Application of probability theory

for machine learning models

Łukasz Podlodowski
lukasz.podlodowski@gmail.com

April, 2016

Outline

Introduction to probability theory

Common distributions

Probability theory and machine learning

Latent Dirichlet Allocation

Example of application of topic models

What does a randomness mean?

- What is a intuition behind concept of the randomness?
  Randomness is the lack of pattern or predictability in events. [1]
- For machine learning methods we need a formal definition of randomness. What is a mathematical definition of the randomness?
- Probability theory is a mathematical framework that allows us to reason about phenomena or experiments whose outcome is uncertain.
- We will talk about probability space which formally is a triple (Ω, F, P)
  - Ω is a **sample space**. This space contains all possible outcomes of experiment, i.e. number of dots on a thrown dice. Typical elements of Ω are often denoted by ω, and are called elementary outcomes, or simply outcomes. The sample space can be finite, countable or uncountable.
  - F is a σ-**field**, which is a collection of subsets of Ω. σ-**field** means that:
    1. $\emptyset \in F$;
    2. $A \in F \rightarrow \Omega - A \in F$;
    3. $\bigcup_{i=1}^{\infty} A_i \in F$.

    We could interpret this field as set of results of interested experiments, i.e. a dice throw which outcome dot number greater than 4.Typical elements of this space are called events or random events. Note that the formal requirements for F do not presuppose correct representation of ,,real" random events space.

Probability space

- P is a probability measure which provide us information about ,,chance" to observe some set of outcomes. In a very beginning approach to probability theory we described probability as proportion of number interested events which cover our requirements to all events. Unfortunately this approach doesn't allow us to use whole analytical instruments so instead of it we want to use some concepts based on measure theory.
- Because of need of formalization of P we use *probability axioms (Kołmogrow Axioms)*:
  1. $P(A) \geqslant 0$;
  2. $P(\Omega) = 1$;
  3. $P\left(\bigcup_{i=1} A_i\right) = \sum_{i=1} P(A_i)$,
     where $A_i \cap A_j = \emptyset$ for $i \neq j$.

Probability space

- Because of this axioms:

$$P\left(\bigcup_{\omega \in A} \{\omega\}\right) = P(A) \text{ (1)}$$

$$P\left(\bigcup_{\omega \in \Omega} \{\omega\}\right) = P(\Omega) = 1 \text{ (2)}$$

- For simplifying notation we often denote $P(\{\omega\})$ as $P(\omega)$.

Probability space

- Let's try to describe probability for every outcome of random generating number from $[0;1]$. Note that since sum is a binary operator we couldn't handle with infinite sequence of adding probabilities. Because of that (1) doesn't allow us to directly handle with this situation. When the sample space Ω is uncountable, the idea of defining the probability of a general subset of Ω in terms of the probabilities of elementary outcomes runs into difficulties. This is the main reason of setting σ-field to probability space definition. The idea is to assign probability value for a whole subset not for a specific element.
- The pair (Ω, F) is called a **measurable space** and the triple (Ω, F, P) is called a **probability space**.

Random variable

- A **random variable** is a measurable function from the set of possible outcomes Ω to some set E, $X : \Omega \to E$. Usually $E = \mathbb{R}$.
- The random variable doesn't represent probability, which as we have already said is represented by measure P. The main purpose of introducing it is to easily describe some numerical properties of outcomes, i.e. a number of people taller than 1.9m in a population or the number of dice throws with number of dots higher than 4.
- Let's X be a random variable which describes a sum of dots which outcomes in a sequence of three throws. "How likely is it that the value of X is equal to 3?" which formally we denote as $P(\{\omega : X(\omega) = 3\})$. For simplifying notation we often will describe it as: $P(X = 3)$

Random variable

- In case of getting random variable's value in process of executing some experiment we will call random variable observable, otherwise we will call it unobservable.
- Collection of all probabilities for each possible value of a random variable allow us to define some object called **probability distribution**. We describe this object by **probability density function (PDF)** for uncountable random variables and **probability mass function (PMF)** for discrete variables. Note that PDF would rather represent probability concentration than direct probability values. PDF can take values higher than 1. Integrate PDF over some area allow to receive a probability of event.
- In context of machine learning based on probability theory we would often use term **sampling distribution** which could be interpreted as collect some observation of random variable instances.

Random variable

- Conditional probability describes probability of observing some value of random variable when some specific values of other random variables was observed.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

could be read as "the probability of X under the condition Y" or "the conditional probability of X given Y".

