$$

How to parameterize an Oriented graphical model?



Conditional Probability tables

- $x_1 \in \{0, 1\}$
- $x_2 \in \{0, 1, 2\}$
- $x_3 \in \{0, 1, 2\}$

		$p(x_3 = k)$		
x_1	x_2	0	1	2
0	0	1	0	0
0	1	1	0	0
0	2	0.1	0	0.9
1	0	1	0	0
1	1	0.5	0.5	0
1	2	0.2	0.3	0.5

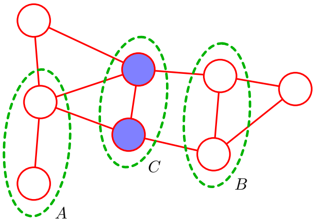
$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \theta_1)\, p(x_2|x_1; \theta_2)\, p(x_3|x_2, x_1; \theta_3)\, p(x_4|x_3, x_2; \theta_4)\, p(x_5|x_3; \theta_5)$$

A Markov random field or *non-oriented graphical model*

Is it possible to define a collection of distributions that somehow factorize according to the graph such that conditional independence coincides exactly with usual separation in the graph, i.e. such that we have the

Global Markov Property

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad \Leftrightarrow C \text{ separates } A \text{ from } B$$



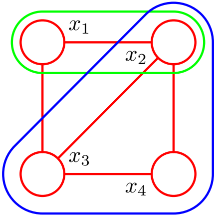
Gibbs distribution

Clique Set of nodes which is fully connected.

Potential The potential $\psi_C(x_C) \geq 0$ is associated to the clique C.

Gibbs distribution

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$$



Partition funtcion

$$Z = \sum_x \prod_C \psi_C(x_C)$$

Potential in exponential form: $\psi_C(x_C) = \exp\{-E(x_C)\}$.
$E(x_C)$ is an *energy term*.
This is then called a *Boltzmann distribution*.

Outline

Bayesian estimation

Bayesians treat the parameter θ as a random variable.

A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters θ, which models his prior belief of the relative plausibility of different values of the parameter.

Bayesian estimation

Bayesians treat the parameter θ as a random variable.

A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters θ, which models his prior belief of the relative plausibility of different values of the parameter.

A posteriori

The observation contribute through the likelihood: $p(x|\theta)$.
The *a posteriori* distribution on the parameters is then

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} \propto p(x|\theta)\,p(\theta)\,.$$

$\rightarrow$ The Bayesian estimator is therefore a probability distribution on the parameters.

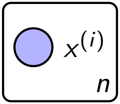
This estimation procedure is called Bayesian inference.

Graphical models associated with an *i.i.d. sample* from $\mathcal{P}_\Theta$

Important to distinguish between:

- constituents of a structured random variable $X = (X_1, \dots, X_d)$ and
- an i.i.d. sample $X^{(1)}, \dots, X^{(n)}$ with $X \sim X^{(i)} = (X_1^{(i)}, \dots, X_d^{(i)})$.

I.i.d. sampling itself corresponds to a graphical model:



Frequentist model

$$\prod_{i=1}^{n} p(x^{(i)}\,;\,\theta)$$

Bayesian formulation

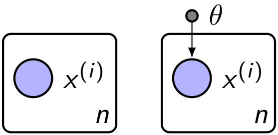
$$p(\theta\,;\,\alpha) \prod_{i=1}^{n} p(x^{(i)}|\theta)$$

Graphical models associated with an *i.i.d. sample* from $\mathcal{P}_\Theta$

Important to distinguish between:

- constituents of a structured random variable $X = (X_1, \ldots, X_d)$ and
- an i.i.d. sample $X^{(1)}, \ldots, X^{(n)}$ with $X \sim X^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$.

