Dirichlet Processes and Nonparametric Bayesian Modelling

Volker Tresp
Motivation

- Infinite models have recently gained a lot of attention in Bayesian machine learning.
- They offer great flexibility and, in many applications, allow a more truthful representation.
- The most prominent representatives are Gaussian processes and Dirichlet processes.
Gaussian Processes: Modeling Functions

- Gaussian processes define a prior over functions

- A sample of a Gaussian process is a function

\[ f(\cdot) \sim \text{GP}(\cdot | \mu(\cdot), k(\cdot, \cdot)) \]

- Gaussian processes are infinite-dimensional generalizations of finite-dimensional Gaussian distributions

- In a typical problem we have samples of the underlying true function and we want to calculate the posterior distribution of the functions and make predictions at a new input (Gaussian process smoothing)

- In a related setting, we can only obtain noisy measurements of the true function (Gaussian process regression)
Dirichlet Processes: Modeling Probability measures

- Dirichlet processes define a prior over probability measures
- A sample of a Dirichlet process is a probability measure
  \[ G \sim \text{DP} \left( \cdot | G_0, \alpha_0 \right) \]
- Infinite-dimensional Dirichlet processes are generalizations to finite Dirichlet distributions
- In a typical problem we have samples of the underlying true probability measure and we want to calculate the posterior probability measure or the predictive distribution for a new sample; (note, that we do not have a measurement of the function, as in the GP case but a sample of the true probability measure; this is the main difference between GP and DP)
- In a related setting, we can only obtain noisy measurements of a sample; this is then a *Dirichlet process mixture model*
Outline

• I: Introduction to Bayesian Modeling
• II: Multinomial Sampling with a Dirichlet Prior
• III: Hierarchical Bayesian Modeling
• IV: Dirichlet Processes
• V: Applications and More on Nonparametric Bayesian Modeling
I: Introduction to Bayesian Modeling
Statistical Approaches to Learning and Statistics

- Probability theory is a branch of mathematics

- Statistics and (statistical) machine learning are attempts to applying probability theory to solving problems in the real world: effectiveness of a medication, text classification, medical expert systems, ...

- There are different approaches to applying probability theory to problems in the real world in a useful way: frequentist statistics, Bayesian statistics, statistical learning theory, ...

- All of them are useful in their own right

- In this tutorial, we take a Bayesian point of view
Review of Some Laws of Probability
Multivariate Distribution

- We start with two (random) variables $X$ and $Y$. A multivariate (probability) distribution is defined as

$$P(x, y) := P(X = x, Y = y) = P(X = x \land Y = y)$$
Conditional Distribution

- Definition of a conditional distribution

\[ P(Y = y | X = x) := \frac{P(X = x, Y = y)}{P(X = x)} \quad \text{where} \quad P(X = x) > 0 \]
Product Decomposition and Chain Rule

From the definition of a conditional distribution we obtain:

- **Product Decomposition**
  \[ P(x, y) = P(x|y)P(y) = P(y|x)P(x) \]

- and the **chain rule**
  \[ P(x_1, \ldots, x_M) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \ldots P(x_M|x_1, \ldots, x_{M-1}) \]
Bayes’ Rule

- **Bayes’ Rule** follows immediately

\[
P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(y|x)P(x)}{P(y)} \quad P(y) > 0
\]
Marginal Distribution

To calculate a **marginal distribution** from a joint distribution, one uses:

\[
P(X = x) = \sum_y P(X = x, Y = y)
\]
Bayesian Reasoning and Bayesian Statistics
Bayesian Reasoning

- Bayesian reasoning is the straightforward application of the rules of probability to real world problems involving uncertain reasoning.

- $P(H = 1)$: assumption about the truth of hypothesis $H$ (a priori probability).

- $P(D|H = 1)$: Probability of observing (Data) $D$, if hypothesis $H$ is true (likelihood);
  $P(D|H = 0)$: Probability of observing (Data) $D$, if hypothesis $H$ is not true (likelihood).