Random variable

- In case when observation of any values of random variable X doesn't have influence of a observed value of other random variable Y we call them **independent**. Independence of random variables X and Y means for subset of sample space (event) value for specific value of Y X does have still the same distribution.

$$P(X|Y) = \frac{P(X \cap Y)}{\text{P(Y)}} = \frac{P(X)P(Y)}{P(Y)} = P(X).$$

normalization factor for evaluation distribution probability in a subspace

Bayes theorem

For simplify some analysis we decompose random process into two parts. A prior probability $P(A)$ represent probability of some process evaluated in basics of collected information before experiment was done. We could interpret it as the initial degree of belief in A. A posterior $P(A|B)$ is a probability after experiment was done (and event B happens), is the degree of belief having accounted for B. $P(B|A)$ is the probability of observing event B given that A is true.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \iff P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Bayes theorem is a common used in machine learning, i.e. classification, for machine learning engineers is fundamental theorem of probability theory.

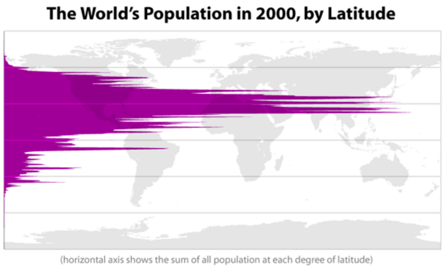
Marginal distribution

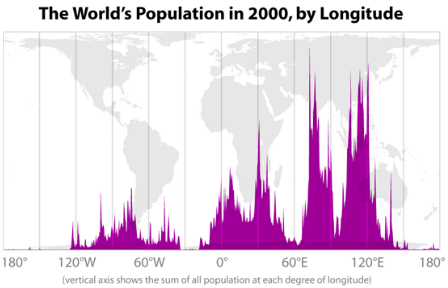
Marginal distribution represents a distribution of some subset of random variables. Marginal is "going to ask about just one (or a few) factor at a time". For continuous distributions:

$$p_X(x) = \int_y p_{X|Y}(x|y)\, p_Y(y)\ \mathrm{d}y = \mathbb{E}_Y\left[p_{X|Y}(x|Y)\right]$$

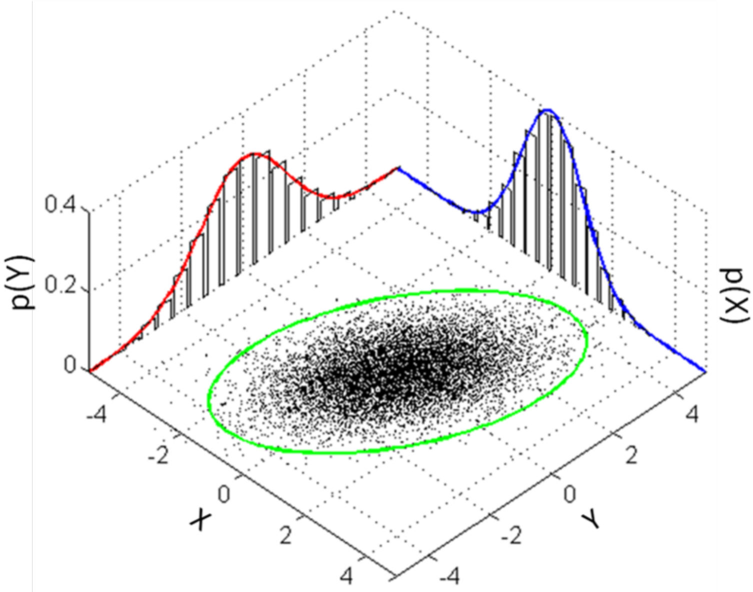
For discrete distribution formula is analogical with exchanging integral into discrete "equivalent" operation of sum.

Marginal distribution



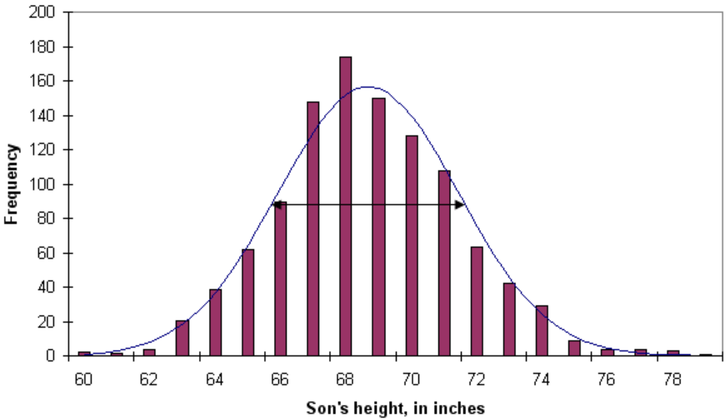


Marginal distribution



Maximum likelihood estimation

- Let's consider us some observation from a normal distribution. How to estimate parameters of this distribution?
- **Maximum likelihood estimation (MLE)** method is a vary used tool for estimating parameters in order to finding distribution which optimize likelihood of sampling our data.