I.i.d. sampling itself corresponds to a graphical model:



Frequentist model

$$\prod_{i=1}^{n} p(x^{(i)}\,;\,\theta)$$

Bayesian formulation

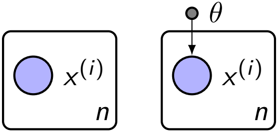
$$p(\theta\,;\,\alpha) \prod_{i=1}^{n} p(x^{(i)}|\theta)$$

Graphical models associated with an *i.i.d. sample* from $\mathcal{P}_\Theta$

Important to distinguish between:

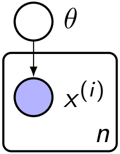
- constituents of a structured random variable $X = (X_1, \dots, X_d)$ and
- an i.i.d. sample $X^{(1)}, \dots, X^{(n)}$ with $X \sim X^{(i)} = (X_1^{(i)}, \dots, X_d^{(i)})$.

I.i.d. sampling itself corresponds to a graphical model:



Frequentist model

$$\prod_{i=1}^{n} p(x^{(i)}\,;\,\theta)$$



Bayesian formulation

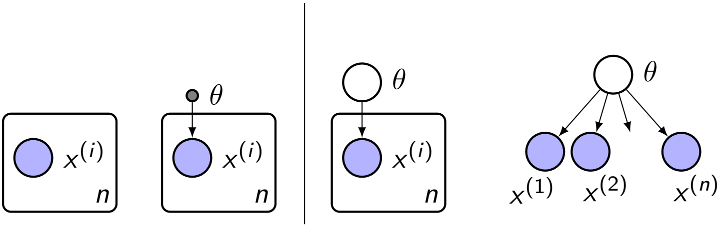
$$p(\theta\,;\,\alpha) \prod_{i=1}^{n} p(x^{(i)}|\theta)$$

Graphical models associated with an *i.i.d. sample* from $\mathcal{P}_\Theta$

Important to distinguish between:

- constituents of a structured random variable $X = (X_1, \ldots, X_d)$ and
- an i.i.d. sample $X^{(1)}, \ldots, X^{(n)}$ with $X \sim X^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$.

I.i.d. sampling itself corresponds to a graphical model:



Frequentist model

$$\prod_{i=1}^{n} p(x^{(i)}\,;\,\theta)$$

Bayesian formulation

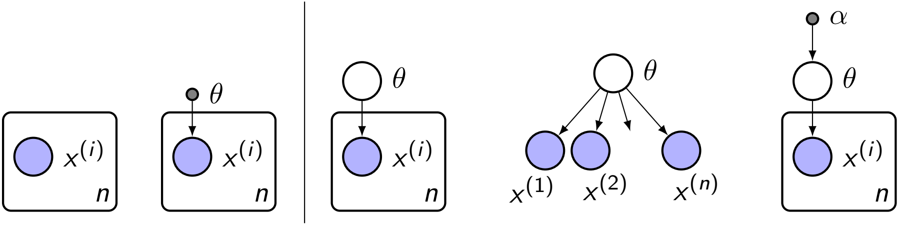
$$p(\theta\,;\,\alpha)\prod_{i=1}^{n} p(x^{(i)}|\theta)$$

Graphical models associated with an *i.i.d. sample* from $\mathcal{P}_\Theta$

Important to distinguish between:

- constituents of a structured random variable $X = (X_1, \ldots, X_d)$ and
- an i.i.d. sample $X^{(1)}, \ldots, X^{(n)}$ with $X \sim X^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$.

I.i.d. sampling itself corresponds to a graphical model:



Frequentist model

$$\prod_{i=1}^{n} p(x^{(i)}\,;\,\theta)$$

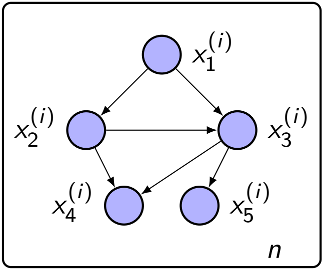
Bayesian formulation

$$p(\theta\,;\,\alpha)\prod_{i=1}^{n} p(x^{(i)}|\theta)$$

Graphical model for an i.i.d. sample II

Exposing the structure

in the frequentist case

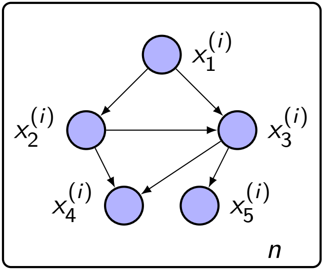


$$p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\prod_{j=1}^{d} p(\, x_j^{(i)} \mid x_{\Pi_j}^{(i)}; \theta_j \,) \right]$$

