Latent Dirichlet Allocation

Soojung Hong

Feb 6. 2017

Contents

- Introduction : Text corpora modeling
- Background and Terminology
- Latent Dirichlet Allocation
- Relationship with other latent variable models
- Inference and Parameter Estimation
- Applications and Empirical Results
- Summary

Text corpora modeling

- **Goal**

  Finding short descriptions of the members "*documents*" of the collections "*corpora*"

  In particular, this descriptions preserve essential statistical relationships

- **Application area**

  classification, novelty detection, summarization and collaborative filtering

- **Relevant approaches for text modeling**

  tf-idf scheme, LSI, pLSI and latent Dirichlet allocation (LDA)

Summary of other approaches

	Advantage	Disadvantage
tf-idf	Reduces documents from arbitrary length to fixed-length lists of numbers $\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$	Relatively small amount of reduction Reveal little about inter-intra document statistical structure
LSI	Reduce the dimensionality and learns latent topics by performing a matrix composition (SVD) on term-document matrix A $A = S\Sigma U^T$	Not clear why LSI for generative model of text
pLSI	Significant step forward probabilistic modeling of text $P(w, d) = \sum_c P(c)P(d\|c)P(w\|c) = P(d)\sum_c P(c\|d)P(w\|c)$	No probabilistic model at the level of documents, If number of parameter grows linearly with size of corpus, it leads overfitting Not clear how to assign probability to document outside of the training set

Details of tf-idf

Term Frequency - Inverse Document Frequency : $\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$

Term Frequency $\mathbf{tf}(t, d)$: How frequently a word occurs in a document

Inverse Document Frequency $\mathbf{idf}(t, D)$: The inverse document frequency for any given term

$$\mathbf{tf}(t, d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t, D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$f_d(t) :=$ frequency of term t in document d

$D :=$ corpus of documents

$$\mathbf{tfidf}(t, d, D) = \mathbf{tf}(t, d) \cdot \mathbf{idf}(t, D)$$

$$\mathbf{tfidf}'(t, d, D) = \frac{\mathbf{idf}(t, D)}{|D|} + \mathbf{tfidf}(t, d, D)$$

http://www.joyofdata.de/blog/tf-idf-statistic-keyword-extraction/

Details of LSI

LSI (also Latent Semantic Analysis) is the analysis of latent semantics in a corpora of text.

A collection of documents can be represented as a huge term-document matrix

Similarity between document to document, word to word, word to document can be measured by such as cosine similarity (similar words = higher cosine similarity)

Two common problems using LSI is polysemy (a word having different meaning in different contexts) and synonymy (a concept having multiple forms of representation i.e. two or more words denoting the same concept).

LSI transforms the original data in a different space so that two documents/words about the same concept are mapped close. LSI achieves this by Singular Value Decomposition (SVD) of term-document matrix A. $A = S\Sigma U^T$

Details of pLSI

pLSI models each word in a document as a sample from mixture model

The mixture components of mixture model are multinomial random variables that can be viewed as representations of `topics'. Thus each word is generated from a single topic.

Each document is represented as a list of mixing proportions for these mixture components "topics".

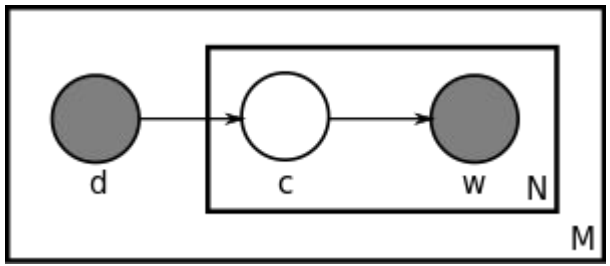


Plate notation representing the pLSA model

d : the document index variable

c : a word's topic drawn from the document's topic distribution $P(c|d)$

w : a word drawn from the word distribution of this word's topic $P(w|c)$

d and w : observable variables

topic c : a latent variable

Assumptions

Bag-of-words (Fundamental probabilistic assumption for dimensionality reduction)

The order of words in document can be neglected.

This is an assumption of exchangeability for the words in a document.

Exchangeability (Documents are exchangeable)

The specific ordering of the documents in a corpus can be neglected

Exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed, it is rather “Conditionally independent and identically distributed”. This conditioning is with respect to an underlying latent parameter of a probability distribution.

Notation and Terminology

Latent variables aim to capture abstract notions such as topics

Word is a basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1,\ldots,V\}$

We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. *v*-th word in the vocabulary is represented by a V-vector *w* such that $w^v = 1$ and $w^u = 0$ for $u \neq v$

Document is a sequence of *N* words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$

Corpus is a collection of M documents denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$

Goal : Find a probabilistic model of a corpus not only assign high probability to members of the corpus but also assign high probability to other similar documents.

LDA

LDA is a generative probabilistic model of a corpus.
The basic idea is that documents are represented as random mixtures over latent topics
Each topic is characterized by a distribution over words.

LDA generative process for each document w in a corpus D:

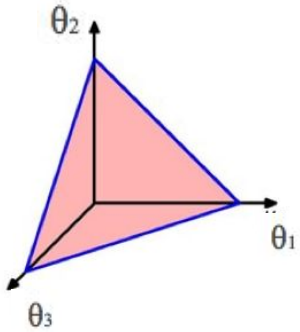
1. Choose N ~ Poisson(ξ)
2. Choose θ ~ Dir(α)
3. For each of the N words w_n:
   (a) Choose a topic z_n ~ Multinomial(θ).
   (b) Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

Several simplifying assumption :

The dimensionality k of the Dirichlet distribution (and thus the dimensionality of topic variable z) is assumed to be known and fixed

The word probabilities are parameterized by $k \times V$ matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$ which for now we treat as a fixed quantity to be estimated. Poisson assumption is not critical to anything. *N* is independent to all other data generating θ and z

LDA (continue)



k-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex

(k vector θ lies in the k-1 simplex if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$

The probability density on this simplex : $p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$

Parameter α is a k-vector with components with $\alpha_i > 0$, and where $\Gamma(x)$ is gamma function.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

LDA (continue)

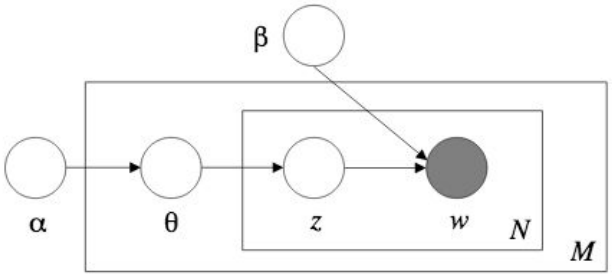
Where $p(z_n \mid \theta)$ is simply θ_i for the unique i such that $z_n^i = 1$

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

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

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus.

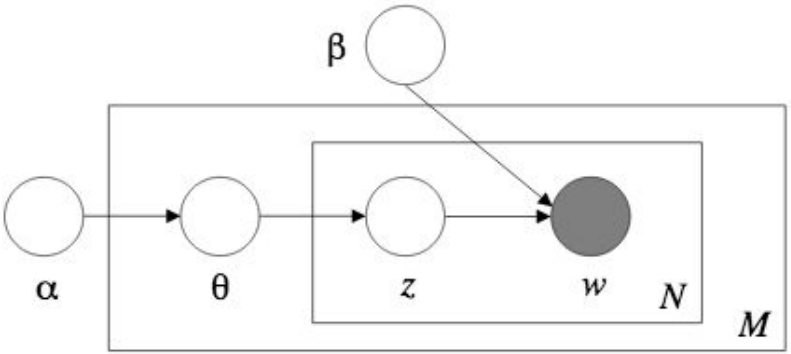
$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$



Graphical Model Distribution of LDA

Three levels to the LDA representation

1. The parameter α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus.
2. Variable θ_d are document-level variables, sampled per document.
3. Variables z_{dn} and w_{dn} are word-level variables and sampled once for each word in each document. (difference between Dirichlet multinomial clustering model)



LDA and Exchangeability

If finite set of random variables $\{z_1, \ldots, z_n\}$ is said to be *exchangeable*, if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N.

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(N)}).$$

An infinite sequence of random variables is infinitely exchangeable, if every infinite subsequence is *exchangeable*.

De Finetti's representation theorem states that "the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were independent and identically distributed, conditioned on that parameter".

In LDA, we assume that words are generated by topics (by fixed conditional distributions) and that those topics are infinitely exchangeable within a document.

By de Finetti's theorem, the probability of a sequence of words and topics must therefore have the form.

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

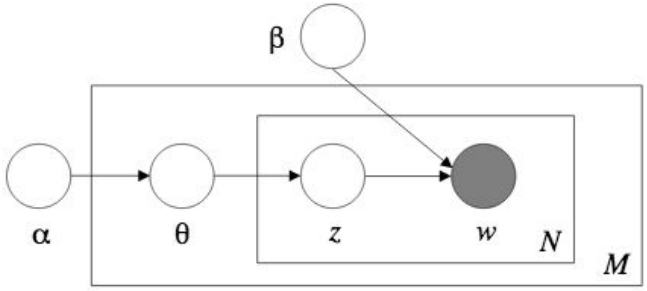
Where θ is the random parameter of a multinomial over topic. We obtain the LDA distribution on documents in Equation below by marginalizing out the topic variables and endowing θ with a Dirichlet distribution.

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

Continuous mixture of unigrams

LDA model in following is more elaborate than the two-level models often studied in the classical hierarchical Bayesian literature.

By marginalizing over the hidden topic variable z, however, we can understand LDA as a two-level model.



The word distribution $p(\mathrm{w} \mid \theta, \beta)$ is $p(w|\theta,\beta) = \sum_z p(w|z,\beta)p(z|\theta)$

Continuous mixture of unigrams (continue)

Given word distribution generative process for a document w :

1. Choose $\theta \sim Dir(\alpha)$
2. For each of the N words wn :

Choose a word w_n from $p(w_n \mid \theta, \beta)$, $p(w \mid \theta, \beta) = \sum_z p(w \mid z, \beta) p(z \mid \theta)$

This process defines the marginal distribution of a document as a continuous mixture distribution,

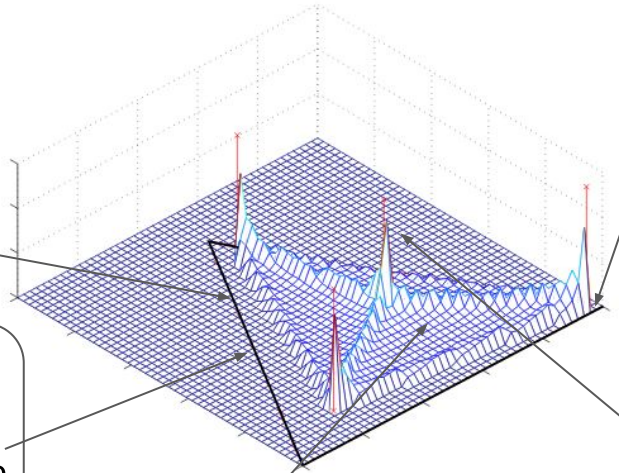
$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} p(w_n \mid \theta, \beta) \right) d\theta,$$

$p(w_n \mid \theta, \beta)$ are mixture components and $p(\theta \mid \alpha)$ are the mixture weights and then marginal out with θ

Interpretation of LDA

Example density on unigram distributions $p(w|\theta,\beta)$ under LDA for three words and four topics

The triangle embedded in the x-y plane is the 2-D simplex representing all possible multinomial distributions over three words.



Each of the vertices corresponds to a deterministic distribution that assigns probability one to one of the words

Midpoint of an edge gives probability 0.5 to two of the words

Centroid of the triangle is the uniform distribution over all three words

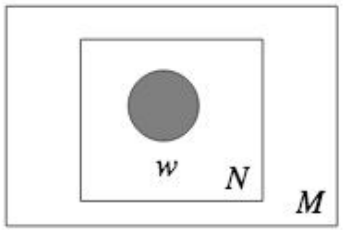
Locations of the multinomial distributions $p(w|z)$ for each of the four topics, and the surface shown on top of the simplex is an example of a density over the $(V-1)$-simplex (multinomial distributions of words) given by LDA.

Comparison : LDA and other latent variable models

(a) Unigram model

The words of every document are drawn independently from single multinomial distribution.

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

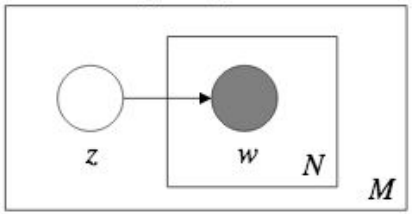


(a) unigram

(b) Mixture of unigram

Augment the unigram model with a discrete random topic variable z and obtain a mixture of unigrams model. Each document is generated by the first choosing topic z and then generating N words independently from the conditional multinomial p(w|z).

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z).$$

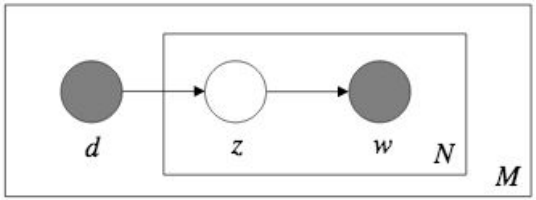


(b) mixture of unigrams

(c) pLSI model

Posits that a document label d and a word w_n are conditionally independent given an unobserved topic z. $p(d, w_n) = p(d) \sum_{z} p(w_n|z)p(z|d)$

pLSI model does capture the possibility that a document may contain multiple topics since p(z|d) serves as the mixture weights of the topics for particular document d.



(c) pLSI/aspect model

Drawback of pLSI

- pLSI does capture the possibility that a document may contain multiple topics :
  - $p(z|d)$ serves as the mixture weights of the topics for a particular document d
  - However, d is a dummy index into the list of documents in the training set. Thus d is a multinomial random variable restricted in training documents

- pLSI is not well-defined generative model of documents :
  - There is no natural way to assign probability to a previously unseen document.
  - The parameters for a k-topic pLSI model are K multinomial distributions of size V and M mixtures over the k-hidden topics. This gives $kV + kM$ parameters and therefore linear growth in M. The linear growth in parameters suggests that the model is prone to overfitting.