- Bayes’ rule:

$$P(H = 1|D) = \frac{P(D|H = 1)P(H = 1)}{P(D)}$$

- Evidence: $P(D) = P(D|H = 1)P(H = 1) + P(D|H = 0)P(H = 0)$
Bayesian Reasoning: Example

- A friend has a new car
- A priori assumption:
  \[ P(\text{Car} = \text{SportsCar}) = 0.5 \]
- I learn that the car has exactly two doors; likelihood:
  \[ P(2\text{Doors}|\text{Car} = \text{SportsCar}) = 1 \quad P(2\text{Doors}|\text{Car} = \neg\text{SportsCar}) = 0.5 \]
- Using Bayes’ theorem:
  \[ P(\text{Car} = \text{SportsCar}|2\text{Doors}) = \frac{1 \times 0.5}{(1 \times 0.5 + 0.5 \times 0.5)} = 0.66 \]
Bayesian Reasoning: Debate

• There is no disagreement that one can define an appropriate likelihood term $P(D|H)$

• There is disagreement, if one should be allowed to define and exploit the prior probability of a hypothesis $P(H)$, since in most cases, this can only present someone’s prior belief

• Non-Bayesians often criticize the necessity to model someone’s prior belief: this appears to be subjective and non-scientific

• To people sympathetic to Bayesian reasoning: the prior distribution can be used to incorporate valuable prior knowledge and constraints (e.g., medical expert system); it is a necessity for obtaining a complete statistical model

• Comment: Since in parametric modeling the assumption about the likelihood function is much more critical than assumptions concerning the prior distribution, the discussion might not be quite to the point
Bayesian Reasoning: Subjective Probabilities

- If one is willing to assign numbers to beliefs then under few assumptions of consistency and if 1 means that one is certain that an event will occur and if 0 means that one is certain that an event will not occur, then these numbers exactly behave as probabilities. Theorem: Any measure of belief is isomorphic to a probability measure (Cox, 1946).

- “One common criticism of the Bayesian definition of probability is that probabilities seem arbitrary. Why should degrees of belief satisfy the rules of probability? On what scale should probabilities be measured? In particular, it makes sense to assign a probability of one (zero) to an event that will (not) occur, but what probabilities do we assign to beliefs that are not at the extremes? Not surprisingly, these questions have been studied intensely. With regards to the first question, many researchers have suggested different sets of properties that should be satisfied by degrees of belief (e.g., Ramsey 1931, Cox 1946, Good 1950, Savage 1954, DeFinetti 1970). It turns out that each set of properties leads to the same rules: the rules of probability. Although each set of properties is in itself compelling, the fact that different sets all lead to the rules of probability provides a particularly strong argument for using probability to measure beliefs.” Heckerman: A Tutorial on Learning With Bayesian Networks
Technicalities in Bayesian Statistics
Basic Approach in Statistical Bayesian Modeling

- Despite the fact that in Bayesian modeling any uncertain quantity of interest is treated as a random variable, one typically distinguishes between parameters and variables; parameters are random variables that are assumed fixed in the domain of interest whereas variables might assume different states in each data point (e.g., object, measurement)

- In a typical setting might have observed data $D$, unknown parameters $\theta$ and a quantity to be predicted $X$. Furthermore, we might have latent variables $H_D$ and $H$ in the training data and in the test point, respectively.

- One first builds a joint model, using the product rule (example)

\[
P(\theta, H_D, D, H, X) = P(\theta)P(D, H_D|\theta)P(X, H|\theta)
\]

$P(\theta)$ is the prior distribution, $P(D, H_D|\theta)$ is the complete data likelihood; we might be interested in $P(X|D)$

- First we marginalize the latent variables $P(D|\theta) = \int P(D, H_D|\theta) \, dH_D$ and obtain the likelihood w.r.t the observed data $P(D|\theta)$
Basic Approach in Statistical Bayesian Modeling (2)

• Thus we obtain the posterior parameters distribution using Bayes rule

\[ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \]

• Then we marginalize the parameters and obtain

\[ P(X, H|D) = \int P(X, H|\theta)P(\theta|D) \, d\theta \]

• Finally, one marginalizes the latent variable in the test point

\[ P(X|D) = \sum_H P(X, H|D) \]

• The demanding operations are the integrals, resp. sums; thus one might say with some justification: the frequentist optimizes (e.g., in the maximum likelihood approach), and the Bayesian integrates
The integrals are often over high-dimensional quantities; typical approaches to solving or approximating the integrals

- Closed-from solutions (exist for some special cases)
- Laplace approximation (leads to an optimization problem)
- Markov Chain Monte Carlo Sampling (e.g., Gibbs sampling, ...) (integration via Monte Carlo)
- Variational approximations (e.g., mean field) (leads to an optimization problem)
- Expectation Propagation
Conclusion on Bayesian Modeling

- Bayesian modeling is the straightforward application of the laws of probability to problems in the real world.