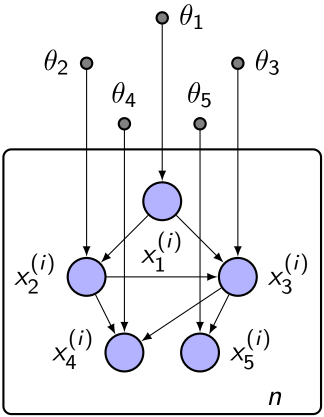
Graphical model for an i.i.d. sample II

Exposing the structure

in the frequentist case



or



$$p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\prod_{j=1}^{d} p(\, x_j^{(i)} \mid x_{\Pi_j}^{(i)}; \theta_j \,) \right]$$

Conjugate priors

A family of prior distribution

$$\mathcal{P}_A = \{p_\alpha(\theta) \mid \alpha \in A\}$$

is said to be **conjugate** to a model $\mathcal{P}_\Theta$, if, for a sample

$$X^{(1)}, \ldots, X^{(n)} \overset{\text{i.i.d.}}{\sim} p_\theta \qquad \text{with} \qquad p_\theta \in \mathcal{P}_\Theta,$$

the distribution q defined by

$$q(\theta) = p(\theta | x^{(1)}, \ldots, x^{(n)}) = \frac{p_\alpha(\theta) \prod_i p_\theta\left(x^{(i)}\right)}{\int p_\alpha(\theta) \prod_i p_\theta\left(x^{(i)}\right) d\theta}$$

is such that

$$q \in \mathcal{P}_A.$$

Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

for

Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\mathbf{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and

Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

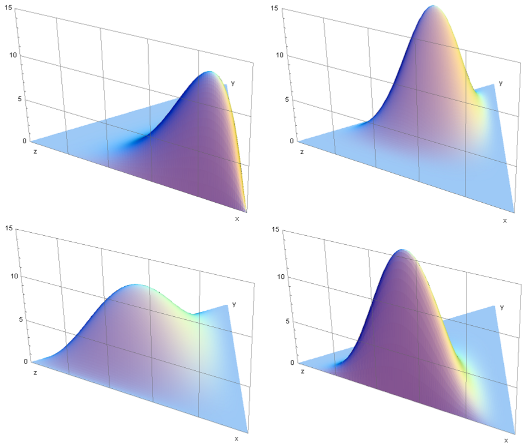
for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\mathbf{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and admitting the density

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_k \Gamma(\alpha_k)} \, \theta_1^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1}$$

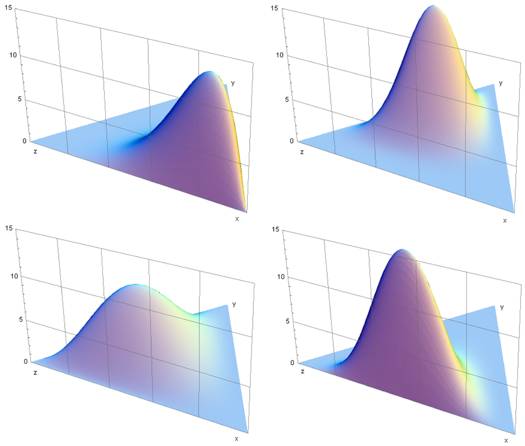
with respect to the uniform measure on the simplex, where

$$\alpha_0 = \sum_k \alpha_k \quad \text{and} \quad \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt$$

Dirichlet distribution II



Dirichlet distribution II

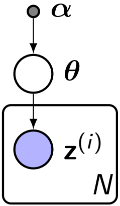


$$\mathbb{E}[\theta_k] = \frac{\alpha_k}{\alpha_0} \quad , \quad \mathrm{Var}(\theta_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0+1)} \quad \text{and} \quad \mathrm{Cov}(\theta_j, \theta_k) = \frac{-\alpha_j \alpha_k}{\alpha_0^2(\alpha_0+1)}$$

with $\alpha_0 = \sum_k \alpha_k$.