LDA overcomes pLSI problem

By treating the topic mixture weights as a k-parameter hidden random variable rather than a large set of individual parameters which are explicitly linked to the training set.

LDA is a well-defined generative model and generalize easily to new documents. Furthermore, the k+kV parameters in k-topic LDA model do not grow with the size of training corpus, therefore doesn't suffer from the issue of overfitting.

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d)$$ **pLSI**

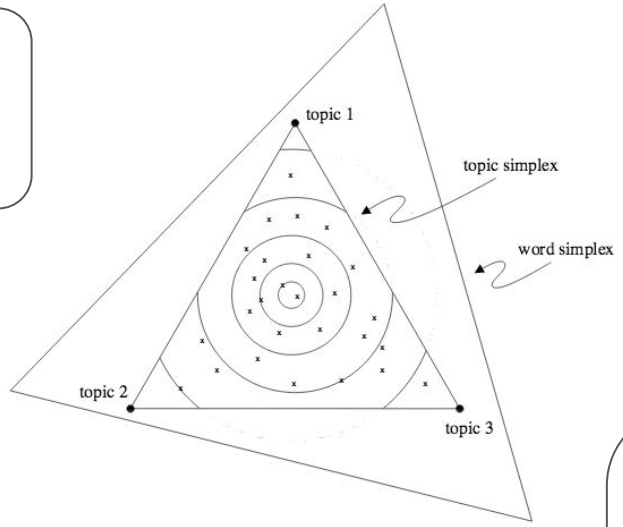
$$(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_N p(z_n | \theta) p(w_n | z_n, \beta)$$ **LDA**

Geometric Interpretation

Difference between LDA and other latent topic models (unigram, mixture of unigrams, pLSI) using geometry of the latent space

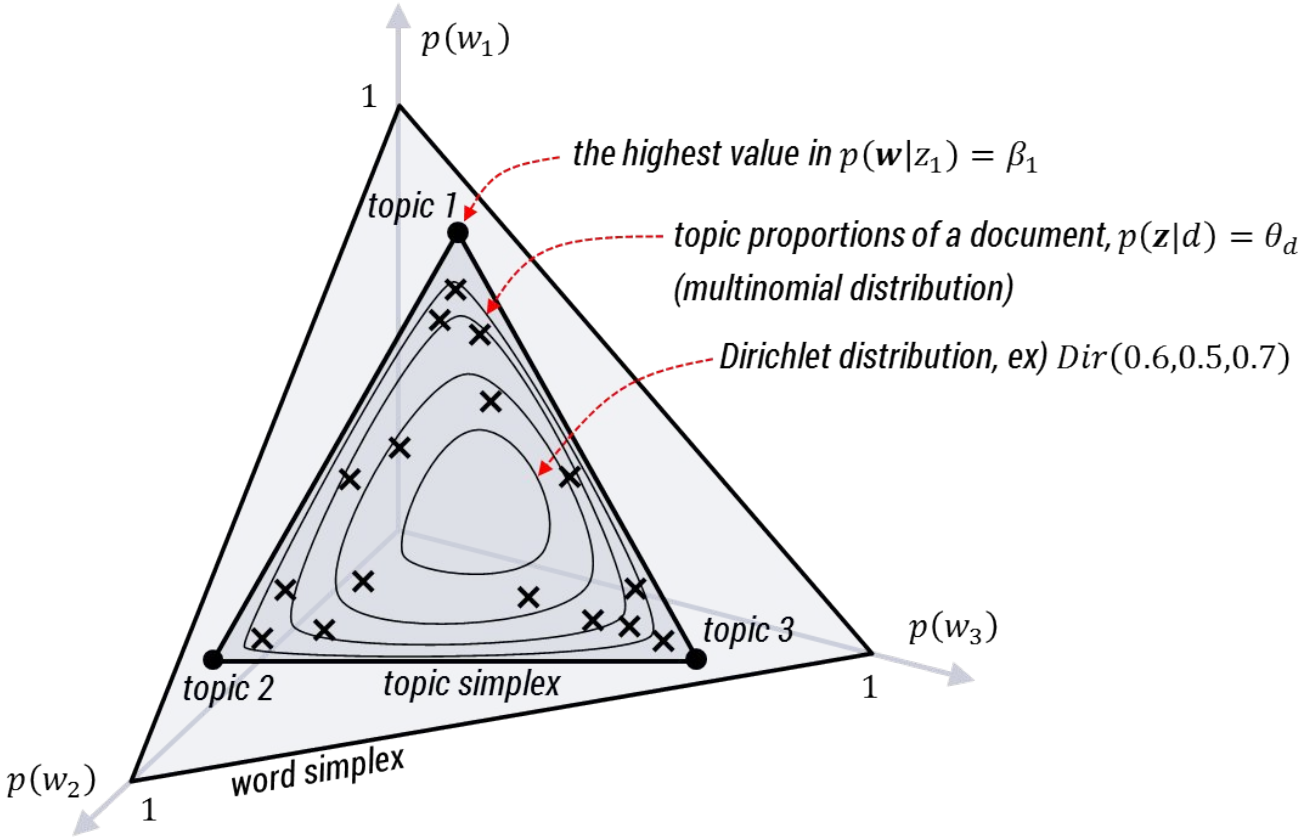
The unigram model finds a single point on the word simplex and posits that all words in the corpus come from the corresponding distribution.

Mixture of unigrams model posits that for each document, one of the k points on the word simplex (that is, one of the corners of the topic simplex) is chosen randomly and all the words of the documents are drawn from the distribution corresponding to that point.



pLSI model posits that each word of a training document comes from randomly selected chosen topic. The topics are themselves drawn from a document-specific distribution over topics.(ie. Points on the topic simplex). There is one such distribution for each document.

LDA posit that each word of both observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter. This



http://parkcu.com/blog/

Inference and Parameter Estimation

The key inferential problem to solve in order to use LDA is computing the posterior distribution of the hidden variables given a document. This distribution is intractable to compute in general.

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}.$$

To normalize the distribution we marginalize over the hidden variables and write following equation

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

in terms of model parameter

$$p(\mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta, \quad (*)$$

⇒ Needs approximate inference algorithm (such as Laplace approximation, Variational Inference)

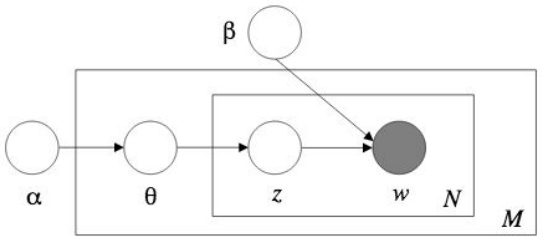
Inference and Parameter Estimation : Variational Inference

The basic idea of convexity-based variational inference is to obtain an adjustable lower bound on the log likelihood. (using Jensen's inequality)

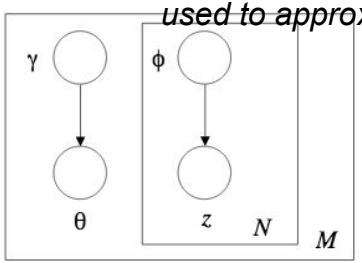
One considers a family of lower bounds, indexed by a set of variational parameters. The variational parameters are chosen by an optimization procedure that attempts to find the tightest possible lower bound.

A simple way to obtain a tractable family of lower bounds is to consider simple modifications of the original graphical model in which some of the edges and nodes are removed.

Graphical model representation of LDA



Graphical model representation of the variational distribution used to approximate the posterior in LDA



The goal of variational inference is to maximize the variational lower-bound with respect to approximate q distribution or minimize $\mathrm{KL}(q||p)$.

The problematic coupling between θ and β arises due to the edges between θ, z and w. By dropping these edges and the w nodes, and endowing the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables. This family is characterized by the following variational distribution:

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n), \qquad (\phi_1, \ldots, \phi_N)$$

where the Dirichlet parameter γ and the multinomial parameters $(\varphi_1,\ldots,\varphi_N)$ are the free variational parameters.

The next step is to set up an optimization problem that determines the values of the variational parameters γ and φ.

Parameter Estimation

Given a corpus of documents $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$

Find parameters α and β that maximize the (marginal) log likelihood of the data.

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \,|\, \alpha, \beta)$$

As described before, the quantity $p(\mathrm{w}|\alpha,\beta)$ cannot be computed tractable. Using variational inference provides us with a tractable lower bound on the log likelihood, which we can maximize with respect to α and β.

We can thus find approximate empirical Bayes estimates for the LDA model via an alternating variational EM procedure that maximizes a lower bound with respect to the variational parameters γ and φ , and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters α and β.

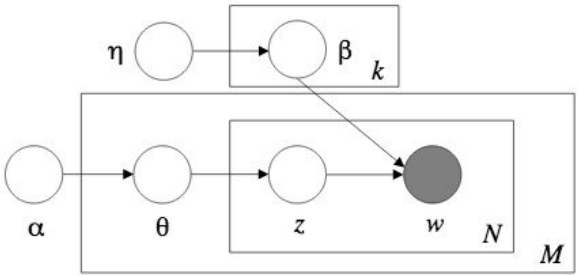
Smoothing

Large vocabulary size that is characteristic of many document corpora problems with sparsity.

Maximum likelihood estimates of the multinomial parameters assign zero probability to new words, and thus zero probability to new documents.

To avoid this problem using “smooth” the multinomial parameters. Assigning positive probability to all vocabulary items whether or not they are observed in the training set.

Proposed solution for smoothing is to simply apply variational inference methods to the extended model that includes Dirichlet smoothing on the multinomial parameter.



Graphical model representation of the smoothed LDA model.

Example Data

Data : 16,000 documents from a subset of the TREC AP corpus

Preparation : Removing a standard list of stop words and we use EM algorithm to find the Dirichlet and conditional multinomial parameters for a 100-topic LDA model.

The top words from some of the resulting multinomial distributions $p(w|z)$ are in following figure. (hopefully) These distributions capture some of the underlying topics in the corpus

"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

Applications and Empirical Results : (1) Document Modeling

Trained a number of latent variable models, including LDA, on two corpora

Comparison of generalization performance of each models. The documents in the corpora are treated as unlabeled. Thus, our goal is density estimation which achieve high likelihood on a held-out test set.

Computed the perplexity of a held-out test set to evaluate the models. The perplexity, used by convention in the language modeling, is monotonically decreasing in the likelihood of the test data. A lower perplexity score indicates better generalization performance.

Formally, a test set of M documents, the perplexity is following.

$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

In our experiment, we used a corpus of scientific abstracts from the C. Elegans community containing,

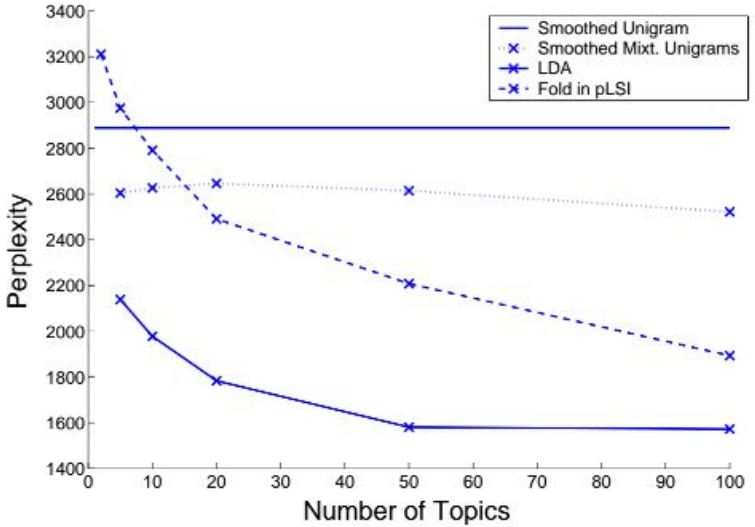
5225 abstracts with 28414 unique terms

A subset of TREC AP corpus containing, 16333 newswire articles with 23075 unique terms.

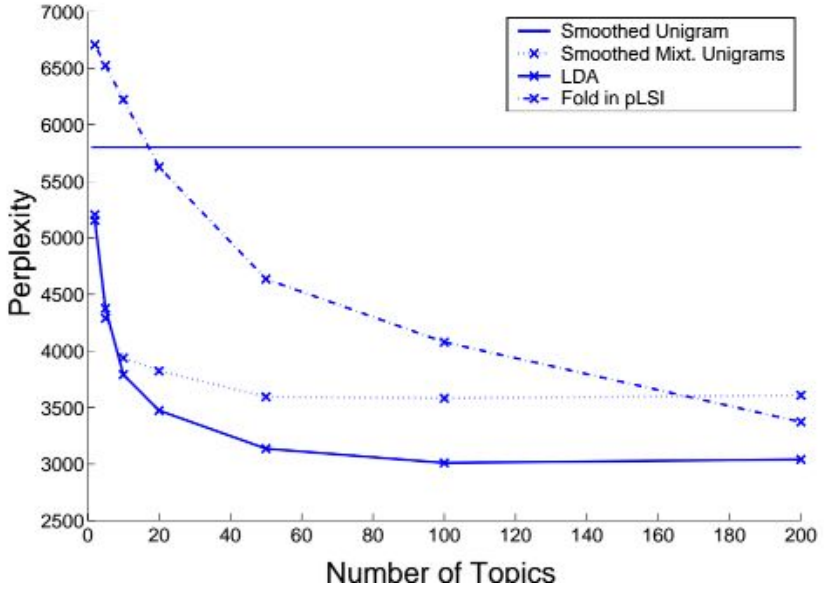
Both case, the 10% of data for test purpose and trained the models on remaining 90%.