- The Bayesian program is quite simple: build a model, get data, perform inference.

- No matter, which paradigm one follows: one should never forget that the assumptions going into any statistical model (in particular in machine learning) are almost always very rough approximations (a cartoon).
II: Multinomial Sampling with a Dirichlet Prior
Likelihood, Prior, Posterior, and Predictive Distribution
Before we introduce the Dirichlet process, we need to get a good understanding of the finite-dimensional case: Multinomial sampling with a Dirichlet prior.

Learning and inference in the finite case find their equivalences in the infinite-dimensional case of Dirichlet processes.

Example: Tossing a Loaded Dice

- Running example: the repeated tossing of a loaded dice
- Let’s assume that we toss a loaded dice; by $\Theta = \theta^k$ we indicate the fact that the toss resulted in showing $\theta^k$
- Let’s assume that we observe in $N$ tosses $N_k$ times $\theta^k$
- A reasonable estimate is then that

$$\hat{P}(\Theta = \theta^k) = \frac{N_k}{N}$$
Multinomial Likelihood

- In a formal model we would assume multinomial sampling; the observed variable $\Theta$ is discrete, having $r$ possible states $\theta^1, \ldots, \theta^r$. The likelihood function is given by

$$P(\Theta = \theta^k | g) = g_k, \quad k = 1, \ldots, r$$

where $g = \{g_2, \ldots, g_r, \}$ are the parameters and $g_1 = 1 - \sum_{k=2}^{r} g_k, \ g_k \geq 0, \forall k$

- Here, the parameters correspond to the physical probabilities

- The sufficient statistics for a data set $D = \{\Theta_1 = \theta_1, \ldots, \Theta_N = \theta_N\}$ are $\{N_1, \ldots, N_r\}$, where $N_k$ is the number of times that $\Theta = \theta^k$ in $D$. (In the following, $D$ will in general stand for the observed data)
The likelihood for the complete data set (here and in the following, \( C \) denotes normalization constants irrelevant for the discussion)

\[
P(D|g) = \text{Multinomial}(\cdot|g) = \frac{1}{C} \prod_{k=1}^{r} g_k^{N_k}
\]

The maximum likelihood estimate is (exercise)

\[
g_k^{ML} = \frac{N_k}{N}
\]

Thus we obtain the very intuitive result that the parameter estimates are the empirical counts. If some or many counts are very small (e.g., when \( N < r \)) many probabilities might be (incorrectly) estimated to be zero; thus, a Bayesian treatment might be more appropriate.
Dirichlet Prior

• In a Bayesian framework, one defines an a priori distribution for $g$. A convenient choice is a conjugate prior, in this case a Dirichlet distribution

$$P(g|\alpha^*) = \text{Dir}(\cdot|\alpha_1^*, \ldots, \alpha_r^*) \equiv \frac{1}{C} \prod_{k=1}^{r} g_k^{\alpha_k^* - 1}$$

• $\alpha^* = \{\alpha_1^*, \ldots, \alpha_r^*\}$, $\alpha_k^* > 0$.

• It is also convenient to re-parameterize

$$\alpha_0 = \sum_{k=1}^{r} \alpha_k^* \quad \alpha_k = \frac{\alpha_k^*}{\alpha_0} \quad k = 1, \ldots, r$$

and $\alpha = \{\alpha_1, \ldots, \alpha_r\}$ such that $\text{Dir}(\cdot|\alpha_1^*, \ldots, \alpha_r^*) \equiv \frac{1}{C} \prod_{k=1}^{r} g_k^{\alpha_0 \alpha_k - 1}$.