Maximum likelihood estimation

- Common approach focus on an optimization of log of the MLE criterion. Since the log function is monotonic it allow to find the same optimal solution as optimization based on direct MLE criterion.
- Application of log-likelihood function allow to simplify some computation and is more resistance to loss precision on processing very small likelihood values on computers.
- MLE approach is common called "classical (frequentist) inference"
- What is wrong with this approach? Nothing, but is not corresponding to our intuitive understanding of problem. Data is assumed to be random, parameter is fixed. From mathematical point of view this approach don't allow to be so easily interpreted in a learning (estimating) process.

Bayesian parameter estimation

- Based on Bayes theorem. We assume that parameters of model are random variables.
- We specify some distribution of joint distribution over data and parameters $p(X, \theta)$

$$p(y, \theta) = p(y|\theta)p(\theta);$$

- We combine the data we have collected with our prior beliefs is done via Bayes' theorem:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} = p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\int p(x|\theta)\, p(\theta)\, d\theta}.$$

Bayesian parameter estimation

- We need to specify prior distribution of θ.
- If the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called **conjugated**.
- Conjugation of distribution in Bayesian statistics is very desired property of distributions. It could simplify analysis and computation.
- Fortunately all distribution from an exponential family have corresponding to them conjugate distributions.

Bernoulli distribution

For simple experiment which could outcome with "success" or "fail" with probability equals p of receiving success we use Bernoulli distribution:

$$P(X = 1) = 1 - P(X = 0) = 1 - q = p.$$

$$f(k; p) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

Binomial distribution

Consider experiment which is a sequence of independent experiments described by Bernoulli distribution. The probability of getting exactly k "successes" in n trials is given by:

$$P(X = k) = \underbrace{\binom{n}{k}}_{\text{number of } k \text{ combinations}} \underbrace{p^k (1-p)^{n-k}}_{\text{sequence of } k \text{ successes}}$$

Poisson distribution

The Poisson distribution is a continuous "generalization" of binomial distribution. In binomial distribution we talk about some steps, in each of them we performed one experiment. What if we would want to swap this discrete steps into continuous time?
Poisson limit theorem:
if $n \to \infty, p \to 0$, such that $np \to \lambda$ then:

$$\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k} \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

λ is interpreted as expected number of "successes" in some period.

Poisson distribution

Note that since $np = \lambda$, we can rewrite $p = \lambda/n$ so:

$$\lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$(1-p)^{n-k} = \left(1-\frac{\lambda}{n}\right)^{n-k} = \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-k} \xrightarrow{n\to\infty} e^{-\lambda}$$

$$\lim_{n\to\infty} \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k k!} = \frac{1}{k!}.$$

Poisson distribution

Finally Poisson distribution can be described by an equation:

$$f(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

factor of "success" events and ordering

factor of "fail" events

Poisson distribution is often used in modeling occurrence of random events in time (i.e. queueing theory). For example to evaluate "how likely is to receive 100k requests for a server in period of one hour?"

Multinomial distribution

Multinomial distribution is generalization of the binomial distribution. Instead of "success" and "fail" we consider more possible outcomes, but the sample space is still finite.

$$P(X_1 = x_1 \text{ and } \dots \text{ and } X_k = x_k) = \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

The probability mass function can be expressed using the gamma function $\Gamma(x)$ as:

$$P(X_1 = x_1 \text{ and } \dots \text{ and } X_k = x_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{k} p_i^{x_i}.$$

Gamma function

We can interpret gamma function $\Gamma(t)$ as continuous "generalization" of factorial. For $t > 0$:

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} \, dx$$

Using integration by parts we can easily show that:

$$\Gamma(t+1) = t\Gamma(t).$$

$$\Gamma(n) = 1 \cdot 2 \cdot 3 \cdots (n-1) = (n-1)!$$

Beta function

Beta function is defined by:

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t$$

We can express beta function by relationship of gamma functions:

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

Dirichlet distribution

Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.

$$f\left(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K\right) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

For $k-1$ dimensional simplex:

$$\begin{aligned} & x_1, \cdots, x_{K-1} > 0, \\ & x_1 + \cdots + x_{K-1} < 1, \\ & x_K = 1 - x_1 - \cdots - x_{K-1}, \\ & \forall_i \alpha_i > 0 \end{aligned}$$

or $f(x;\alpha) = 0$ if x is not a PMF.