Perplexity Results

- 5225 abstracts
- 28414 unique terms



- 6333 newswire articles
- 23075 unique terms



Mixture of unigram model and pLSI both suffer serious overfitting issue.
LDA can easily assign probability to a new document without overfitting

Applications and Empirical Results : Document classification

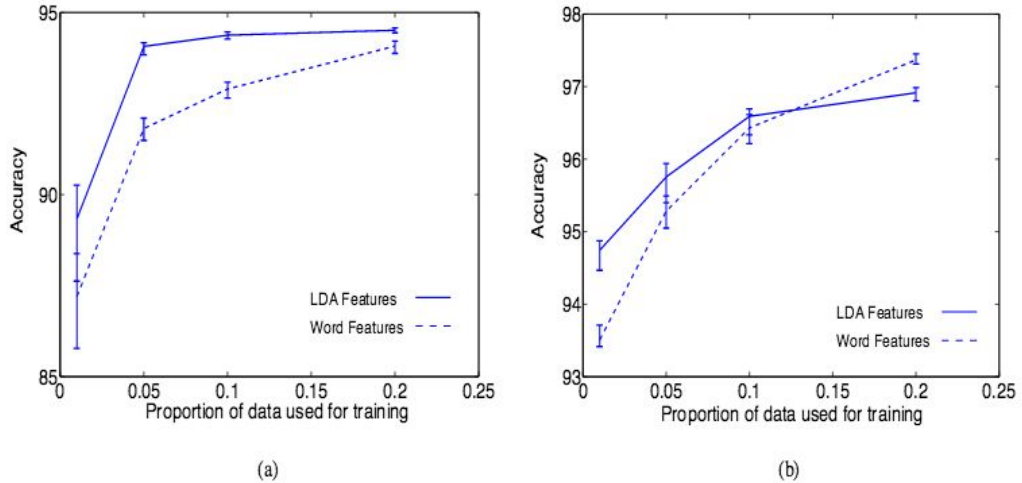
The goal of text classification problem : classify a document into two or more mutually exclusive classes. As in any classification problem, we may wish to consider generative approaches or discriminative approaches. In particular, by using one LDA module for each class, we obtain a generative model for classification.

A challenging aspect of the document classification problem is the choice of features. Treating individual words as features yields a rich but very large feature set. One way to reduce feature set is to use an LDA model for dimensionality reduction. In particular, LDA reduces any document to a fixed set of real-valued features - the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$) associated with the document.

Two binary classification experiments using Reuter-21578 dataset

The dataset contains 8000 documents and 15,818 words.

Estimated the parameters of an LDA model on all the documents, without reference to their true class label. Then trained a support vector machine (SVM) on the low-dimensional representations provided by LDA and compared this SVM to an SVM trained on all the word features. (i.e. low-dimensional features by LDA vs. all features, then use SVM classification)



Applications and Empirical Results : Collaborative filtering

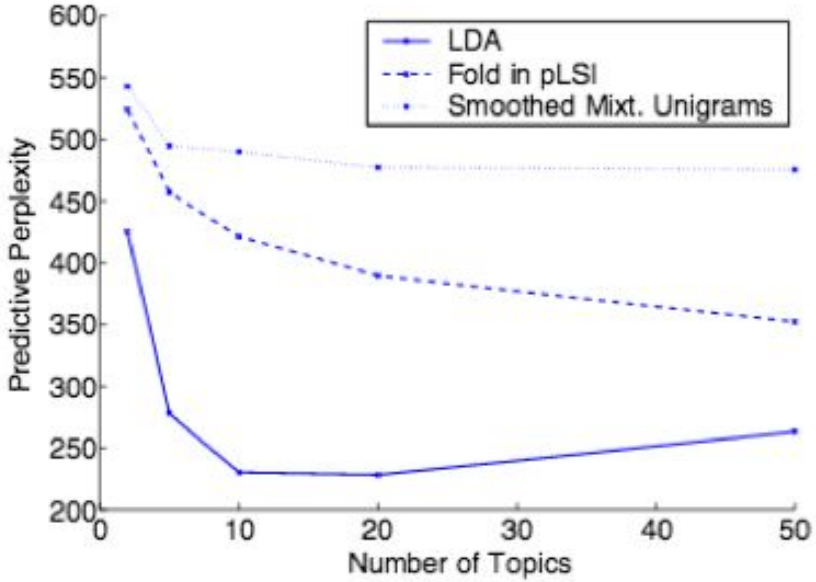
Final experiment on the EachMovie collaborative filtering data.

A collection of users indicates their preferred movie choice. A user and the movie chosen are analogous to a document and the words in the document (respectively).

The collaborative filtering task is as follows :

1. Train a model on fully observed set of users
2. For each unobserved user, we are shown all but one of movies preferred by that user and are asked to predict what the held-out movie is. The different algorithms are evaluated according to the likelihood they assign to the held-out movie.
3. predictive perplexity on M test users to be :

$$predictive\text{-}perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_{d,N_d} \mid \mathbf{w}_{d,1:N_d-1})}{M}\right\}.$$



Result for collaborative filtering on EachMovie data

3300 training users and
390 testing users

Mixture of unigram

$$p(w|\mathbf{w}_{\text{obs}}) = \sum_z p(w|z)p(z|\mathbf{w}_{\text{obs}}).$$

LDA model

$$p(w|\mathbf{w}_{\text{obs}}) = \int \sum_z p(w|z)p(z|\theta)p(\theta|\mathbf{w}_{\text{obs}})d\theta,$$

Summary

Latent Dirichlet Allocation is a generative probabilistic model for collections of data.

LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics.

Each topic is in turn, modeled as an infinite mixture over an underlying set of topic probability.

This paper presents efficient approximate inference techniques based on variational methods and EM algorithm for empirical Bayes parameter estimation.

The paper reports results in document modeling, text classification and collaborative filtering, compared to mixture of unigrams model and pLSI model.

==========

Reference : Jensen's Inequality

Jensen's inequality to obtain an adjustable lower bound on the log likelihood

Jensen's inequality generalizes the statement that the secant line of a convex function lies *above* the graph of the function, which is Jensen's inequality for two points: the secant line consists of weighted means of the convex function,

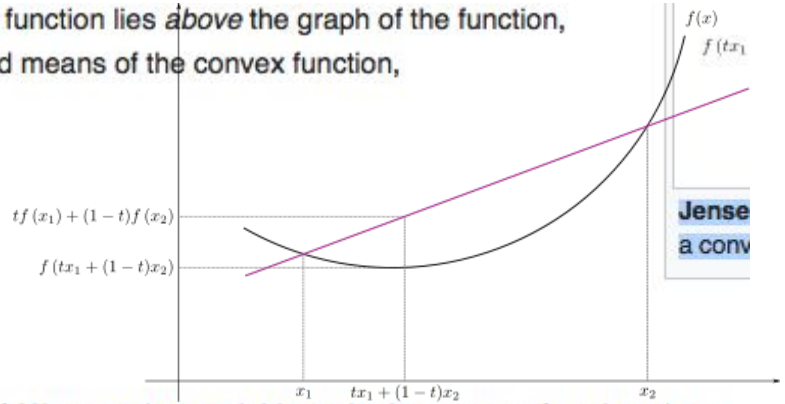
$$tf(x_1) + (1-t)f(x_2),$$

while the graph of the function is the convex function of the weighted means,

$$f\left(tx_1 + (1-t)x_2\right).$$

Thus, Jensen's inequality is

$$f\left(tx_1 + (1-t)x_2\right) \leq tf(x_1) + (1-t)f(x_2).$$



In the context of probability theory, it is generally stated in the following form: if X is a random variable and φ is a convex function, then

$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\varphi(X)\right].$$

Reference : Variational Inference

In many statistical problems, we can't directly calculate the posterior because the normalization constant is intractable. This happens often in latent variable models. For example assume that X represents a set of observations and Z represents a set of latent variables. If we are interested in the posterior P(Z|X), we know that

$$P(Z|X) = \frac{P(Z,X)}{\int_z P(Z,X)}$$

but often times we can't calculate the denominator. One popular approach is MCMC, where we can sample exactly from the true posterior distribution; however, convergence can be prohibitively slow if we have many parameters to sample.

$$P(Z|X) \approx Q(Z|V) = \prod_i Q(Z_i|V_i)$$

This is where variational inference comes in handy. Variational inference **seeks to approximate the true posterior, P(Z|X),** with an approximate variational distribution, which we can calculate more easily. For notation, let V be the parameters of the variational distribution.

$$V^\star = \arg\min_V KL(Q(Z|V)||P(Z|X)$$

Reference : Kullback-Leibler Divergence

To measure the difference between two probability distributions over the same variable x a measure, called the Kullback-Leibler divergence. The KL divergence, which is closely related to relative entropy, information divergence, and information for discrimination, is a non-symmetric measure of the difference between two probability distributions p(x) and q(x).

Specifically, the Kullback-Leibler (KL) divergence of q(x) from p(x), denoted $D_{KL}(p(x), q(x))$, is a measure of the information lost when q(x) is used to approximate p(x)

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Typically p(x) represents the “true” distribution of data, observations, or a precisely calculated theoretical distribution. The measure q(x) typically represents a theory, model, description, or approximation of p(x).

KL- divergence is a non-symmetric measure of the difference between two probability distribution q and p.

Reference : Kullback-Leibler Divergence

To measure the difference between two probability distributions over the same variable x a measure, called the Kullback-Leibler divergence. The KL divergence, which is closely related to relative entropy, information divergence, and information for discrimination, is a non-symmetric measure of the difference between two probability distributions p(x) and q(x).

Specifically, the Kullback-Leibler (KL) divergence of q(x) from p(x), denoted $D_{KL}(p(x), q(x))$, is a measure of the information lost when q(x) is used to approximate p(x)

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Typically p(x) represents the “true” distribution of data, observations, or a precisely calculated theoretical distribution. The measure q(x) typically represents a theory, model, description, or approximation of p(x).

KL- divergence is a non-symmetric measure of the difference between two probability distribution q and p.

Parameter

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/introductory-concepts/basic-concepts/parameter-estimates/

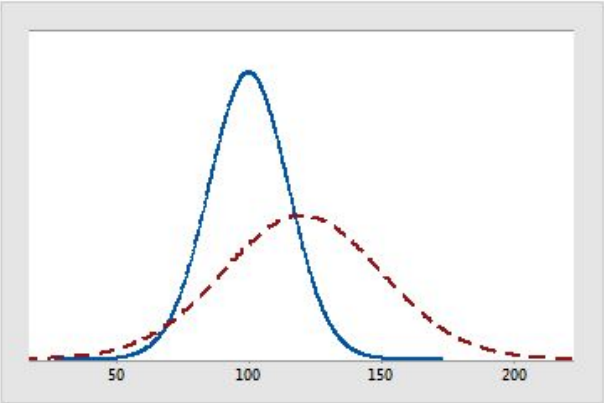
arameters are descriptive measures of an entire population used as the inputs for a probability distribution function (PDF) to generate distribution curves. Parameters are usually signified by Greek letters to distinguish them from sample statistics. For example, the population mean is represented by the Greek letter mu (μ) and the population standard deviation by the Greek letter sigma (σ). Parameters are fixed constants, that is, they do not vary like variables. However, their values are usually unknown because it is infeasible to measure an entire population.

Each distribution is entirely defined by several specific parameters, usually between one and three. The following table provides examples of the parameters required for three distributions. The parameter values determine the location and shape of the curve on the plot of distribution, and each unique combination of parameter values produces a unique distribution curve.

parameter

Distribution	Parameter 1	Parameter 2	Parameter 3
Chi-square	Degrees of freedom		
Normal	Mean	Standard deviation	
3-Parameter Gamma	Shape	Scale	Threshold

For example, a normal distribution is defined by two parameters, the mean and standard deviation. If these are specified, the entire distribution is precisely known.



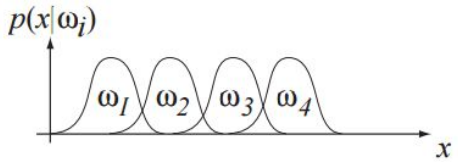
The solid line represents a normal distribution with a mean of 100 and a standard deviation of 15. The dashed line is also a normal distribution, but it has a mean of 120 and a standard deviation of 30.

Reference : Generative model

https://www.ee.columbia.edu/~dpwe/e6820/lectures/L03-ml.pdf

Generative models

- Describe the data using structured probabilistic models
- Observations are random variables whose distribution depends on model parameters
- Source distributions $p(x \mid \theta_i)$
  - reflect variability in features
  - reflect noise in observation
  - generally have to be estimated from data (rather than known in advance)



Three things to do with generative models

- Evaluate the probability of an observation, possibly under multiple parameter settings

$$p(x), \quad p(x \mid \theta_1), \quad p(x \mid \theta_2), \quad \ldots$$

- Estimate model parameters from observed data

$$\hat{\theta} = \operatorname*{argmin}_{\theta} C(\theta^*, \theta \mid x)$$

- Run the model forward to generate new data

$$\tilde{x} \sim p(x \mid \hat{\theta})$$

Binomial and Multinomial

Binomial distribution: the number of successes in a sequence of independent *yes/no* experiments (Bernoulli trials).

$$P(X = x \mid n, p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

Multinomial: suppose that each experiment results in one of k *possible outcomes* with probabilities $p_1, \ldots, p_k$; Multinomial models the distribution of the histogram vector which indicates how many time each outcome was observed over N trials of experiments.

$$P(x_1, \ldots, x_k \mid n, p_1, \ldots, p_k) = \frac{N!}{\prod_{i=1}^{k} x_i!} p_i^{x_i}, \quad \sum_i x_i = N, x_i \geq 0$$

https://www.cs.cmu.edu/~epxing/Class/10701-08s/recitation/dirichlet.pdf

(Reference :Mixture Model)

Simple distribution : Bernoulli or Gaussian
Often the data we're trying to model is much more complex. (e.g. multimodal. This means that there are several different modes, or regions of high probability mass, and regions of smaller probability mass in between.)

In case when the distribution has two modes. In this situation, we might model the data in terms of a mixture of several components, where each component has a simple parametric form (such as a Gaussian).
n other words, we assume each data point belongs to one of the components, and we try to infer the distribution for each component separately. The model itself doesn't "know" anything about cities (latent variable), In general, we won't always know precisely what meaning should be attached to the latent variables. In order to represent this mathematically, we formulate the model in terms of latent variables, usually denoted z. These are variables which are never observed, and where we don't know the correct values in advance. They are roughly analogous to hidden units, in that the learning algorithm needs to figure out what they should represent, without a human specifying it by hand. Variables which are always observed, or even sometimes observed, are referred to as observables. In the above example, the city is the latent variable and the temperature is the observable.