• The meaning of $\alpha$ becomes apparent when we note that

$$P(\Theta = \theta^k|\alpha^*) = \int P(\Theta = \theta^k|g)P(g|\alpha^*) \, dg = \int g_k \text{Dir}(g|\alpha^*) \, dg = \alpha_k$$
The posterior distribution is again a Dirichlet with
\[
P(g|D, \alpha^*) = \text{Dir}(\cdot|\alpha_1^* + N_1, \ldots, \alpha_r^* + N_r)
\]
(Incidentally, this is an inherent property of a conjugate prior: the posterior comes from the same family of distributions as the prior)

The probability for the next data point (after observing \(D\))
\[
P(\Theta_{N+1} = \theta^k|D, \alpha^*) = \int g_k \text{Dir}(g|\alpha_1^* + N_1, \ldots, \alpha_r^* + N_r) dg = \frac{\alpha_0 \alpha_k + N_k}{\alpha_0 + N}
\]

We see that with increasing \(N_k\) we obtain the same result as with the maximum likelihood approach and the prior becomes negligible
Dirichlet Distributions for $\text{Dir}(\cdot|\alpha_1^*, \alpha_2^*, \alpha_3^*)$

$$\text{Dir}(\cdot|\alpha_1^*, \ldots, \alpha_r^*)$$

$$\equiv \frac{1}{C} \prod_{k=1}^{r} g_k^{\alpha_k^* - 1}$$

(From Ghahramani, 2005)
Generating Samples from g and $\theta$
Generative Model

- Our goal is now to use the multinomial likelihood model with a Dirichlet prior as a generative model.

- This means that we want to "generate" loaded dices according to our Dirichlet prior and "generate" virtual tosses from those virtual dices.

- The next slide shows a graphical representation.
Graphical Model: multinomial sampling with a Dirichlet prior

Plate representation
First Approach: Sampling from $g$

- The first approach is to first generate a sample $g$ from the Dirichlet prior.

- This is not straightforward but algorithms for doing that exist; (one version involves sampling from independent gamma distributions using shape parameters $\alpha_1^*, \ldots, \alpha_r^*$ and normalizing those samples) (later in the DP case, this sample can be generated using the stick breaking presentation).

- Given a sample $g$ it is trivial to generate independent samples for the tosses with

$$P(\Theta = \theta^k | g) = g_k$$
Second Approach: Sampling from $\Theta$ directly

- We can also take the other route and sample from $\Theta$ directly.
- Recall the probability for the next data point (after observing $D$)

$$P(\Theta_{N+1} = \theta^k | D) = \frac{\alpha_0 \alpha_k + N_k}{\alpha_0 + N}$$

We can use the same formula, only that now $D$ are previously generated samples; this simple equation is of central importance and will reappear in several guises repeatedly in the tutorial.

- Thus there is no need to generate an explicit sample from $g$ first.
- Note, that with probability proportional to $N$, we will sample from the empirical distribution with $P(\Theta = \theta^k) = N_k/N$ and with probability proportional to $\alpha_0$ we will generate a sample according to $P(\Theta = \theta^k) = \alpha_k$
Second Approach: Sampling from $\Theta$ directly (2)

- Thus a previously generated sample increases the probability that the same sample is generated at a later stage; in the DP model this behavior will be associated with the Pólya urn representation and the Chinese restaurant process
\[ P(\Theta_{N+1} = \theta^k | D) = \frac{\alpha_0 \alpha_k + N_k}{\alpha_0 + N} \quad \text{with } \alpha_0 \to 0: \text{ A Paradox?} \]

- If we let \( \alpha_0 \to 0 \), the first generated sample will dominate all samples generated thereafter: they will all be identical to the first sample; but note that independent of \( \alpha_0 \) we have \( P(\Theta = \theta^k) = \alpha_k \).

- Note also that

\[
\lim_{\alpha_0 \to 0} P(g|\alpha^*) \propto \prod_{k=1}^{r} \frac{1}{g_k}
\]

such that distributions with many zero-entries are heavily favored.

- Here is the paradox: the generative model will almost never produce a fair dice but if actual data would indicate a fair dice, the prior is immediately and completely ignored.

- Resolution: The Dirichlet prior with a small \( \alpha_0 \) favors extreme solutions, but this prior belief is very weak and is easily overwritten by data.

- This effect will reoccur with the DP: if \( \alpha_0 \) is chosen to be small, sampling heavily favors clustered solutions.
The Beta-distribution is a two-dimensional Dirichlet with two parameters $\alpha$ and $\beta$; for small parameter values, we see that extreme solutions are favored.
Noisy Observations
Noisy Observations

• Now we want to make the model slightly more complex; we assume that we cannot observe the results of the tosses $\Theta$ directly but only (several) derived quantities (e.g., noisy measurements) $X$ with some $P(X|\Theta)$. Let $D_k = \{x_{k,j}\}_{j=1}^{M_k}$ be the observed measurements of the $i$-th toss and let $P(x_{k,j}|\theta_k)$ be the probability distribution (several unreliable persons inform you about the results of the tosses).