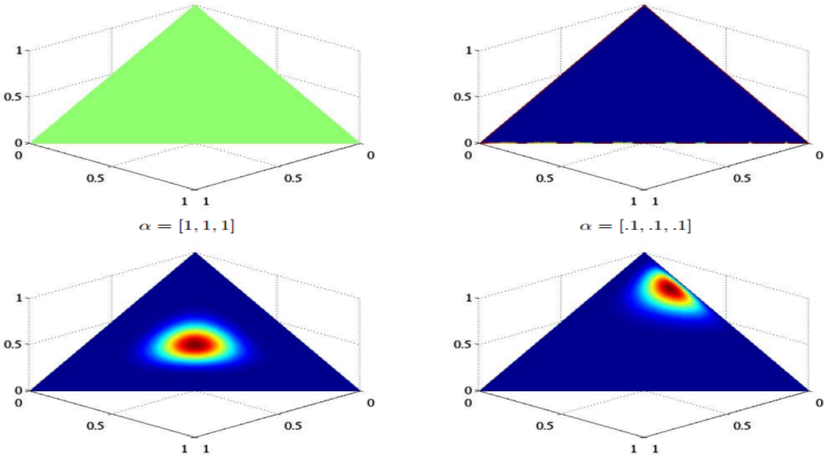
Dirichlet distribution

Where $\mathrm{B}(\boldsymbol{\alpha})$ is a multivariate beta function:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K).$$

From our point of view fact that Dirichlet distribution is a conjugate prior to the multinomial distribution is very important.

Dirichlet distribution - figures from [2]



$\alpha = [1, 1, 1]$

$\alpha = [.1, .1, .1]$

$\alpha = [10, 10, 10]$

$\alpha = [2, 5, 15]$

Figure 1: Density plots (blue = low, red = high) for the Dirichlet distribution over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α. When $\alpha = [c, c, c]$ for some $c > 0$, the density is symmetric about the uniform pmf (which occurs in the middle of the simplex), and the special case $\alpha = [1, 1, 1]$ shown in the top-left is the uniform distribution over the simplex. When $0 < c < 1$, there are sharp peaks of density almost at the vertices of the simplex and the density is miniscule away from the vertices. The top-right plot shows an example of this case for $\alpha = [.1, .1, .1]$, one sees only blue (low density) because all of the density is crammed up against the edge of the probability simplex (clearer in next figure). When $c > 1$, the density becomes concentrated in the center of the simplex, as shown in the bottom-left. Finally, if α is not a constant vector, the density is not symmetric, as illustrated in the bottom-right.

Dirichlet distribution - figures from [2]

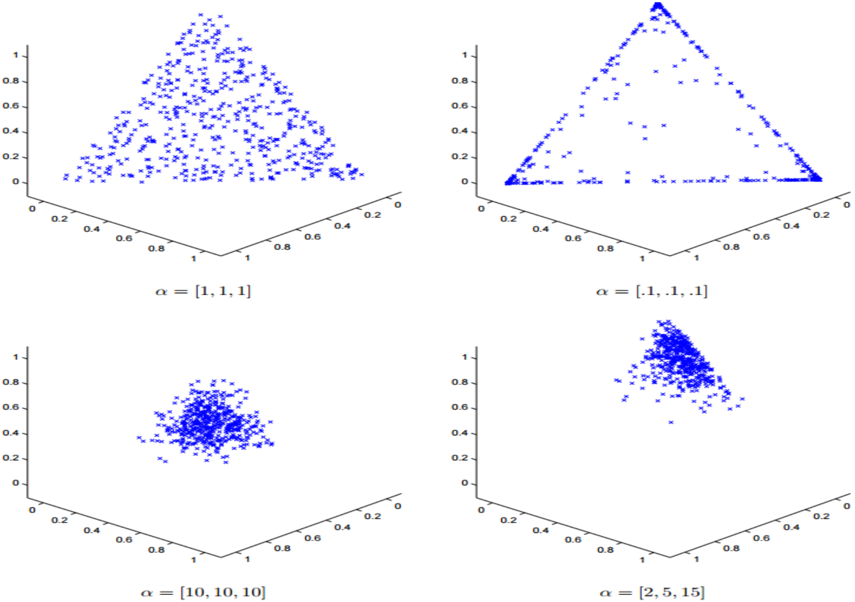
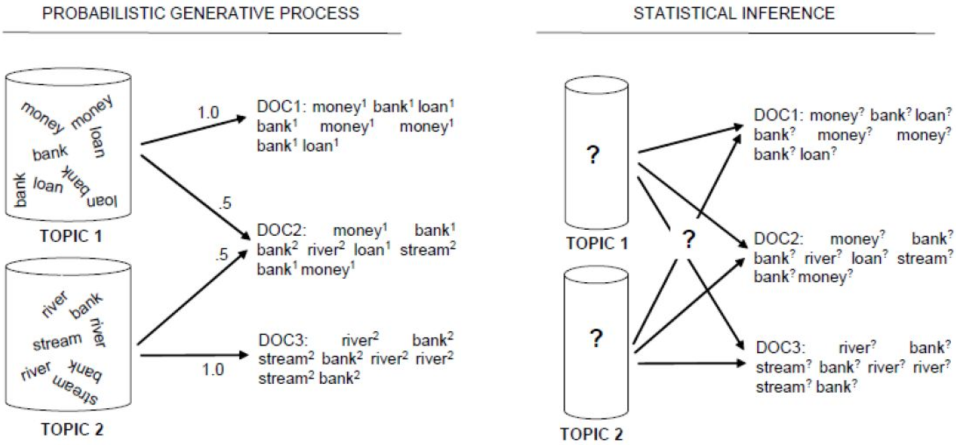