(Reference :Mixture Model)

In mixture models, the latent variable corresponds to the mixture component. It takes values in a discrete set, which we'll denote {1, . . . , K}. (For now, assume K is fixed; we'll talk later about how to choose it.) In general, a mixture model assumes the data are generated by the following process:

first we sample z,

and then we sample the observables x from a distribution which depends on z,

i.e. $p(z, x) = p(z)\, p(x \mid z)$.

In mixture models, p(z) is always a multinomial distribution.

p(x | z) can take a variety of parametric forms, If we assume it's a Gaussian distribution. We refer to such a model as a mixture of Gaussians

(Reference :Mixture Model)

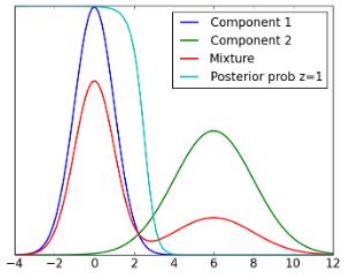


Figure 2: An example of a univariate mixture of Gaussians model.

For the general case,

$$z \sim \text{Multinoimal}(\boldsymbol{\pi}) \tag{4}$$

$$x \,|\, z = k \sim \text{Gaussian}(\mu_k, \sigma_k). \tag{5}$$

Here, $\boldsymbol{\pi}$ is a vector of probabilities (i.e. nonnegative values which sum to 1) known as the **mixing proportions**.

In general, we can compute the probability density function (PDF) over $\mathbf{x}$ by **marginalizing out**, or summing out, z:

$$p(\mathbf{x}) = \sum_z p(z)\, p(\mathbf{x} \,|\, z) \tag{6}$$

$$= \sum_{k=1}^{K} \Pr(z = k)\, p(\mathbf{x} \,|\, z = k) \tag{7}$$

Beta distribution

Beta Distribution

$$p(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- $p \in [0, 1]$: considering p as the parameter of a Binomial distribution, we can think of Beta is a "distribution over distributions" (binomials).
- Beta function simply defines binomial coefficient for continuous variables. (likewise, Gamma function defines factorial in continuous domain.)

$$B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \simeq \begin{pmatrix} \alpha - 1 \\ \alpha + \beta - 2 \end{pmatrix}$$

- Beta is the conjugate prior of Binomial.

Bayesian Model

Bayesian hierarchical modelling is a statistical model written in multiple levels (hierarchical form) that estimates the parameters of the posterior distribution using the Bayesian method. The sub-models combine to form the hierarchical model, and the Bayes' theorem is used to integrate them with the observed data, and account for all the uncertainty that is present. The result of this integration is the posterior distribution, also known as the updated probability estimate, as additional evidence on the prior distribution is acquired.

In Bayesian statistics, the **posterior probability** of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence or background is taken into account. Similarly, the **posterior probability distribution** is the probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey. "Posterior", in this context, means after taking into account the relevant evidence related to the particular case being examined.

The posterior probability is the probability of the parameters θ given the evidence X: $p(\theta|X)$.

It contrasts with the likelihood function, which is the probability of the evidence given the parameters: $p(X|\theta)$.

The two are related as follows:

Let us have a prior belief that the probability distribution function is $p(\theta)$ and observations x with the likelihood $p(x|\theta)$, then the posterior probability is defined as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$[1]

The posterior probability can be written in the memorable form as

$$\text{Posterior probability} \propto \text{Likelihood} \times \text{Prior probability}.$$

approximate inference techniques

Approximate inference methods make it possible to learn realistic models from big data by trading off computation time for accuracy, when exact learning and inference are computationally intractable.

Major methods classes[edit]

- Variational Bayesian methods
- Expectation propagation
- Markov random fields
- Bayesian networks
    - Variational message passing
- loopy and generalized belief propagation

variational methods

http://www.mit.edu/~9.520/spring11/slides/class19_approxinf.pdf

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	Likelihood function of θ
$P(\theta)$	Prior probability of θ
$P(\theta\|\mathcal{D})$	Posterior distribution over θ

Variational Inference

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$.

We can lower bound the marginal likelihood using Jensen's inequality:

$$\begin{aligned}
\ln p(\mathcal{D}) &= \ln \int p(\mathcal{D}, \theta) d\theta = \ln \int q(\theta) \frac{P(\mathcal{D}, \theta)}{q(\theta)} d\theta \\
&\geq \int q(\theta) \ln \frac{p(\mathcal{D}, \theta)}{q(\theta)} d\theta = \underbrace{\int q(\theta) \ln p(\mathcal{D}, \theta) d\theta + \underbrace{\int q(\theta) \ln \frac{1}{q(\theta)} d\theta}_{\text{Entropy functional}}}_{\text{Variational Lower-Bound}} \\
&= \ln p(\mathcal{D}) - \mathrm{KL}(q(\theta)||p(\theta|D)) = \mathcal{L}(q)
\end{aligned}$$

where $\mathrm{KL}(q||p)$ is a Kullback–Leibler divergence. It is a non-symmetric measure of the difference between two probability distributions q and p.

The goal of variational inference is to maximize the variational lower-bound w.r.t. approximate q distribution, or minimize $\mathrm{KL}(q||p)$.

EM algorithm

https://www.ee.columbia.edu/~dpwe/e6820/lectures/L03-ml.pdf

Expectation-maximization (EM)

- General procedure for estimating model parameters when some are unknown
  e.g. which GMM component generated a point
- Iteratively updated model parameters θ to maximize Q, the expected log-probability of observed data x and hidden data z

$$Q(\theta, \theta_t) = \int_z p(z \mid x, \theta_t) \log p(z, x \mid \theta)$$

  - E step: calculate $p(z \mid x, \theta_t)$ using θ_t
  - M step: find θ that maximizes Q using $p(z \mid x, \theta_t)$
  - can prove $p(x \mid \theta)$ non-decreasing
  - hence maximum likelihood model
  - local optimum—depends on initialization

empirical Bayes parameter estimation

Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data. This approach stands in contrast to standard Bayesian methods, for which the prior distribution is fixed before any data are observed. Despite this difference in perspective, empirical Bayes may be viewed as an approximation to a fully Bayesian treatment of a hierarchical model wherein the parameters at the highest level of the hierarchy are set to their most likely values, instead of being integrated out. Empirical Bayes, also known as maximum marginal likelihood,[1] represents one approach for setting hyperparameters.

SVD (singular value decomposition)

Reference : https://datajobs.com/data-science-repo/SVD-Tutorial-%5BKirk-Baker%5D.pdf (This great reference has matrix example to calculate the SVD)

Singular value decomposition (SVD) can be looked at from three mutually compatible points of view.

On the one hand, we can see it as a method for transforming correlated variables into a set of uncorrelated ones that better expose the various relationships among the original data items. At the same time, SVD is a method for identifying and ordering the dimensions along which data points exhibit the most variation. This ties into the third way of viewing SVD, which is that once we have identied where the most variation is, it's possible to find the best approximation of the original data points using fewer dimensions. Hence, SVD can be seen as a method for data reduction.

These are the basic ideas behind SVD: taking a high dimensional, highly variable set of data points and reducing it to a lower dimensional space that exposes the substructure of the original data more clearly and orders it from most variation to the least. What makes SVD practical for NLP applications is that you can simply ignore variation below a particular threshhold to massively reduce your data but be assured that the main relationships of interest have been preserved

Reduced singular value decomposition is the mathematical technique underlying a type of document retrieval and word similarity method variously called Latent Semantic Indexing or Latent Semantic Analysis.

The insight underlying the use of SVD for these tasks is that it takes the original data, usually consisting of some variant of a word × document matrix, and breaks it down into linearly independent components. These components are in some sense an abstraction away from the noisy correlations found in the original data to sets of values that best approximate the underlying structure of the dataset along each dimension independently. Because the majority of those components are very small, they can be ignored, resulting in an approximation of the data that contains substantially fewer dimensions than the original. SVD has the added benet that in the process of dimensionality reduction, the representation of items that share substructure become more similar to each other, and items that were dissimilar to begin with may become more dissimilar as well. In practical terms, this means that documents about a particular topic become more similar even if the exact same words don't appear in all of them.

Latent Semantic Indexing

Reference : http://www1.se.cuhk.edu.hk/~seem5680/lecture/LSI-Eg.pdf

mixture model

Reference :

Notation for Mixture Model : https://en.wikipedia.org/wiki/Mixture_model

Explanation > http://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf

In general, our strategy for unsupervised learning will be to formulate a probabilistic model which postulates certain unobserved random variables (called latent variables) which correspond to things we're interested in inferring. We then try to infer the values of the latent variables for all the data points, as well as parameters which relate the latent variables to the observations. In this lecture, we'll look at one type of latent variable model, namely mixture models.

Mixture Model (good definition)

http://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf

we looked at some methods for learning probabilistic models which took the form of simple distributions (e.g. Bernoulli or Gaussian). But often the data we're trying to model is much more complex. For instance, it might be multimodal. This means that there are several different modes, or regions of high probability mass, and regions of smaller probability mass in between.

For instance, suppose we've collected the high temperatures for every day in March 2014 for both Toronto and Miami, Florida, but forgot to write down which city is associated with each temperature. The values are plotted in Figure 1; we can see the distribution has two modes. In this situation, we might model the data in terms of a mixture of several components, where each component has a simple parametric form (such as a Gaussian).

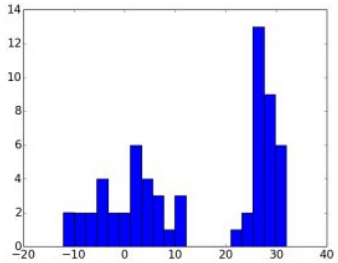


Figure 1: A histogram of daily high temperatures in °C for Toronto and Miami in March 2014. The distribution clearly has two modes.

Mixture Model (continue)

In other words, we assume each data point belongs to one of the components, and we try to infer the distribution for each component separately. In this example, we happen to know that the two mixture components should correspond to the two cities. The model itself doesn't "know" anything about cities, though: this is just something we would read into it at the very end, when we analyze the results. In general, we won't always know precisely what meaning should be attached to the latent variables. In order to represent this mathematically, we formulate the model in terms of latent variables, usually denoted z. These are variables which are never observed, and where we don't know the correct values in advance. They are roughly analogous to hidden units, in that the learning algorithm needs to figure out what they should represent, without a human specifying it by hand. Variables which are always observed, or even sometimes observed, are referred to as observables. In the above example, the city is the latent variable and the temperature is the observable. In mixture models, the latent variable corresponds to the mixture component. It takes values in a discrete set, which we'll denote $\{1, \ldots, K\}$. (For now, assume K is fixed; we'll talk later about how to choose it.) In general, a mixture model assumes the data are generated by the following process: first we sample z, and then we sample the observables x from a distribution which depends on z, i.e. $p(z, x) = p(z)\, p(x \mid z)$. In mixture models, p(z) is always a multinomial distribution. $p(x \mid z)$ can take a variety of parametric forms, but for this lecture we'll assume it's a Gaussian distribution. We refer to such a model as a mixture of Gaussians.

multinomial distribution

$$p = \frac{n!}{(n_1!)(n_2!)(n_3!)} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

The binomial distribution allows one to compute the probability of obtaining a given number of binary outcomes. For examp it can be used to compute the probability of getting 6 heads out of 10 coin flips. The flip of a coin is a binary outcome becau it has only two possible outcomes: heads and tails. The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes. For example, suppose that two chess players had played numerous games and it was determined that the probability that Player A would win is 0.40, the probability that Player B would win is 0.35, and the probability that the game would end in a draw is 0.25. The multinomial distribution can be used answer questions such as: "If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?" The following formula gives the probability of obtaining a specific set of outcomes when there are three possible outcomes for each event:

where

p is the probability,
n is the total number of events
n_1 is the number of times Outcome 1 occurs,
n_2 is the number of times Outcome 2 occurs,
n_3 is the number of times Outcome 3 occurs,
p_1 is the probability of Outcome 1
p_2 is the probability of Outcome 2, and
p_3 is the probability of Outcome 3.

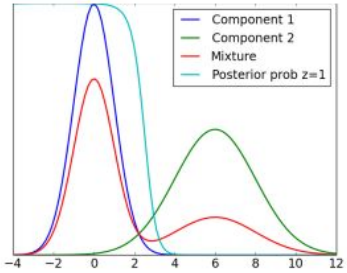


Figure 2: An example of a univariate mixture of Gaussians model.

Figure 2 shows an example of a mixture of Gaussians model with 2 components. It has the following generative process:

- With probability 0.7, choose component 1, otherwise choose component 2
- If we chose component 1, then sample x from a Gaussian with mean 0 and standard deviation 1
- If we chose component 2, then sample x from a Gaussian with mean 6 and standard deviation 2

This can be written in a more compact mathematical notation:

$$z \sim \text{Multinomial}(0.7, 0.3) \tag{1}$$
$$x \,|\, z = 1 \sim \text{Gaussian}(0, 1) \tag{2}$$
$$x \,|\, z = 2 \sim \text{Gaussian}(6, 2) \tag{3}$$

For the general case,

$$z \sim \text{Multinoimal}(\boldsymbol{\pi}) \tag{4}$$
$$x \,|\, z = k \sim \text{Gaussian}(\mu_k, \sigma_k). \tag{5}$$

Here, $\boldsymbol{\pi}$ is a vector of probabilities (i.e. nonnegative values which sum to 1) known as the **mixing proportions**.

In general, we can compute the probability density function (PDF) over $\mathbf{x}$ by **marginalizing out**, or summing out, z:

$$p(\mathbf{x}) = \sum_z p(z)\, p(\mathbf{x} \,|\, z) \tag{6}$$
$$= \sum_{k=1}^{K} \Pr(z = k)\, p(\mathbf{x} \,|\, z = k) \tag{7}$$

Note: Equations 6 and 7 are two different ways of writing the PDF; the first is more general (since it applies to other latent variable models), while the second emphasizes the meaning of the clustering model itself. In the example above, this gives us:

$$p(x) = 0.7 \cdot \text{Gaussian}(0, 1) + 0.3 \cdot \text{Gaussian}(6, 2). \tag{8}$$

This PDF is a **convex combination**, or **weighted average**, of the PDFs of the component distributions. The PDFs of the component distributions, as well as the mixture, are shown in Figure 2.