• Again we might be interested in inferring the property of the dice by calculating $P(g|D)$ (the probabilities of the properties of the dice) or in the probability of the actual tosses $P(\Theta_1, \ldots, \Theta_N|D)$.

• This is now a problem with missing data (the $\Theta$ are missing); since it is relevant also for DP, we will only discuss approaches based on Gibbs sampling but we want to mention that the popular EM algorithm might also be used to obtain a point estimate of $g$.

• The next slide shows a graphical representation.
Graphical Model: multinominal sampling with a Dirichlet prior and noisy observations
Inference based on Markov Chain Monte Carlo Sampling

• What we have learned about the model based on the data is incorporated in the predictive distribution

\[
P(\Theta_{N+1}|D) = \sum_{\theta_1, \ldots, \theta_N} P(\Theta_1, \ldots, \Theta_N|D)P(\Theta_{N+1}|\Theta_1, \ldots, \Theta_N)
\]

\[
\approx \frac{1}{S} \sum_{s=1}^{S} P(\Theta_{N+1}|\theta_1^s, \ldots, \theta_N^s)
\]

where (Monte Carlo approximation) \(\theta_1^s, \ldots, \theta_N^s \sim P(\Theta_1, \ldots, \Theta_N|D)\)

• In contrast to before, we now need to generate samples from the posterior distribution; ideally, one would generate samples independently, which is often infeasible

• In Markov chain Monte Carlo (MCMC), the next generated sample is only dependent on the previously generated sample (in the following we drop the \(s\) label in \(\theta^s\))
Gibbs Sampling

- Gibbs sampling is a specific form of an MCMC process.

- In Gibbs sampling we initialize all variables in some appropriate way, and replace a value \( \Theta_k = \theta_k \) by a sample of \( P(\Theta_k | \{ \Theta_i = \theta_i \}_{i \neq k}, D) \). One continuous to do this repeatedly for all \( k \). Note, that \( \Theta_k \) is dependent on its data \( D_k = \{ x_{k,j} \}_j \) but is independent of the remaining data given the samples of the other \( \Theta \).

- The generated samples are from the correct distribution (after a burn in phase); a problem is that subsequent samples are not independent, which would be a desired property; it is said that the chain does not mix well.

- Note that we can integrate out \( g \) so we never have to sample from \( g \); this form of sampling is called collapsed Gibbs sampling.
Gibbs Sampling (2)

• We obtain (note, that $N_l$ are the counts without considering $\Theta_k$)

$$P(\Theta_k = \theta^l | \{\Theta_i = \theta_i\}_{i \neq k}, D) = P(\Theta_k = \theta^l | \{\Theta_i = \theta_i\}_{i \neq k}, D_k)$$

$$= \frac{1}{C} P(\Theta_k = \theta^l | \{\Theta_i = \theta_i\}_{i \neq k}) P(D_k | \Theta_k = \theta^l)$$

$$= \frac{1}{C} (\alpha_0 \alpha_l + N_l) P(D_k | \Theta_k = \theta^l)$$

$$C = \sum_l (\alpha_0 \alpha_l + N_l) P(D_k | \Theta_k = \theta_l)$$
Auxiliary Variables, Blocked Sampling and the Standard Mixture Model
Introducing an Auxiliary Variable $Z$

- The figure shows a slightly modified model; here the auxiliary variables $Z$ have been introduced with states $z^1, \ldots, z^r$.

- We have

$$P(Z = z^k | g) = g_k, \quad k = 1, \ldots, r$$

$$P(\Theta = \theta^j | Z = z^k) = \delta_{j,k}, \quad k = 1, \ldots, r$$

- If the $\theta$ are fixed, this leads to the same probabilities as in the previous model and we can again use Gibbs sampling.
Graphical Model: multinomial sampling with a Dirichlet prior and noisy observations including an auxiliary variable Z
Collapsing and Blocking

• So far we had used a collapsed Gibbs sampler, which means that we never explicitly sampled from $g$.

• This is very elegant but has the problem that the Gibbs sampler does not mix very well.

• One often obtains better sampling by using a non-collapsed Gibbs sample, i.e., by sampling explicitly from $g$.