Figure 2: Plots of sample pmfs drawn from Dirichlet distributions over the probability simplex in $\mathbb{R}^3$ for various values of the parameter α.

Does God play Dice?

- As we mentioned randomness doesn't have formal math definition. Intuitively we could understand randomness as lack of knowledge about deterministic rules in some process, of course it does not mean that something which we interpret as random doesn't have any pattern.
- We could say probability theory focus on description some processes based on their outcomes without any deeper analysis of reason, semantic or deterministic rules which made observer outcome.
- Note that this is exactly what we require in machine learning models.

Generative topic models



Latent Dirichlet Allocation - generative process

- Let's assume that document from corpus D could be generated by following process:
  1. Choose number of words $N \sim Poisson(\xi)$
  2. Choose topic mixture $\theta \sim \mathrm{Dir}(\alpha)$
  3. For each of the N words w_n:
     - 3.1 Choose a topic $z_n \sim \mathrm{Multinomial}(\theta_i)$.
     - 3.2 Choose a word from $p(w_n|z_n, \beta)$ where $w_n \sim \mathrm{Multinomial}(\varphi_{z_n})$, which is a multinomial probability conditioned on the topic z_n
- The dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed.
- The Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Note that N is independent of all the other data generating variables (θ and z)

Latent Dirichlet Allocation

- The word probabilities are parametrized by a $k \times V$ matrix β where $\beta_{i,j} = p(w_j = 1 | z_i = 1)$, which we treat as a fixed quantity that is to be estimated.
- Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N words w and corresponding to them topics z is given by:

$$p(\theta, z, w | \alpha, \beta) = \underbrace{p(\theta|\alpha)}_{\text{topic mixture, parameters of word distribution}} \underbrace{\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)}_{\text{words and corresponding topics}}$$

Latent Dirichlet Allocation

- Integrating over θ and summing over z, we obtain the marginal distribution of a document:

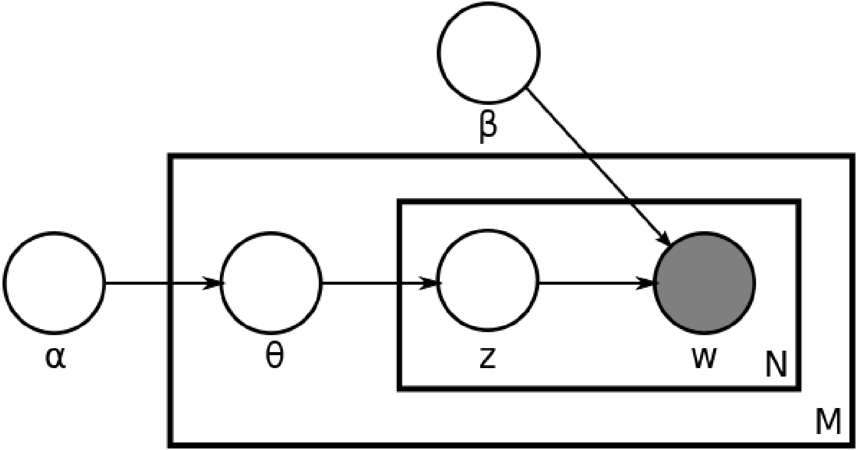
$$p(w|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n,\beta) \right) d\theta$$

- Taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

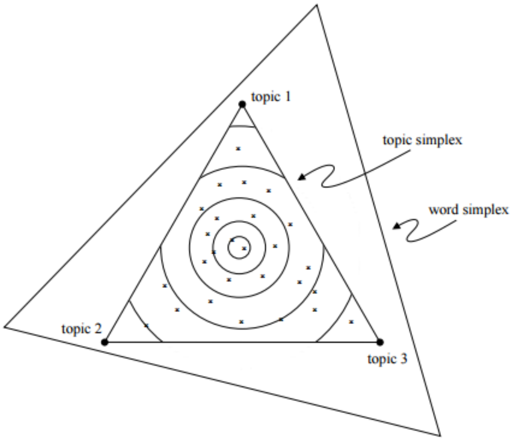
$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d$$