The general problem of grouping data points into clusters, where data points in the same cluster are more similar than data points in different clusters, is known as **clustering**. Learning a mixture model is one approach to clustering, but we should mention that there are a number of other approaches, most notably an algorithm called K-means[1].

Exchangeable random variables

In statistics, an **exchangeable sequence of random variables** (also sometimes **interchangeable**)[1] is a sequence such that future samples behave like earlier samples, meaning formally that any order (of a finite number of samples) is equally likely. This formalizes the notion of "the future being predictable on the basis of past experience." It is closely related to the use of independent and identically distributed random variables in statistical models. Exchangeable sequences of random variables arise in cases of simple random sampling.

https://users.soe.ucsc.edu/~draper/greenland-draper-1998.pdf

Reference : Exchangeability and de finetti theorem

http://www.stats.ox.ac.uk/~steffen/teaching/grad/definetti.pdf

N ~ Poisson(?)

A Poisson distribution is the probability distribution that results from a Poisson experiment.

A **Poisson experiment** is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures.
- The average number of successes (μ) that occurs in a specified region is known.
- The probability that a success will occur is proportional to the size of the region.
- The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

- *e*: A constant equal to approximately 2.71828. (Actually, *e* is the base of the natural logarithm system.)
- μ: The mean number of successes that occur in a specified region.
- *x*: The actual number of successes that occur in a specified region.
- P(*x*; μ): The **Poisson probability** that exactly *x* successes occur in a Poisson experiment, when the mean number of successes is μ.

Poisson Distribution

A **Poisson random variable** is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a **Poisson distribution**.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

Poisson Formula. Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ. Then, the Poisson probability is:

$$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$$

where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to μ .
- The variance is also equal to μ .

Possion example

Example 1

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- μ = 2; since 2 homes are sold per day, on average.
- x = 3; since we want to find the likelihood that 3 homes will be sold tomorr
- e = 2.71828; since *e* is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$

$P(3; 2) = (2.71828^{-2}) (2^3) / 3!$

$P(3; 2) = (0.13534) (8) / 6$

$P(3; 2) = 0.180$

Thus, the probability of selling 3 homes tomorrow is 0.180 .

2 The Poisson Distribution

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

If we let X = The number of events in a given interval,

Then, if the mean number of events per interval is λ

The probability of observing x events in a given interval is given by

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, 3, 4, \ldots$$

Note e is a mathematical constant. e ≈ 2.718282. There should be a button on your calculator e^x that calculates powers of e.

If the probabilities of X are distributed in this way, we write

$$X \sim Po(\lambda)$$

λ is the **parameter** of the distribution. We *say* X follows a Poisson distribution with parameter λ

Note A Poisson random variable can take on any positive integer value. In contrast, the Binomial distribution always has a finite upper limit.

http://www.stats.ox.ac.uk/~marchini/teaching/L5/L5.notes.pdf

Bernoulli distribution

Definition

Bernoulli random variables are characterized as follows.

Definition Let X be a discrete random variable. Let its support be

$$R_X = \{0, 1\}$$

Let $p \in (0, 1)$. We say that X has a **Bernoulli distribution** with parameter p if its probability mass function is

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{if } x \notin R_X \end{cases}$$

A random variable having a Bernoulli distribution is also called a Bernoulli random variable.

Note that, by the above definition, any indicator function is a Bernoulli random variable.

Beta Distribution

Good reference for Beta Distribution

https://www.statlect.com/probability-distributions/beta-distribution

The Beta distribution is a continuous probability distribution having two parameters. One of its most common uses is to model one's uncertainty about the probability of success of an experiment.

Beta Function : The Beta function is a function of two variables that is often found in probability theory and mathematical statistics (for example, as a normalizing constant in the probability density functions of the F distribution and of the Student's t distribution).

The following is a possible definition of the Beta function:

> **Definition** The **Beta function** is a function $B : \mathbb{R}^2_{++} \to \mathbb{R}$ defined as follows:
>
> $$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$
>
> where $\Gamma(\ \ )$ is the Gamma function.

Beta distribution

The Beta distribution is a continuous probability distribution having two parameters. One of its most common uses is to model one's uncertainty about the probability of success of an experiment.

Suppose a probabilistic experiment can have only two outcomes, either success, with probability X, or failure, with probability $1 - X$. Suppose also that X is unknown and all its possible values are deemed equally likely. This uncertainty can be described by assigning to X a uniform distribution on the interval $[0, 1]$. This is appropriate because X, being a probability, can take only values between 0 and 1; furthermore, the uniform distribution assigns equal probability density to all points in the interval, which reflects the fact that no possible value of X is, a priori, deemed more likely than all the others. Now, suppose that we perform n independent repetitions of the experiment and we observe k successes and $n - k$ failures. After performing the experiments, we naturally want to know how we should revise the distribution initially assigned to X, in order to properly take into account the information provided by the observed outcomes. In other words, we want to calculate the conditional distribution of X, conditional on the number of successes and failures we have observed. The result of this calculation is a Beta distribution. In particular, the conditional distribution of X, conditional on having observed k successes out of n trials, is a Beta distribution with parameters $k + 1$ and $n - k + 1$.

The Beta distribution is characterized as follows.

Definition Let X be an absolutely continuous random variable. Let its support be the unit interval:

$$R_X = [0, 1]$$

Let $\alpha, \beta \in \mathbb{R}_{++}$. We say that X has a **Beta distribution** with shape parameters α and β if its probability density function is

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where $B(\quad)$ is the Beta function.

Binomial Coefficient

http://math

Binomial Coefficient

DOWNLOAD Wolfram Notebook EXPLORE THIS TOPIC IN The MathWorld Classroom

The binomial coefficient $\binom{n}{k}$ is the number of ways of picking k *unordered* outcomes from n possibilities, also known as a combination or combinatorial number. The symbols ${}_nC_k$ and $\binom{n}{k}$ are used to denote a binomial coefficient, and are sometimes read as "n choose k."

$\binom{n}{k}$ therefore gives the number of k-subsets possible out of a set of n distinct items. For example, The 2-subsets of $\{1, 2, 3, 4\}$ are the six pairs $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, and $\{3, 4\}$, so $\binom{4}{2} = 6$. The number of lattice paths from the origin $(0, 0)$ to a point (a, b) is the binomial coefficient $\binom{a+b}{a}$ (Hilton and Pedersen 1991).

The value of the binomial coefficient for nonnegative n and k is given explicitly by

$$ {}_nC_k \equiv \binom{n}{k} \equiv \frac{n!}{(n-k)!\,k!},$$

where $z!$ denotes a factorial. Writing the factorial as a gamma function $z! = \Gamma(z+1)$ allows the binomial coefficient to be generalized to noninteger arguments (including complex x and y) as

$$\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\,\Gamma(x-y+1)}.$$

For nonnegative integer arguments, the gamma function reduces to factorials, leading to

$$\binom{n}{k} = \begin{cases} \dfrac{n!}{k!\,(n-k)!} & \text{for } 0 \le k < n \\ 0 & \text{otherwise,} \end{cases}$$

which is Pascal's triangle. Using the symmetry formula

$$\frac{\Gamma(s-a+1)}{\Gamma(s-b+1)} = (-1)^{b-a}\frac{\Gamma(b-s)}{\Gamma(a-s)}$$

for integer a, b and complex s, this definition can be extended to negative integer arguments, making it continuous at all integer arguments as well as continuous for all complex arguments except for negative integer x and noninteger y, in which case it is infinite (Kronenburg 2011). This definition, given by

$$\binom{n}{k} = \begin{cases} (-1)^k \binom{-n+k-1}{k} & \text{for } k \ge 0 \\ (-1)^{n-k} \binom{-k-1}{\quad} & \text{for } k \le n \end{cases}$$

Gamma function

https://www.statlect.com/mathematical-tools/gamma-function

The Gamma function is a generalization of the factorial function to non-integer numbers.

Recall that, if $n \in \mathbb{N}$, its factorial $n!$ is

$$n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$$

so that $n!$ satisfies the following recursion:

$$n! = (n-1)! \cdot n$$

The Gamma function $\Gamma(z)$ satisfies a similar recursion:

$$\Gamma(z) = \Gamma(z-1) \cdot (z-1)$$

but it is defined also when z is not an integer.

Dirichlet Distribution

$$p(P = \{p_i\} \mid \alpha_i) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \prod_i p_i^{\alpha_i - 1}$$

- $\sum_i p_i = 1, p_i \geq 0$
- Two parameters: the scale (or concentration) $\sigma = \sum_i \alpha_i$, and the base measure $(\alpha'_1, \ldots, \alpha'_k), \alpha'_i = \alpha_i / \sigma$.
- A generalization of Beta:
  - Beta is a distribution over binomials (in an interval $p \in [0, 1]$);
  - Dirichlet is a distribution over Multinomials (in the so-called simplex $\sum_i p_i = 1$; $p_i \geq 0$).
- Dirichlet is the conjugate prior of multinomial.

Simplex

Simplex

From Wikipedia, the free encyclopedia

For other uses, see Simplex (disambiguation).

In geometry, a **simplex** (plural: **simplexes** or **simplices**) is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions. Specifically, a ***k*-simplex** is a *k*-dimensional polytope which is the convex hull of its *k* + 1 vertices. More formally, suppose the *k* + 1 points $u_0, \ldots, u_k \in \mathbb{R}^k$ are affinely independent, which means $u_1 - u_0, \ldots, u_k - u_0$ are linearly independent. Then, the simplex determined by them is the set of points

$$C = \left\{ \theta_0 u_0 + \cdots + \theta_k u_k \;\middle|\; \sum_{i=0}^{k} \theta_i = 1 \text{ and } \theta_i \geq 0 \text{ for all } i \right\}.$$

For example, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and a 4-simplex is a 5-cell. A single point may be considered a 0-simplex, and a line segment may be considered a 1-simplex. A simplex may be defined as the smallest convex set containing the given vertices.

A **regular simplex**[1] is a simplex that is also a regular polytope. A regular *n*-simplex may be constructed from a regular (*n* − 1)-simplex by connecting a new vertex to all original vertices by the common edge length.

In topology and combinatorics, it is common to "glue together" simplices to form a simplicial complex. The associated combinatorial structure is called an abstract simplicial complex, in which context the word "simplex" simply means any finite set of vertices.

A regular 3-simplex or tetrahedron

Probability density function

Reference : https://www.statlect.com/glossary/probability-density-function

The distribution of a continuous random variable can be characterized through its probability density function (**pdf**). The probability that a continuous random variable takes a value in a given interval is equal to the integral of its probability density function over that interval, which in turn is equal to the area of the region in the xy-plane bounded by the x-axis, the pdf and the vertical lines corresponding to the boundaries of the interval.

Definition The probability density function of a continuous random variable X is a function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$\mathrm{P}(X \in [a, b]) = \int_a^b f_X(x)dx$$

for any interval $[a, b] \subseteq \mathbb{R}$.

The set of values x for which $f_X(x) > 0$ is called the support of X.

Example

Suppose a random variable X has probability density function

$$f_X(x) = \begin{cases} \frac{1}{4}x & \text{if } x \in [1, 3] \\ 0 & \text{otherwise} \end{cases}$$

To compute the probability that X takes a value in the interval $[1, 2]$,you need to integrate the probability density function over that interval:

$$\begin{aligned} \mathrm{P}(X \in [1, 2]) &= \int_1^2 f_X(x)dx \\ &= \int_1^2 \frac{1}{4}xdx \\ &= \left[\frac{1}{8}x^2\right]_1^2 \\ &= \frac{4}{8} - \frac{1}{8} = \frac{3}{8} \end{aligned}$$

Conjugate Prior

Good reference : http://lesswrong.com/lw/5sn/the_joys_of_conjugate_priors/

Greek alphabet : http://www.rapidtables.com/math/symbols/greek_alphabet.htm

Exponent

http://www.cs.c

http://www.cs.princeton.

Exponential family comprises a set of flexible distribution ranging both continuous and discrete random variables. The members of this family have many important properties which merits discussing them in some general format. Many of the probability distributions that we have studied so far are specific members of this family:

- Gaussian: $\mathbb{R}^p$
- Multinomial: *categorical*
- Bernoulli: binary $\{0, 1\}$
- Binomial: counts of success/failure
- Von mises: sphere
- Gamma: $\mathbb{R}^+$
- Poisson: $\mathbb{N}^+$
- Laplace: $\mathbb{R}^+$
- Exponential: $\mathbb{R}^+$
- Beta: $(0, 1)$
- Dirichlet: Δ (Simplex)
- Weibull: $\mathbb{R}^+$
- Weishart: symmetric positive-definite matrices

All these distributions follow the general format:

$$p(x|\eta) = h(x) \exp\left(\eta^\top t(x) - a(\eta)\right); \tag{1}$$

where, η is called "natural parameter", $t(x)$ is "sufficient statistic" (a statistic is a function of data), $h(x)$ is the "underlying measure" and $a(\eta)$ is called "log normalizer", which ensures that the distribution integrates to one. Hence,

$$a(\eta) = \log \int h(x) \exp\left(\eta^\top t(x)\right) dx.$$

Conjugate prior, prior distribution, posterior distribution

First off, let's talk about the multinomial distribution. To keep things simple, we'll consider the case of three dimensions d=3. Suppose X is a vector which represents n draws of a random variable with three possible outcomes. e.g. when n=10, an example draw of X could be x = [4,4,2], which indicates outcome 1 occurred 4 times, outcome 2 occurred 4 times and outcome 3 occurred 2 times.

Assuming the probabilities of the outcome i is p_i, the multinomial distribution describes the prob. mass distribution of X.

$$\text{Multi}(n_1, n_2, n_3; p_1, p_2, p_3) = \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

The expected value of X is $\vec{p} = (p_1, p_2, p_3)$, so if you want a probabilistic model to describe some real data that is 80% outcome 1, 19% outcome 2 and 1% outcome 3 you use the multinomial distribution with parameters $\vec{p} = [0.8, 0.19, 0.01]$.
A simple learning task would be to learn the inverse mapping: given some data X, estimate (learn) the \vec{p} which generated X.