• The advantage is that given $g$, one can independently sample from the auxiliary variables in a block (thus the term *blocked* Gibbs sampler).
The Blocked Gibbs Sampler

One iterates

- We generate samples from $Z_k|g, D_k$ for $k = 1, \ldots, N$
- We generate a sample from $g|Z_1, \ldots, Z_N \sim \text{Dir}(\alpha_1^* + N_1, \ldots, \alpha_1^* + N_r)$

where $N_k$ is the number of times that $Z_k = z^k$ in the current sample.
Relationship to a standard Mixture Model: Learning $\theta$

- We can now relate our model to a standard mixture model; note, that this is not the same model any more.

- The main difference is that now we treat the the $\theta_k$ as random variables; this corresponds to the situation where $Z$ would tell us which side of the dice is up and $\theta_k$ would correspond to a value associated with the $k$-th face.

- We now need to put a prior on $\theta_k$ with hyperparameters $h$ and learn $\theta_k$ from data (see figure)!

- A reasonable prior for the probabilities might

$$P(\pi | \alpha_0) = \text{Dir}(\cdot | \alpha_0/r, \ldots, \alpha_0/r)$$

- As a special case: when $M_k = 1$, and typically $r << N$, this corresponds to a typical mixture model; a mixture model is a probabilistic version of (soft) clustering.

- Example: if the $P(X|\Theta)$ is a Gaussians distribution with parameters $\Theta$, we obtain a Gaussian mixture model.
Relationship to a standard Mixture Model: Learning $\theta$ (2)

- Gibbs sampling as before can be used but needs to be extended to also generate sample for $\Theta$

- Again, this is a slightly difference model; in the case of infinite models, although, we can indeed define an infinite mixture model which exactly corresponds to the infinite version of the previously defined model!
A finite mixture model
Conclusions for the Multinomial Model with a Dirichlet Prior

- We applied the Bayesian program to a model with a multinomial likelihood and Dirichlet prior.
- We discussed a number of variations on inference, in particular variations on Gibbs sampling.
- But one might argue that we are still quite restrictive in the sense that if one is not interested in loaded dices or gambling in general this might all be not so relevant.
- In the next section we show that by a process called Dirichlet enhancement, the Dirichlet model is the basis for nonparametric modeling in a very general class of hierarchical Bayesian models.
III: Hierarchical Bayesian Modeling and Dirichlet Enhancement
Hierarchical Bayesian Modeling
Hierarchical Bayesian Modelling

• In hierarchical Bayesian modeling both parameters and variables are treated equally as random variables (as we have done in the multinomial model).

• In the simplest case we would assume that there are random variables that might take on specific values in each instance. Example: diagnosis and length of stay of a person in a given hospital typically differs in different patients.

• Then we would assume that there are variables, which we would model as being constant (but unknown) in a domain. These would typically be called parameters. Example: average length of stay given the diagnosis in a given hospital.
The Standard Hierarchy

- The figure shows the standard Bayesian model for supervised learning; as a concrete example let’s assume the goal is to predict the preference for an object \( y \) given object features \( x \) and given parameters \( \theta \). The parameters have a prior distribution with parameters \( g \), which itself originates from a distribution with parameters \( \alpha \).

- The hierarchical probability model is

\[
P(\alpha)P(g|\alpha)P(\theta|g) \prod_{j=1}^{M} P(y_j|x_j, \theta)
\]
A standard hierarchical model
The Standard Hierarchy (2)

- The hyperparameters can be integrated out and one obtains

\[ P(\theta, D) = P(\theta)P(D|\theta) = P(\theta) \prod_{j=1}^{M} P(y_j|x_j, \theta) \]

with

\[ P(\theta) = \int P(\alpha)P(g|\alpha)P(\theta|g)d\alpha dg \]

- The effect of the prior vanishes when sufficient data are available: The posterior probability gets increasingly dominated by the likelihood function; thus the critical term to specify by the user is the functional form of the likelihood! One then needs to do an a posteriori analysis and check if the assumptions about the likelihood were reasonable...
Consider the situation of learning a model for predicting the outcome for patients with a particular disease based on patient information. Due to differences in patient mix and hospital characteristics such as staff experiences the models are different for different hospitals but also will share some common effects. This can be modeled by assuming that the model parameters originate from a particular distribution of parameters that can be learned from data from a sufficiently large number of hospitals. If applied to a new hospital, this learned distribution assumes the role of a learned prior.