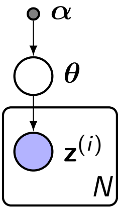
Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with



Bayesian estimation of a multinomial random variable

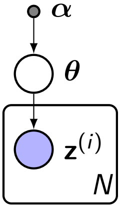
Consider the simple Bayesian Dirichlet-Multinomial model with



- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with

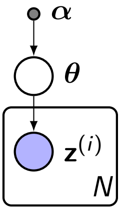


- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with



- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

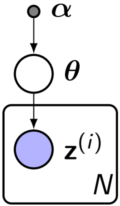
$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) =$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with



- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

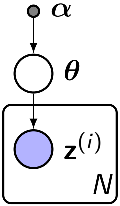
$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})}$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with



- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

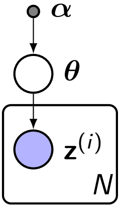
$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})} \propto \prod_k \theta_k^{\alpha_k + \sum_n z_{nk} - 1}$$

Bayesian estimation of a multinomial random variable

Consider the simple Bayesian Dirichlet-Multinomial model with



- A Dirichlet prior on the parameter of the multinomial: $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
- A multinomial random variable $\mathbf{z} \sim \mathcal{M}(1, \boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_k}$$

Let $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ be an i.i.d. sample distributed like $\mathbf{z}$.
We have

$$p(\boldsymbol{\theta}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}) = \frac{p(\boldsymbol{\theta}) \prod_n p(\mathbf{z}^{(n)}|\theta)}{p(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})} \propto \prod_k \theta_k^{\alpha_k + \sum_n z_{nk} - 1}$$

So that $(\theta|(Z)) \sim \mathrm{Dir}((\alpha_1 + N_1, \ldots, \alpha_K + N_K))$ with $N_k = \sum_n z_{nk}$

Laplace smoothing

To obtain *point estimates* in the Bayesian setting, we can compute posterior expectations:

$$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)}].$$

Laplace smoothing

To obtain *point estimates* in the Bayesian setting, we can compute posterior expectations:

$$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)}].$$

We have

$$\boxed{\mathbb{E}[\theta_k|\mathbf{Z}] = \frac{N_k + \alpha_k}{N + \alpha_0} = \frac{N}{N + \alpha_0}\,\frac{N_k}{N} + \frac{\alpha_0}{N + \alpha_0}\,\frac{\alpha_k}{\alpha_0}},$$

with $N_k = \sum_n \mathbf{z}_{nk}$ and $\alpha_0 = \sum_k \alpha_k$.

Laplace smoothing

To obtain *point estimates* in the Bayesian setting, we can compute posterior expectations:

$$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)}].$$

We have

$$\boxed{\mathbb{E}[\theta_k|\mathbf{Z}] = \frac{N_k + \alpha_k}{N + \alpha_0} = \frac{N}{N+\alpha_0}\,\frac{N_k}{N} + \frac{\alpha_0}{N+\alpha_0}\,\frac{\alpha_k}{\alpha_0}},$$

with $N_k = \sum_n \mathbf{z}_{nk}$ and $\alpha_0 = \sum_k \alpha_k$.

This can be useful in to smooth count estimates in IR and NLP since data is very sparse. There exists however other smoothing methods.