The Dirichlet distribution is a probability distribution over the space of multinomial distributions, i.e., to generate data X from a Dirichlet distribution with parameters $\alpha_1, \alpha_2, \alpha_3$, you first draw a $\vec{p} \sim Dir(\vec{\alpha})$, and then draw $X \sim Multi(\vec{p})$.
We have introduced one level of indirection in the model for X: instead of saying directly what $\vec{p}$ generated X, we choose the hyperparameters $\alpha_1, \alpha_2, \alpha_3$ which dictate what prob. distributions are likely to occur, and then draw samples X according to the random $\vec{p}$s.

The addition of the Dirichlet distribution to the model allows us to specify our prior beliefs about what data X is likely to occur. The *Bayesian* learning problem is now to estimate $\vec{p}$ from X, given our prior beliefs $\alpha_1, \alpha_2, \alpha_3$.

Now for the intuition part. How do we encode our intuition into the hyperparameters $\alpha_1, \alpha_2, \alpha_3$. There are two regimes to consider $\alpha_i \geq 1$ and $\alpha_i < 1$. In the first case, the alphas represent pseudo counts --- when trying to learn \vec{p} from data X and a prior with $\alpha_1 = 20$, you are saying that the observed count of outcome 1 is n_1 + 20. Think of this as padding the counts... (Accountants pad the accounts, Bayesians pad the counts).

Since you are asking the LDA channel, I will assume you are more interested in the case $\alpha_i < 1$. Unfortunately, there is no intuition to be had here! The only thing setting small hyperparameters does is enforce sparsity in the multinomial distribution. The Dirichlet prior in LDA can therefore be interpreted as a regularization method, like \ell_1 normalization.

Oh and in since I already invested time in typesetting this, here is the definition of the Dirichlet distribution:

$$\text{Dir}(\vec{p}; \vec{\alpha}) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1}$$

About Dirichlet Distribution
(Very good description)

https://www.quora.com/What-is-an-intuitive-explanation-of-the-Dirichlet-distribution

Multinomial distribution

https://www.statlect.com/probability-distributions/multinomial-distribution

The multinomial distribution is a generalization of the binomial distribution. If you perform n times an experiment that can have only two outcomes (either success or failure), then the number of times you obtain one of the two outcomes (success) is a binomial random variable. If you perform n times an experiment that can have K outcomes (K can be any natural number) and you denote by X_i the number of times that you obtain the i-th outcome, then the random vector X defined as

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_K \end{bmatrix}$$

is a multinomial random vector.

A multinomial distribution can be seen as a sum of mutually independent Multinoulli random variables. This connection between the multinomial and Multinoulli distributions will be illustrated in detail in the rest of this lecture and will be used to demonstrate several properties of the multinomial distribution. For this reason, it is highly recommended to read the lecture entitled Multinoulli distribution before reding the following sections.

Definition

Multinomial random vectors are characterized as follows.

Definition Let X be a $K \times 1$ discrete random vector. Let $n \in \mathbb{N}$. Let the support of X be the set of $K \times 1$ vectors having non-negative integer entries summing up to n:

$$R_X = \left\{ x \in \{0, 1, 2, \dots, n\}^K : \sum_{i=1}^{K} x_i = n \right\}$$

Let $p_1, \dots, p_K$ be K strictly positive numbers such that

$$\sum_{i=1}^{K} p_i = 1$$

We say that X has a **multinomial distribution** with probabilities $p_1, \dots, p_K$ and number of trials n, if its joint probability mass function is

$$p_X(x_1, \dots, x_K) = \begin{cases} \binom{n}{x_1, x_2, \dots, x_K} \prod_{i=1}^{K} p_i^{x_i} & \text{if } (x_1, \dots, x_K) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

where $\binom{n}{x_1, x_2, \dots, x_K}$ is the multinomial coefficient.

Related with Multinomial Distribution

Multinoulli distribution

The Multinoulli distribution (sometimes also called categorical distribution) is a generalization of the Bernoulli distribution. If you perform an experiment that can have only two outcomes (either success or failure), then a random variable that takes value 1 in case of success and value 0 in case of failure is a Bernoulli random variable. If you perform an experiment that can have K outcomes and you denote by X_i a random variable that takes value 1 if you obtain the i-th outcome and 0 otherwise, then the random vector X defined as

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_K \end{bmatrix}$$

is a Multinoulli random vector. In other words, when the i-th outcome is obtained, the i-th entry of the Multinoulli random vector X takes value 1, while all other entries take value 0.

In what follows the probabilities of the K possible outcomes will be denoted by $p_1, \dots, p_K$.

(continue of multinoulli distribution)

Definition

The distribution is characterized as follows.

Definition Let X be a $K \times 1$ discrete random vector. Let the support of X be the set of $K \times 1$ vectors having one entry equal to 1 and all other entries equal to 0:

$$R_X = \left\{ x \in \{0,1\}^K : \sum_{j=1}^{K} x_j = 1 \right\}$$

Let $p_1, \ldots, p_K$ be K strictly positive numbers such that

$$\sum_{j=1}^{K} p_j = 1$$

We say that X has a **Multinoulli distribution** with probabilities $p_1, \ldots, p_K$ if its joint probability mass function is

$$p_X(x_1, \ldots, x_K) = \begin{cases} \prod_{j=1}^{K} p_j^{x_j} & \text{if } (x_1, \ldots, x_K) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

If you are puzzled by the above definition of the joint pmf, note that when $(x_1, \ldots, x_K) \in R_X$ and $x_i = 1$ because the i-th outcome has been obtained, then all other entries are equal to 0 and

$$\begin{aligned} \prod_{j=1}^{K} p_j^{x_j} &= p_1^{x_1} \cdot \ldots \cdot p_{i-1}^{x_{i-1}} \cdot p_i^{x_i} \cdot p_{i+1}^{x_{i+1}} \cdot \ldots \cdot p_K^{x_K} \\ &= p_1^0 \cdot \ldots \cdot p_{i-1}^0 \cdot p_i^1 \cdot p_{i+1}^0 \cdot \ldots \cdot p_K^0 \\ &= 1 \cdot \ldots \cdot 1 \cdot p_i^1 \cdot 1 \cdot \ldots \cdot 1 \\ &= p_i \end{aligned}$$

Hyperparameter

In Bayesian statistics, a **hyperparameter** is a parameter of a prior distribution; the term is used to distinguish them from parameters of the model for the underlying system under analysis.

For example, if one is using a beta distribution to model the distribution of the parameter *p* of a Bernoulli distribution, then:

- *p* is a parameter of the underlying system (Bernoulli distribution), and
- *α* and *β* are parameters of the prior distribution (beta distribution), hence *hyper*parameters

(* prior distribution : In Bayesian statistical inference, a **prior probability distribution**, often simply called the **prior**, of an uncertain quantity is the probability distribution that **would express one's beliefs about this quantity before some evidence is taken into** account. For example, the prior could be the probability distribution representing the relative proportions of voters who will vote for a particular politician in a future election. The unknown quantity may be a parameter of the model or a latent variable rather than an observable variable.)

(* Beta distribution : In probability theory and statistics, the **beta distribution** is a family of continuous probability distributions defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by *α* and *β*, that appear as exponents of the random variable and control the shape of the distribution.In probability theory and statistics, the **beta distribution** is a family of continuous probability distributions defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by *α* and *β*, that appear as exponents of the random variable and control the shape of the distribution.

Beta distribution

The beta distribution has been applied to model the behavior of random variables limited to intervals of finite length in a wide variety of disciplines.

In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions. For example, the beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success such as the probability that a space vehicle will successfully complete a specified mission. The beta distribution is a suitable model for the random behavior of percentages and proportions.)

Probability density function [edit]

The probability density function (pdf) of the beta distribution, for $0 \le x \le 1$, and shape parameters $\alpha, \beta > 0$, is a power function of the variable x and of its reflection $(1 - x)$ as follows:

$$\begin{aligned} f(x;\alpha,\beta) &= \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$

where $\Gamma(z)$ is the gamma function. The beta function, B, is a normalization constant to ensure that the total probability integrates to 1. In the above equations x is a realization—an observed value that actually occurred—of a random process X.

Conjugate prior probability distribution

In Bayesian probability theory, 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 **conjugate distributions,** and the prior is called a **conjugate prior** for the likelihood function. For example, the Gaussian family is conjugate to itself (or *self-conjugate*) with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian. This means that the Gaussian distribution is a conjugate prior for the likelihood that is also Gaussian.

Consider the general problem of inferring a distribution for a parameter θ given some datum or data x. From Bayes' theorem, the posterior distribution is equal to the product of the likelihood function $\theta \mapsto p(x \mid \theta)$ and prior $p(\theta)$, normalized (divided) by the probability of the data $p(x)$:

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{\int p(x \mid \theta')\, p(\theta')\, d\theta'}.$$

Let the likelihood function be considered fixed; the likelihood function is usually well-determined from a statement of the data-generating process. It is clear that different choices of the prior distribution $p(\theta)$ may make the integral more or less difficult to calculate, and the product $p(x|\theta) \times p(\theta)$ may take one algebraic form or another. For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Such a choice is a *conjugate prior*.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise numerical integration may be necessary. Further, conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution.

All members of the exponential family have conjugate priors.[3]

Exponential Family Distribution

Exponential Family : In probability and statistics, an **exponential family** is a set of probability distributions of a certain form, specified below. This special form is chosen for mathematical convenience, on account of some useful algebraic properties, as well as for generality, as exponential families are in a sense very natural sets of distributions to consider.

Scalar parameter [edit]

A single-parameter exponential family is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form

$$f_X(x \mid \theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\theta))$$

where $T(x)$, $h(x)$, $\eta(\theta)$, and $A(\theta)$ are known functions.

An alternative, equivalent form often given is

$$f_X(x \mid \theta) = h(x)g(\theta)\exp(\eta(\theta)\cdot T(x))$$

or equivalently

$$f_X(x \mid \theta) = \exp(\eta(\theta)\cdot T(x) - A(\theta) + B(x))$$

The value θ is called the parameter of the family.

Note that x is often a vector of measurements, in which case $T(x)$ may be a function from the space of possible values of x to the real numbers. More generally, $\eta(\theta)$ and $T(x)$ can each be vector-valued such that $\boldsymbol{\eta(\theta)}' \cdot \boldsymbol{T(x)}$ is real-valued.

The exponential families include many of the most common distributions. Among many others, the family includes the following:

- normal
- exponential
- gamma
- chi-squared
- beta
- Dirichlet
- Bernoulli
- categorical
- Poisson
- Wishart
- Inverse Wishart

A number of common distributions are exponential families, but only when certain parameters are fixed and known. For example:

- binomial (with fixed number of trials)
- multinomial (with fixed number of trials)
- negative binomial (with fixed number of failures)

Notice that in each case, the parameters which must be fixed determine a limit on the size of observation values.

Examples of common distributions that are *not* exponential families are Student's *t*, most mixture distributions, and even the family of uniform distributions when the bounds are not fixed. See the section below on examples for more discussion.

Conjugate Prior

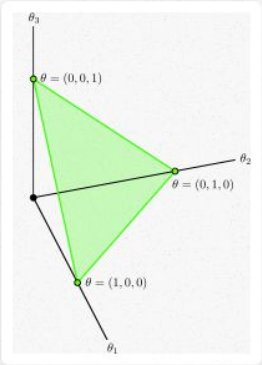
http://lesswrong.com/lw/5sn/the_joys_of_conjugate_priors/

(from paper) k-dimensional Dirichlet random variable theta can take values in (k-1)-simplex

For example with 3 outcomes dices :

http://blog.bogatron[illegible]utions/

$\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3\}$, we can treat θ_1, θ_2, and θ_3 each as an independent variable and $\boldsymbol{\theta}$ as a vector in a 3-dimensional space. Since the multinomial distribution requires that these three variables sum to 1, we know that the allowable values of $\boldsymbol{\theta}$ are confined to a plane. Furthermore, since each value θ_i must be greater than or equal to zero, the set of all allowable values of $\boldsymbol{\theta}$ is confined to an equilateral triangle (a 2-simplex) as shown below.



What we want is to know the probability density at each point on this triangle. That is where the Dirichlet distribution can help us. The Dirichlet distribution defines a probability density for a vector-valued input having the same characteristics as our multinomial parameter ($\boldsymbol{\theta}$). Let's start with the formula for the probability density associated with a Dirichlet distribution:

$$\text{Dir}\,(\boldsymbol{\alpha}) \rightarrow \text{p}\,(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma\,(\boldsymbol{\alpha}_i)} \prod_{i=1}^{k} \boldsymbol{\theta}_i^{\boldsymbol{\alpha}_i - 1}$$

De Finetti Theorem ??

Reference :
https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture1.pdf

http://stats.stackexchange.com/questions/34465/what-is-so-cool-about-de-finettis-representation-theorem

http://wwwf.imperial.ac.uk/~das01/MyWeb/M3S3/Handouts/DeFinetti.pdf

Marginalizing

Marginalizing amounts to integrating or summing out from joint distributions of several variables to the distribution of a smaller number of variables, usually to one variable. So for example if X and Y have the joint desnsity f(x,y) to get the maginal density for X you compute g(x)=∫f(x,y)dy integrated over all possible values of y. You do the analogous thing with sums for discrete variables.

Because you are computing a marginal distribution, that is, a marginal density. Say the density of (X,Z) is $f(x,z)$ then, integrating over z "marginalizing out z" gives

$$\int f(x,z)dz=f(x)$$

where $f(x)$ is the marginal density of X.

Continuous mixture distribution & mixture weight

http://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/mixtureModels.pdf

1 Introduction

Most of the density functions we have considered so far are **unimodal**, that is, they have at most one peak. However, often we want to be able to represent densities with multiple modes. A common way to do this is to create a **mixture model**, which is a convex combination of pdf's:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|k) \tag{1}$$

where K is the (fixed) number of mixture component, π is a vector of **mixing weights**, and $p(x|k)$ are the densities for each component. We consider some examples below.

Exchangeable

Definition : Exchangeability
A *finite* sequence of random variables $X_1, X_2, \ldots, X_n$ is *(finitely) exchangeable* with (joint) probability measure P, if, for any permutation π of indices

$$P(X_1, X_2, \ldots, X_n) = P(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$$

For example, the random variables (X_1, X_2, X_3, X_4) are exchangeable if

$$P(X_1, X_2, X_3, X_4) = P(X_2, X_4, X_1, X_3) = P(X_1, X_3, X_2, X_4) = \cdots$$

An *infinite* sequence, $X_1, X_2, \ldots$, is *infinitely exchangeable* if any finite subset of the sequence is finitely exchangeable.