Latent Dirichlet Allocation

- We can distinguish three levels:
  1. α - sampled once per corpus
  2. θ - sampled once per document
  3. w, z - sampled once per word



Latent Dirichlet Allocation



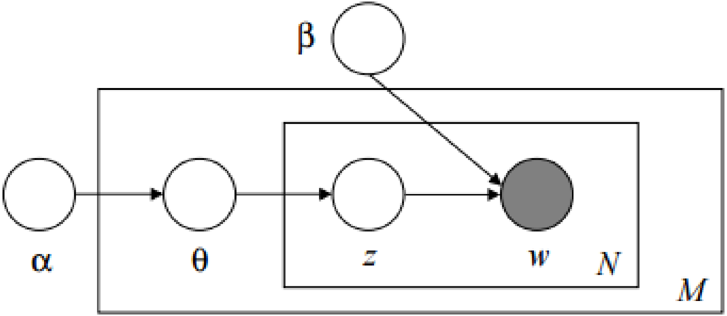
The topic simplex for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word (respectively) has probability one. The three points of the topic simplex correspond to three different distributions over words. The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by x. LDA places a smooth distribution on the topic simplex denoted by the contour lines.

LDA - inference

- The posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo.
- We will focus on variational approximation orignaly proposed by Blei in [3].
- The basic idea of convexity-based variational inference is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood [4][3]. Essentially, one considers a family of lower bounds, indexed by a set of variational parameters. The variational parameters are chosen by an optimization procedure that attempts to find the tightest possible lower bound.

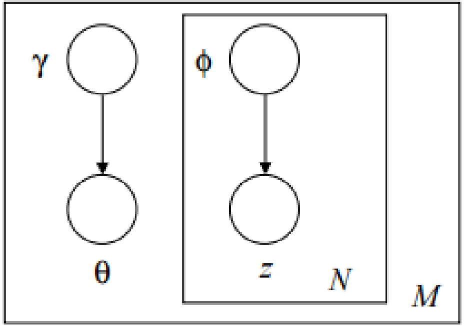
LDA - inference

- Application of variational inference need to specify new variational distribution which allow to simplify optimization process.
- Graphical model representation of LDA:



LDA - inference

Graphical model representation of the variational distribution used to approximate the posterior in LDA (simple modification of the original graphical model in which some of the edges and nodes are removed):



LDA - inference

Variational distribution:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi n)$$

where the Dirichlet parameter γ and the multinomial parameters $(\phi_1, \ldots, \phi_N)$ are the free variational parameters. The optimization problem is defined by:

$$(\gamma^*, \phi^*) = \underset{\gamma,\phi}{\operatorname{argmin}} D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta))$$

where the D is a e Kullback-Leibler (KL) divergence between the variational distribution and the actual posterior distribution .

LDA - inference

One method to minimize this function is to use an iterative fixed-point method, yielding update equations of:

$$\phi_{ni} \propto \beta_{iw_n} \exp \mathbb{E}_q[\log(\theta_i|\gamma)]$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

as shown in [3] $\mathbb{E}_q[\log(\theta i|\gamma)]$ could be computed as:

$$\mathbb{E}_q[\log(\theta_i|\gamma)] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)$$

Ψ is a log of gamma function, which is computable via Taylor approximations.

LDA - inference

Variational EM method take form for E-step:

(1) initialize $\phi_{ni}^{0} := 1/k$ for all i and n

(2) initialize $\gamma_i := \alpha_i + N/k$ for all i

(3) **repeat**

(4)   **for** $n = 1$ **to** N

(5)   **for** $i = 1$ **to** k

(6) $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$

(7)   normalize ϕ_{n}^{t+1} to sum to 1.

(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_{n}^{t+1}$

(9) **until** convergence

LDA - inference

M-step: Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β. This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step.
The β update is based on fact that:

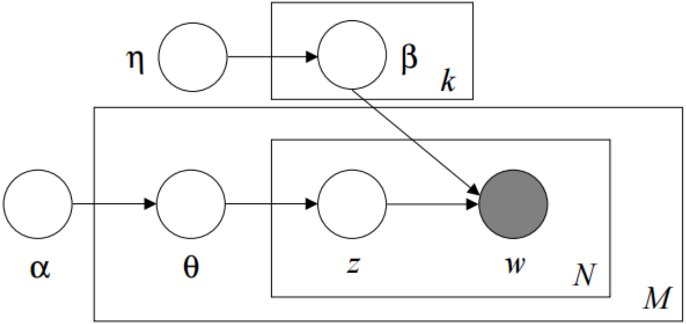
$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dni}$$

The α update uses a linear-scaling Newton-Rhapson algorithm to determine the optimal alpha, with updates carried out in log-space (assuming a uniform α):