Outline

The bag-of-word model, a vector-space representation of documents

Given

- a vocabulary of size d,

The bag-of-word model, a vector-space representation of documents

Given

- a vocabulary of size d,

Represent a document consisting of N words

$$(w_1, \ldots, w_N)$$

The bag-of-word model, a vector-space representation of documents

Given

- a vocabulary of size d,

Represent a document consisting of N words

$$(w_1, \ldots, w_N)$$

as x the vector of counts, or the vector of frequencies of the number of appearances of each of the words (possibly corrected with *tf-idf*):

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{N}_+^d, \quad \text{or } [0,1]_+^d, \quad \text{or } \mathbb{R}^d.$$

The bag-of-word model, a vector-space representation of documents

Given

- a vocabulary of size d,

Represent a document consisting of N words

$$(w_1, \dots, w_N)$$

as x the vector of counts, or the vector of frequencies of the number of appearances of each of the words (possibly corrected with *tf-idf*):

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{N}_+^d, \quad \text{or } [0,1]_+^d, \quad \text{or } \mathbb{R}^d.$$

Document collection

$$X = \begin{bmatrix} | & & | \\ x^{(1)} & \dots & x^{(M)} \\ | & & | \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & & x_1^{(M)} \\ \vdots & \ddots & \vdots \\ x_d^{(1)} & & x_d^{(M)} \end{bmatrix} \in \mathbb{R}^{d \times M}$$

The "Naive Bayes" model for classification

Data

- Class label: $C \in \{1, \ldots, K\}$
- Class indicator vector $Z \in \{0, 1\}^K$
- Features $X_j, \quad j = 1, \ldots, D$
  (e.g. word presence)

The "Naive Bayes" model for classification

Model

$$p(\mathbf{z}) = \prod_k \pi_k^{z_k}$$

Data

- Class label: $C \in \{1, \ldots, K\}$
- Class indicator vector $Z \in \{0, 1\}^K$
- Features $X_j, \quad j = 1, \ldots, D$
  (e.g. word presence)

The "Naive Bayes" model for classification

Data

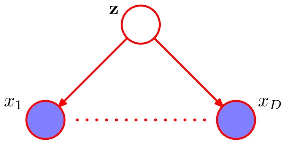
- Class label: $C \in \{1, \dots, K\}$
- Class indicator vector $Z \in \{0,1\}^K$
- Features $X_j, \quad j = 1, \dots, D$
  (e.g. word presence)

Model

$$p(\mathbf{z}) = \prod_k \pi_k^{z_k}$$

Which model for

$$p(x_1, \dots, x_D | z_k = 1) \; ?$$



"Naive" hypothesis

$$p(x_1, \dots, x_D | z_k = 1) = \prod_{j=1}^{D} p(x_j \mid z_k = 1; b_{jk}) = \prod_{j=1}^{D} b_{jk}^{x_j} (1 - b_{jk})^{1-x_j}$$

Naive Bayes (continued)

Learning (estimation)

$$\hat{\pi} = \underset{\pi:\pi^\top \mathbf{1}=1}{\operatorname{argmax}} \prod_{k,i} \pi_k^{z_k^{(i)}} \qquad \hat{b}_{jk} = \underset{b_{jk}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_j^{(i)} | z^{(i)} = k; b_{jk})$$

Prediction:

$$\hat{z} = \operatorname{argmax}_z \frac{\prod_{j=1}^{D} p(x_j|z)p(z)}{\sum_{z'} \prod_{j=1}^{D} p(x_j|z')p(z')}$$

Naive Bayes (continued)

Learning (estimation)

$$\hat{\pi} = \underset{\pi:\pi^\top \mathbf{1}=1}{\operatorname{argmax}} \prod_{k,i} \pi_k^{z_k^{(i)}} \qquad \hat{b}_{jk} = \underset{b_{jk}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_j^{(i)} | z^{(i)} = k; b_{jk})$$

Prediction:

$$\hat{z} = \operatorname{argmax}_z \frac{\prod_{j=1}^{D} p(x_j|z)p(z)}{\sum_{z'} \prod_{j=1}^{D} p(x_j|z')p(z')}$$

Properties

- Ignores the correlation between features
- Prediction requires only to use Bayes rule
- The model can be learnt in parallel
- Complexity in $\mathcal{O}(nD)$