Parameter Estimation

The term *parameter estimation* refers to the process of using sample data (in reliability engineering, usually times-to-failure or success data) to estimate the parameters of the selected distribution. Several parameter estimation methods are available. This section presents an overview of the available methods used in life data analysis. More specifically, we start with the relatively simple method of Probability Plotting and continue with the more sophisticated methods of Rank Regression (or Least Squares), Maximum Likelihood Estimation and Bayesian Estimation Methods.

http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/introductory-concepts/basic-concepts/parameters/

Parameters are descriptive measures of an entire population. However, their values are usually unknown because it is infeasible to measure an entire population. Because of this, you can take a random sample from the population to obtain parameter estimates. One goal of statistical analyses is to obtain estimates of the population parameters along with the amount of error associated with these estimates. These estimates are also known as sample statistics. A fitted distribution line is a curve based on the parameter estimates instead of on the true parameter values.

There are several types of parameter estimates:

- Point estimates are the single, most likely value of a parameter. For example, the point estimate of population mean (the parameter) is the sample mean (the parameter estimate).
- Confidence intervals are a range of values likely to contain the population parameter.

For an example of parameter estimates, suppose you work for a spark plug manufacturer that is studying a problem in their spark plug gap. It would be too costly to measure every single spark plug that is made. Instead, you randomly sample 100 spark plugs and measure the gap in millimeters. The mean of the sample is 9.2. This is the point estimate for the population mean (μ), and it informs you that the most likely value for the average gap for *all* spark plugs is 9.2 You also create a 95% confidence interval for μ which is

Variational inference

https://www.quora.com/What-is-variational-inference

In many interesting statistical problems, we can't directly calculate the posterior because the normalization constant is intractable. This happens often in latent variable models. For example assume that X represents a set of observations and Z represents a set of latent variables. If we are interested in the posterior P(Z|X), we know that

$$P(Z|X)=P(Z,X)$$

$$\int_z P(Z,X)$$

but often times we can't calculate the denominator.

One popular approach is MCMC, where we can sample exactly from the true posterior distribution; however, convergence can be prohibitively slow if we have many parameters to sample. This is where variational inference comes in handy. Variational inference **seeks to approximate the true posterior, P(Z|X),** with an approximate variational distribution, which we can calculate more easily. For notation, let V be the parameters of the variational distribution.

Jensen’s inequality to obtain an adjustable lower bound on the log likelihood

Jensen’s Inequality :

Jensens's inequality concerns the expected value of convex and concave transformations of a random variable.

Statement

The following is a formal statement of the inequality.

> **Proposition** Let X be an integrable random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function such that
>
> $$Y = g(X)$$
>
> is also integrable. Then, the following inequality, called Jensen's inequality, holds:
>
> $$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$$

Proof

If the function g is strictly convex and X is not almost surely constant, then we have a strict inequality:

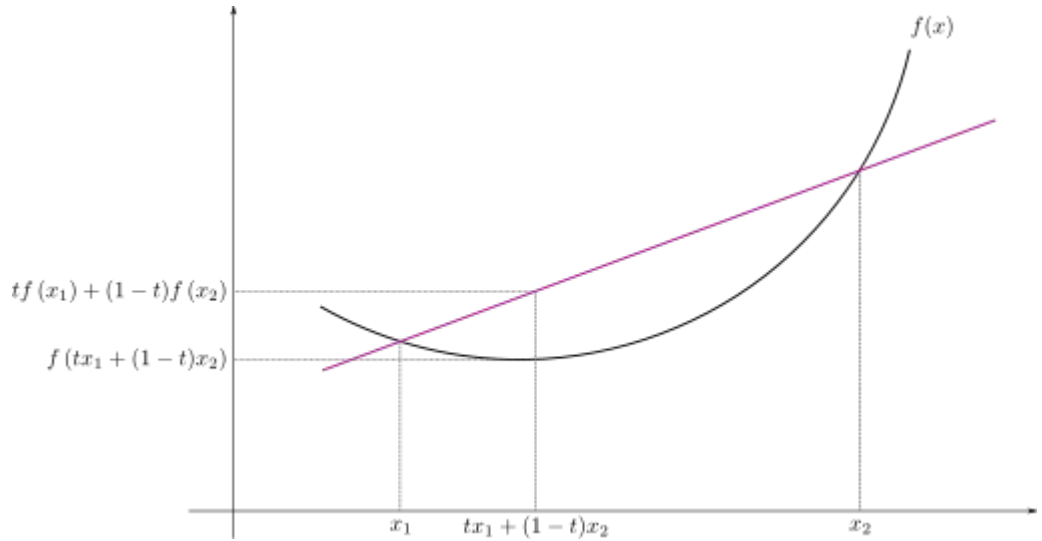
$$\mathrm{E}[g(X)] > g(\mathrm{E}[X])$$

Proof

If the function g is concave, then

$$\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$$

Jensen's inequality generalizes the statement that a secant line of a convex function lies above the graph.



Jensen's inequality generalizes the statement that the secant line of a convex function lies *above* the graph of the function, which is Jensen's inequality for two points: the secant line consists of weighted means of the convex function,

$$tf(x_1) + (1-t)f(x_2),$$

while the graph of the function is the convex function of the weighted means,

$$f\left(tx_1 + (1-t)x_2\right).$$

Thus, Jensen's inequality is

$$f\left(tx_1 + (1-t)x_2\right) \leq tf(x_1) + (1-t)f(x_2).$$

In the context of probability theory, it is generally stated in the following form: if X is a random variable and φ is a convex function, then

$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\varphi(X)\right].$$

Log likelihood

Likelihood Function :

In statistics, a **likelihood function** (often simply the **likelihood**) is a function of the parameters of a statistical model given data. Likelihood functions play a key role in statistical inference, especially methods of estimating a parameter from a set of statistics. In informal contexts, "likelihood" is often used as a synonym for "probability." In statistics, a distinction is made depending on the roles of outcomes vs. parameters. *Probability* is used before data are available to describe possible future outcomes given a fixed value for the parameter (or parameter vector). *Likelihood* is used after data are available to describe a function of a parameter (or parameter vector) for a given *outcome.*

Good reference : https://onlinecourses.science.psu.edu/stat504/node/27

One of the most fundamental concepts of modern statistics is that of likelihood. In each of the discrete random variables we have considered thus far, the distribution depends on one or more parameters that are, in most statistical applications, unknown. In the Poisson distribution, the parameter is λ. In the binomial, the parameter of interest is p (since n is typically fixed and known).

Likelihood is a tool for summarizing the data's evidence about unknown parameters. Let us denote the unknown parameter(s) of a distribution generically by θ. Since the probability distribution depends on θ, we can make this dependence explicit by writing f(x) as $f(x$; θ). For example, in the Bernoulli distribution the parameter is θ = π , and the distribution is

$$f(x;\pi)=\pi^{x}(1-\pi)^{1-x} \quad x=0,1$$

(2)

Loglikelihood

In most cases, for various reasons, but often computational convenience, we work with the loglikelihood

$$l(\theta|x) = \log L(\theta|x)$$

which is defined up to an arbitrary additive constant.

For example, the binomial loglikelihood is

$$l(\pi|x) = x \log \pi + (n - x) \log(1 - \pi).$$

In many problems of interest, we will derive our loglikelihood from a sample rather than from a single observation. If we observe an independent sample $x_1, x_2, \ldots, x_n$ from a distribution $f(x|\theta)$, then the overall likelihood is the product of the individual likelihoods:

$$\begin{aligned} L(\theta|x) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \prod_{i=1}^{n} L(\theta|x_i) \end{aligned}$$

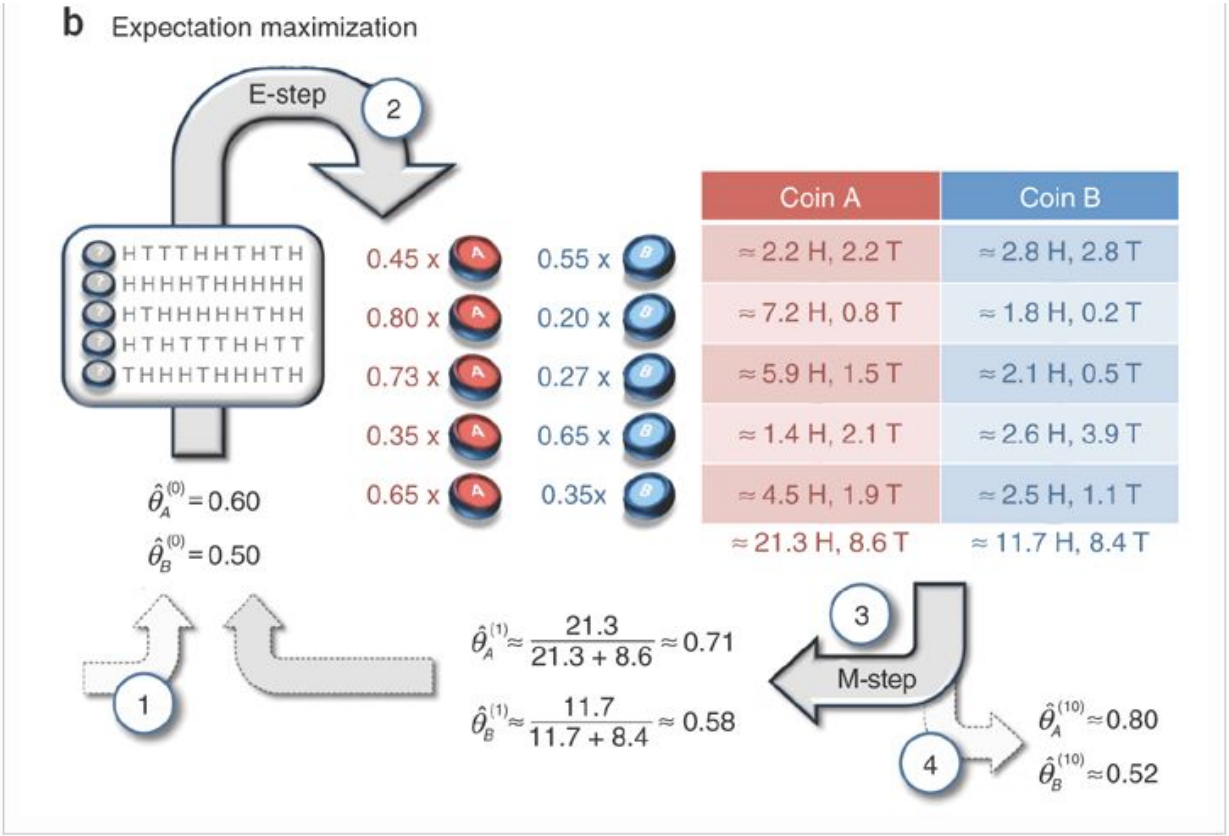
and the loglikelihood is:

$$\begin{aligned} l(\theta|x) &= \log \prod_{i=1}^{n} f(x_i|\theta) \\ &= \sum_{i=1}^{n} \log f(x_i|\theta) = \sum_{i=1}^{n} l(\theta|x_i). \end{aligned}$$

Variational parameter

Expectation Maximization algorithm

http://www.nature.com/nbt/journal/v26/n8/fig_tab/nbt1406_F1.html



(**a**) Maximum likelihood estimation. For each set of ten tosses, the maximum likelihood procedure accumulates the counts of heads and tails for coins A and B separately. These counts are then used to estimate the coin biases. (**b**) Expectation maximization. 1. EM starts with an initial guess of the parameters. 2. In the E-step, a probability distribution over possible completions is computed using the current parameters. The counts shown in the table are the expected numbers of heads and tails according to this distribution. 3. In the M-step, new parameters are determined using the current completions. 4. After several repetitions of the E-step and M-step, the algorithm converges.

Now consider a more challenging variant of the parameter estimation problem in which we are given the recorded head counts x but not the identities z of the coins used for each set of tosses. We refer to z as hidden variables or latent factors. Parameter estimation in this new setting is known as the incomplete data case. This time, computing proportions of heads for each coin is no longer possible, because we don't know the coin used for each set of tosses. However, if we had some way of completing the data (in our case, guessing correctly which coin was used in each of the five sets), then we could reduce parameter estimation for this problem with incomplete data to maximum likelihood estimation with complete data.
One iterative scheme for obtaining completions could work as follows: starting from some initial parameters, $^{(t)} = {}_{A}^{(t)}, {}_{B}^{(t)}$, determine for each of the five sets whether coin A or coin B was more likely to have generated the observed flips (using the current parameter estimates). Then, assume these completions (that is, guessed coin assignments) to be correct, and apply the regular maximum likelihood estimation procedure to get $^{(t+1)}$. Finally, repeat these two steps until convergence. As the estimated model improves, so too will the quality of the resulting completions.

Parametric Estimation problem

Large vocabulary size that is characteristic of many document corpora creates problems of sparsity. A new document is very likely to contain words that did not appear in any of the documents in a training corpus.

Before starting, let us recall the main elements of a parametric estimation problem:

- a sample ξ is used to make statements about the probability distribution that generated the sample;
- the sample ξ is regarded as the realization of a random vector Ξ, whose unknown joint distribution function, denoted by $F_{\Xi}(\xi)$, is assumed to belong to a set of distribution functions Φ, called statistical model;
- the model Φ is put into correspondence with a set $\Theta \subset \mathbb{R}^p$ of real vectors; Θ is called the parameter space and its elements $\theta \in \Theta$ are called parameters;
- the parameter associated with the unknown distribution function $F_{\Xi}(\xi)$ that actually generated the sample is denoted by θ_0 and it is called the true parameter (if several different parameters are put into correspondence with $F_{\Xi}(\xi)$, θ_0 can be any one of them);
- a predefined rule (a function) that associates a parameter estimate $\hat{\theta} = \hat{\theta}(\xi)$ to each ξ in the support of Ξ is called an estimator (the symbol $\hat{\theta}$ is often used to denote both the estimate and the estimator and the meaning is usually clear from the context).