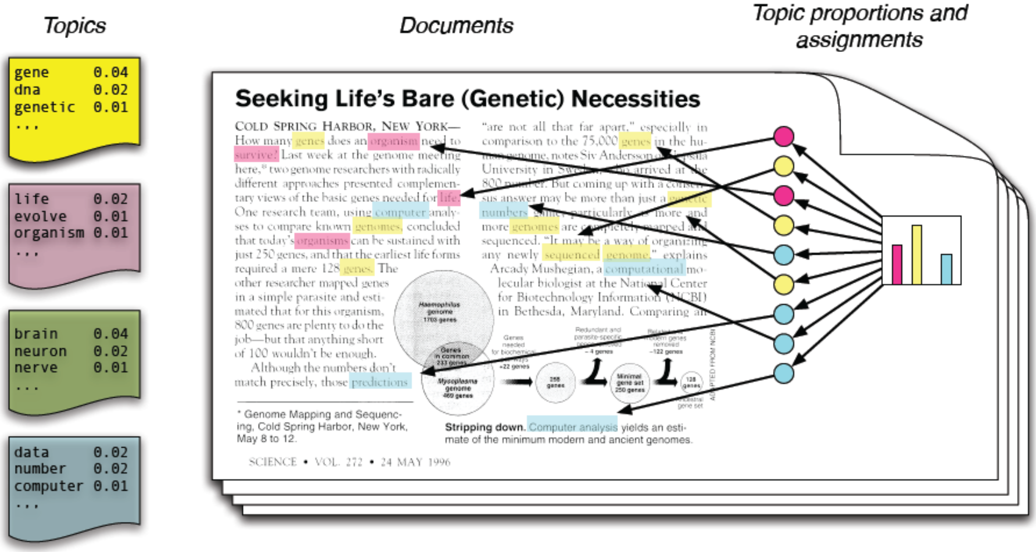
$$\log(\alpha^{t+1}) = \log(\alpha^t) - \frac{\frac{dL}{d\alpha}}{\frac{d^2L}{d\alpha^2}\alpha + \frac{dL}{d\alpha}}$$

Smoothed LDA

For very large corpora frequently occurs problem of sparsity. It is very likely to contain words that did not appear in any of the documents in a training corpus. Maximum likelihood estimates of the multinomial parameters assign zero probability to such words, and thus zero probability to new documents. Smoothed version of LDA model is based on Dirichlet smoothing. Each row in β matrix is treated as each row is independently drawn from an exchangeable Dirichlet distribution.



Application of topic models



Application of topic models

Topic models are common used in many problems. We can get examples of application it in [5], [6] or [7].

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Application of topic models

Topic 247 word	prob.	Topic 5 word	prob.	Topic 43 word	prob.	Topic 56 word	prob.
DRUGS	.069	RED	.202	MIND	.081	DOCTOR	.074
DRUG	.060	BLUE	.099	THOUGHT	.066	DR.	.063
MEDICINE	.027	GREEN	.096	REMEMBER	.064	PATIENT	.061
EFFECTS	.026	YELLOW	.073	MEMORY	.037	HOSPITAL	.049
BODY	.023	WHITE	.048	THINKING	.030	CARE	.046
MEDICINES	.019	COLOR	.048	PROFESSOR	.028	MEDICAL	.042
PAIN	.016	BRIGHT	.030	FELT	.025	NURSE	.031
PERSON	.016	COLORS	.029	REMEMBERED	.022	PATIENTS	.029
MARIJUANA	.014	ORANGE	.027	THOUGHTS	.020	DOCTORS	.028
LABEL	.012	BROWN	.027	FORGOTTEN	.020	HEALTH	.025
ALCOHOL	.012	PINK	.017	MOMENT	.020	MEDICINE	.017
DANGEROUS	.011	LOOK	.017	THINK	.019	NURSING	.017
ABUSE	.009	BLACK	.016	THING	.016	DENTAL	.015
EFFECT	.009	PURPLE	.015	WONDER	.014	NURSES	.013
KNOWN	.008	CROSS	.011	FORGET	.012	PHYSICIAN	.012
PILLS	.008	COLORED	.009	RECALL	.012	HOSPITALS	.011