Maximum Likelihood

The main elements of a maximum likelihood estimation problem are the following:

- a sample ξ, that we use to make statements about the probability distribution that generated the sample;
- the sample ξ is regarded as the realization of a random vector Ξ, whose distribution is unknown and needs to be estimated;
- there is a set $\Theta \subseteq \mathbb{R}^p$ of real vectors (called the parameter space) whose elements (called parameters) are put into correspondence with the possible distributions of Ξ; in particular:
  - if Ξ is a discrete random vector, we assume that its joint probability mass function $p_\Xi(\xi;\theta_0)$ belongs to a set of joint probability mass functions $p_\Xi(\xi;\theta)$ indexed by the parameter θ; when the joint probability mass function is considered as a function of θ for fixed ξ, it is called **likelihood** (or **likelihood function**) and it is denoted by
  $$L(\theta;\xi) = p_\Xi(\xi;\theta)$$
  - if Ξ is an absolutely continuous random vector, we assume that its joint probability density function $f_\Xi(\xi;\theta_0)$ belongs to a set of joint probability density functions $f_\Xi(\xi;\theta)$ indexed by the parameter θ; when the joint probability density function is considered as a function of θ for fixed ξ, it is called likelihood and it is denoted by
  $$L(\theta;\xi) = f_\Xi(\xi;\theta)$$
- we need to estimate the true parameter θ_0, which is associated with the unknown distribution that actually generated the sample (we rule out the possibility that several different parameters are put into correspondence with true distribution).

Maximum likelihood estimator

A maximum likelihood estimator $\widehat{\theta}$ of θ_0 is obtained as a solution of a maximization problem:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; \xi)$$

In other words, $\widehat{\theta}$ is the parameter that maximizes the likelihood of the sample ξ. $\widehat{\theta}$ is called the **maximum likelihood estimator** of θ.

In what follows, the symbol $\widehat{\theta}$ will be used to denote both a maximum likelihood estimator (a random variable) and a maximum likelihood estimate (a realization of a random variable): the meaning will be clear from the context.

The same estimator $\widehat{\theta}$ is obtained as a solution of

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ln[L(\theta; \xi)]$$

i.e., by maximizing the natural logarithm of the likelihood function. Solving this problem is equivalent to solving the original one, because the logarithm is a strictly increasing function. The logarithm of the likelihood is called **log-likelihood** and it is denoted by

$$l(\theta; \xi) = \ln[L(\theta; \xi)]$$

Binomial Distribution

For *n* trials, *k* successes, and success probability *p*:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$ Prob. mass function

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Estimation problem: If we observe *n* and *k*, **what is *p*?**

Laplace Smoothing

https://www.quora.com/Could-someone-explain-Laplacian-smoothing-or-1-up-smoothing

Suppose you are looking at outcomes of a die. Let us say you get the following outcomes of each number, in 10 throws:

One : 1

Two : 3

Three : 1

Four : 0

Five : 3

Six : 2

Now, the probabilities without the smoothing are

Prior and Posterior

Very good explanation : https://www.quora.com/What-is-the-difference-between-the-prior-and-the-posterior-in-statistics

Say you have a quantity of interest: θ.

1. The prior is a probability distribution that represents your uncertainty over θ**before** you have sampled any data and attempted to estimate it - usually denoted $\pi(\theta)$.
2. The posterior is a probability distribution representing your uncertainty over θ**after** you have sampled data - denoted $\pi(\theta|X)$. It is a conditional distribution because it conditions on the observed data.

From Bayes' theorem we relate the two:

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\sum_{\theta_i} f(X|\theta_i)\pi(\theta_i)}$$

Taking a step back, we widely accept that the process of having beliefs about the

Density Estimation

Let $X_1,\ldots,X_n$ be a sample from a distribution P with density p. The goal of nonparametric density estimation is to estimate p with as few assumptions about p as possible. We denote the estimator by $\widehat{p}$. The estimator will depend on a smoothing parameter h and choosing h carefully is crucial. To emphasize the dependence on h we sometimes write $\widehat{p}_h$.

Density estimation used for: regression, classification, clustering and unsupervised prediction. For example, if $\widehat{p}(x,y)$ is an estimate of $p(x,y)$ then we get the following estimate of the regression function:

$$\widehat{m}(x) = \int y\widehat{p}(y|x)dy$$

where $\widehat{p}(y|x) = \widehat{p}(y,x)/\widehat{p}(x)$. For classification, recall that the Bayes rule is

$$h(x) = I(p_1(x)\pi_1 > p_0(x)\pi_0)$$

where $\pi_1 = \mathbb{P}(Y = 1)$, $\pi_0 = \mathbb{P}(Y = 0)$, $p_1(x) = p(x|y = 1)$ and $p_0(x) = p(x|y = 0)$. Inserting sample estimates of π_1 and π_0, and density estimates for p_1 and p_0 yields an estimate of the Bayes rule. For clustering, we look for the high density regions, based on an estimate of the density. Many classifiers that you are familiar with can be re-expressed this way. Unsupervised prediction is discussed in Section 9.

Perplexity

In information theory, **perplexity** is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the sample.

There are Perplexity in probability distribution and probability model and perplexity for word

Perplexity of probability distribution

Perplexity of a probability distribution [edit]

The perplexity of a discrete probability distribution p is defined as

$$2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$$

where $H(p)$ is the entropy of the distribution and x ranges over events.

Perplexity of a random variable X may be defined as the perplexity of the distribution over its possible values x.

In the special case where p models a fair k-sided die (a uniform distribution over k discrete events), its perplexity is k. A random variable with perplexity k has the same uncertainty as a fair k-sided die, and one is said to be "k-ways perplexed" about the value of the random variable. (Unless it is a fair k-sided die, more than k values will be possible, but the overall uncertainty is no greater because some of these values will have probability greater than $1/k$, decreasing the overall value while summing.)

Perplexity is sometimes used as a measure of how hard a prediction problem is. This is not always accurate. If you have two choices, one with probability 0.9, then your chances of a correct guess are 90 percent using the optimal strategy. The perplexity is $2^{-0.9 \log_2 0.9 - 0.1 \log_2 0.1} = 1.38$. The inverse of the perplexity (which, in the case of the fair k-sided die, represents the probability of guessing correctly), is 1/1.38 = 0.72, not 0.9.

The perplexity is the exponentiation of the entropy, which is a more clearcut quantity. The entropy is a measure of the expected, or "average", number of bits required to encode the outcome of the random variable, using a theoretical optimal variable-length code, cf. the next section. It can equivalently be regarded as the expected information gain from learning the outcome of the random variable, where information is measured in bits.

Perplexity of probability model

Perplexity of a probability model [edit]

A model of an unknown probability distribution p, may be proposed based on a training sample that was drawn from p. Given a proposed probability model q, one may evaluate q by asking how well it predicts a separate test sample $x_1, x_2, ..., x_N$ also drawn from p. The perplexity of the model q is defined as

$$b^{-\frac{1}{N}\sum_{i=1}^{N}\log_b q(x_i)}$$

where b is customarily 2. Better models q of the unknown distribution p will tend to assign higher probabilities $q(x_i)$ to the test events. Thus, they have lower perplexity: they are less surprised by the test sample.

The exponent above may be regarded as the average number of bits needed to represent a test event x_i if one uses an optimal code based on q. Low-perplexity models do a better job of compressing the test sample, requiring few bits per test element on average because $q(x_i)$ tends to be high.

The exponent may also be regarded as a cross-entropy,

$$H(\tilde{p}, q) = -\sum_{x} \tilde{p}(x) \log_2 q(x)$$

where $\tilde{p}$ denotes the empirical distribution of the test sample (i.e., $\tilde{p}(x) = n/N$ if x appeared n times in the test sample of size N).

Perplexity per word

In natural language processing, perplexity is a way of evaluating language models. A language model is a probability distribution over entire sentences or texts.

Using the definition of perplexity for a probability model, one might find, for example, that the average sentence x_i in the test sample could be coded in 190 bits (i.e., the test sentences had an average log-probability of -190). This would give an enormous model perplexity of 2^{190} per sentence. However, it is more common to normalize for sentence length and consider only the number of bits per word. Thus, if the test sample's sentences comprised a total of 1,000 words, and could be coded using a total of 7.95 bits per word, one could report a model perplexity of $2^{7.95} = 247$ *per word.* In other words, the model is as confused on test data as if it had to choose uniformly and independently among 247 possibilities for each word.

The lowest perplexity that has been published on the Brown Corpus (1 million words of American English of varying topics and genres) as of 1992 is indeed about 247 per word, corresponding to a cross-entropy of $\log_2 247 = 7.95$ bits per word or 1.75 bits per letter [1] using a trigram model. It is often possible to achieve lower perplexity on more specialized corpora, as they are more predictable.

Again, simply guessing that the next word in the Brown corpus is the word "the" will have an accuracy of 7 percent, not 1/247 = 0.4 percent, as a naive use of perplexity as a measure of predictiveness might lead one to believe. This guess is based on the unigram statistics of the Brown corpus, not on the trigram statistics, which yielded the word perplexity 247. Using trigram statistics would further improve the chances of a correct guess.

Generative approaches, Discriminative approaches

Reference https://www.quora.com/What-are-the-differences-between-generative-and-discriminative-machine-learning

1. Generative Models:

One way is to model p(x, y) directly. Once we do that, we can obtain p(y|x) by simply conditioning on x. And we can then use decision theory to determine class membership i.e. we can use loss matrix, etc. to determine which class the point belongs to (such an assignment would minimize the expected loss). For e.g. in Naive Bayes model, you can learn p(y), the prior class probabilities from the data. You can also learn p(x|y) from the data using say maximum likelihood estimation (or you can Bayes estimator, if you will). Once you have p(y) and p(x|y), p(x, y) is not difficult to find out.

2. Discriminative Models:

Instead of modeling p(x, y), we can directly model p(y|x), for e.g. in logistic regression p(y|x) is assumed to be of the form 1 / (1 + exp(-sigma(wi. xi))). All we have to do in such a case is to learn weights that would minimize the squared loss.

3. Encoding a Function:

We find a function f(.) that directly maps x to a class. Decision trees do that.

And I just want to add one simple trick that works for me: whenever an algorithm involves assuming, calculating or estimating the distribution of Y, it is generative, or simply put, **if the algorithm cares about the distribution of Y, it is generative, if not, then it is a discriminative.**

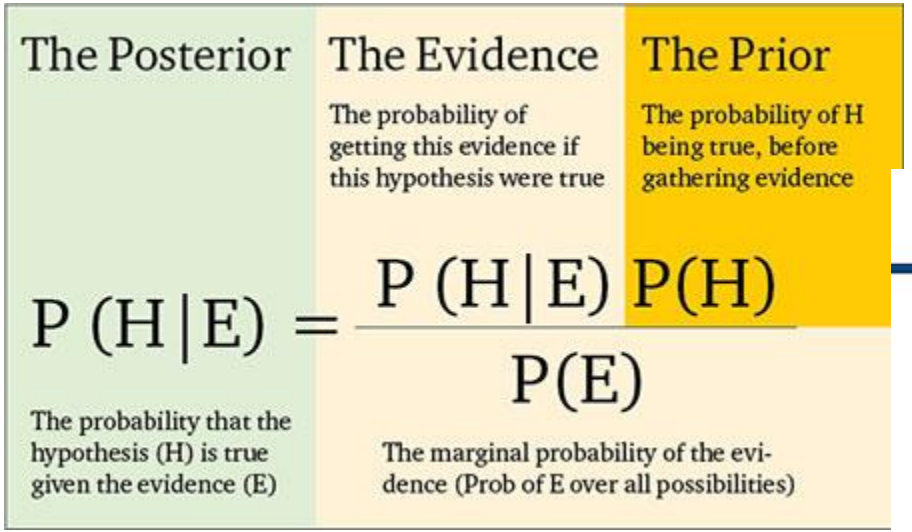
For example, for logistic regression, you don't really care about the distribution of Y, well maybe you don't want it to be too unbalance (too many 0, too few 1) to work properly, but that is not the fundamental problem.

Whereas in Normal Discriminant Analysis, you would require or assume the distribution of Y to be multinomial, and in Naive Bayes Classifier, it would involve estimating the distribution of Y in the algorithm, as seen in the equation (2) of the paper here.

Why Dirichlet distribution the prior over multinomial distribution

http://stats.stackexchange.com/questions/44494/why-is-the-dirichlet-distribution-the-prior-for-the-multinomial-distribution

Bayesian



Bayes Theorem

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

- P(A|B) is conditional probability of observing A given B is true
- P(B|A) is conditional probability of observing B given A is true
- P(A) and P(B) are probabilities of A and B without conditioning on each other

http://stats.stackexchange.com/questions/44494/why-is-the-dirichlet-distribution-the-prior-for-the-multinomial-distribution

The Dirichlet distribution is a conjugate prior for the multinomial distribution. This means that if the prior distribution of the multinomial parameters is Dirichlet then the posterior distribution is also a Dirichlet distribution (with parameters different from those of the prior). The benefit of this is that (a) the posterior distribution is easy to compute and (b) it in some sense is possible to quantify how much our beliefs have changed after collecting the data.

It can certainly be discussed whether these are good reasons to choose a particular prior, as these criteria are unrelated to actual prior beliefs... Nevertheless, conjugate priors are popular, as they often are reasonably flexible and convenient to use for the reasons stated above.

For the special case of the multinomial distribution, let $(p_1,\ldots,p_k)$

be the vector of multinomial parameters (i.e. the probabilities for the different categories). If

$$(p_1,\ldots,p_k)\sim\text{Dirichlet}(\alpha_1,\ldots,\alpha_k)$$

prior to collecting the data, then, given observations $(x_1,\ldots,x_k)$ in the different categories,

$$(p_1,\ldots,p_k)|\,|\,|(x_1,\ldots,x_k)\sim\text{Dirichlet}(\alpha_1+x_1,\ldots,\alpha_k+x_k).$$

The uniform distribution is actually a special case of the Dirichlet distribution, corresponding to the case $\alpha_1=\alpha_2=\cdots=\alpha_k=1$

. So is the least-informative Jeffreys prior, for which $\alpha_1=\cdots=\alpha_k=1/2$. The fact that the Dirichlet class includes these natural