Application of topic models

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

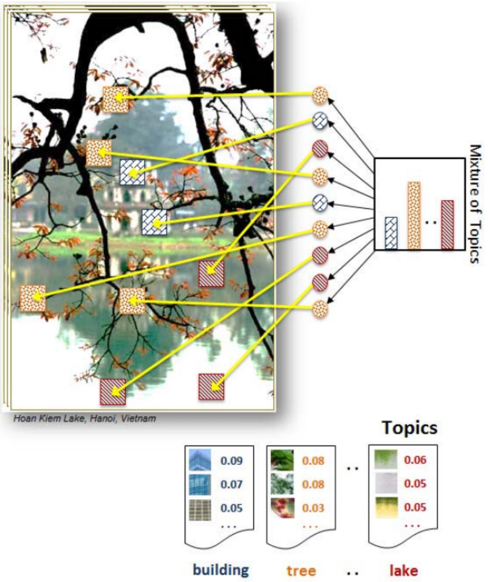
Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

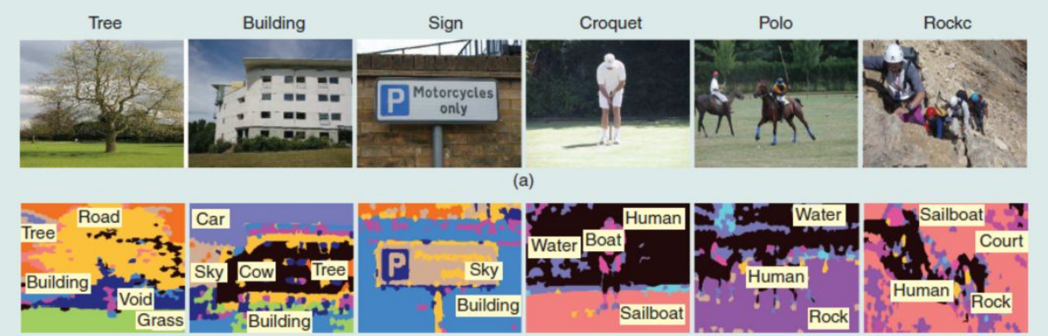
Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Application of topic models



Application of topic models



Application of topic models

Topic index	Typical word pairs
Topic 1	cars†, prototype†, tracks†, street, turn, marsh‡, roofs, bengal§, forest‡, tiger§
Topic 4	plane†, jet†, sky†, sun, birds§, fly§, clouds§, snow, sand†, dunes†
Topic 27	snow†, ice†, polar†, frozen†, bear, mountain§, water, rocks§, grass, sky
Topic 48	island†, beach†, sand, sea†, water†, sky, people, kauai†, sunset, buildings
Topic 72	ocean†, coral†, fish†, rocks§, reefs§, water, orchid, boat§, sky, fan
Topic 1	water, sky†, tree, people, clouds†, grass, mountain, buildings, sun, snow
Topic 9	sky†, jet†, plane†, mountain, tree, water, sun, people, clouds, buildings
Topic 30	tree†, grass†, flowers§, people, field, house, mountain, sky, water, garden§
Topic 41	ice†, people, mountain§, sky, frost, snow§, clouds, water, rocks§, landscape
Topic 67	cars, buildings§, street†, people, sidewalk, lights†, window†, post, store‡, shops‡

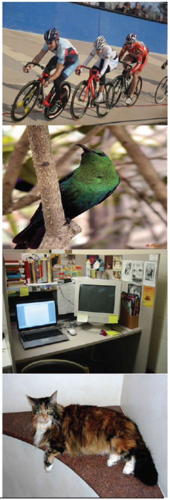
Application of topic models

Groundtruth: bike, velodrome, racing
corrLDA: bike, people, blue, sky
corrCTM: velodrome, bike, people, racing, cycling

Groundtruth: bird, natural, blue, green
corrLDA: bird, animal, sky, flying
corrCTM: bird, park, animal, natural, plant

Groundtruth: computer, desk, office
corrLDA: monitor, computer, desk
corrCTM: monitor, computer, office, desk, chair

Groundtruth: cat
corrLDA: cat, pet, cute, black, puppy
corrCTM: cat, kitty, cute, pet



Groundtruth: bus, yellow
corrLDA: bus, trip, airplane
corrCTM: bus, station, railway

Groundtruth: family, house, car
corrLDA: sky, bird, flying
corrCTM: blue, sky, airplane, green

Figure: Taken from [7]. Note that is a modification of original LDA model for catching correlation between two kinds of words (in this example text and visual)

Oxford English Dictionary.
Oxford University Press, 2010.

A. K. Bela A. Frigyik and M. R. Gupta, "Introduction to the dirichlet distribution and related processes,"

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*.
Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

N. Rasiwasia and N. Vasconcelos, "Latent dirichlet allocation models for image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2665–2679, Nov 2013.

T. H. Tran and S. Choi, "Supervised multi-modal topic model for image annotation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5979–5983, May 2014.

X. Xu, A. Shimada, and R. i. Taniguchi, “Correlated topic model for image annotation,” in *Frontiers of Computer Vision, (FCV), 2013 19th Korea-Japan Joint Workshop on*, pp. 201–208, Jan 2013.