- A preference model for items (movies, books); the preference model is individual for each person.

- The probability of a word is document specific; the word probabilities come out off a cluster of similar word documents.

- The figure shows a graphical representation
An object oriented hierarchical model
Discussion of the Extended Hierarchical Model

- Inference and learning is more difficult but in principle nothing new (Gibbs sampling might be applied)

- Let’s look at the convergence issues;

- As before, $P(\theta_k|D_k)$ will converge to a point mass with $M_k \rightarrow \infty$

- With increasing numbers of situations and data for the situations, also $g$ will converge to a point mass at some $\hat{g}$

- This means that for a new object $N + 1$, we can inherit the learned prior distribution

  $$P(\theta_{N+1}|D_1, \ldots, D_N) \approx P(\theta_{N+1}|\hat{g})$$
Towards a Nonparametric Approach: Dirichlet Enhancement
Model Check in Hierarchical Bayesian Modelling

- In the standard model, the likelihood was critical and should be checked to be correct.
- In a hierarchical Bayesian model, in addition, the learned prior

\[ P(\theta_{N+1}|D_1, \ldots, D_N) \approx P(\theta_{N+1}|\hat{g}) \]

should be checked; this distribution is critical for the sharing strength effect and the assumed functional form of the prior becomes much more important! Also note that \( \theta \) is often high dimensional (whereas the likelihood often reduces to evaluating scalar probabilities, e.g., in the case of additive independent noise).

- A simple parametric prior is typically too inflexible to represent the true distribution.
- Thus one needs nonparametric distributions as priors such as derived from the the Dirichlet Process; the figure illustrates the point.
Prior Distribution

Posterior Distribution

Set of max. likl. estimates;

$\theta^{ML}(N_D \to \infty)$

$P(\theta | h_{post})$ with $N_D = 100$

Set of max. likl. estimates where a nonparametric distribution might be appropriate
Dirichlet Enhancement: The Key Idea

• Let’s assume that we consider only discrete $\theta \in \{\theta^1, \ldots, \theta^r\}$ with a very large $r$

• Now we can re-parameterize the prior distribution in terms of a multinomial model with a Dirichlet prior

\[
P(\Theta = \theta^k | g) = g_k, \quad k = 1, \ldots, r
\]

\[
P(g | \alpha^*) = \text{Dir}(\cdot | \alpha^*_1, \ldots, \alpha^*_r) \equiv \frac{1}{C} \prod_{k=1}^{r} g_k^{\alpha^*_k - 1}
\]

• We might implement our noninformative prior belief in various forms; for example, one might sample $\theta_i$ from $P(\theta_i)$ and set $\alpha^*_i = \alpha_0, \forall i$
Uninformed Gaussian prior

Samples used for Dirichlet enhancement
Thus we have obtained a model that technically is equivalent the the multinomial likelihood model with a Dirichlet prior and noisy measurements as discussed in the last section.

The process of replacing the original prior by a prior using the Dirichlet Process is sometimes referred to a Dirichlet enhancement.

For inference in the model we can immediately apply Gibbs sampling.
Towards Dirichlet Processes

- Naturally there are computational problems if we let \( r \rightarrow \infty \)
- Technically, we have two options:
  - We introduce an auxiliary variables \( Z \) as before and use a *standard mixture model* where a reasonable small \( r \) might be used; this might not be appropriate if the distribution is not really clustered
  - We let \( r \rightarrow \infty \), which leads us to nonparametric Bayesian modeling and the Dirichlet process
- In the latter case we obtain a Dirichlet process prior and the corresponding model is called a Dirichlet process mixture (DPM)
IV: Dirichlet Processes
Basic Properties
Dirichlet Process

- We have studied the multinomial model with a Dirichlet prior and extended the model to the case of noisy measurements.

- We have studied the hierarchical Bayesian model and found that in the case of repeated trials it makes sense to employ Dirichlet enhancement.

- We have concluded that one can pursue two paths:
  - Either one assumes a finite mixture model and one permits the adaptation of the parameters.
  - Or uses an infinite model and makes the transition from a Dirichlet distribution to a Dirichlet process (DP).

- In this section we study the transition to the DP.

- The Dirichlet Process is a generalization of the Dirichlet distribution; whereas a Dirichlet distribution is a distribution over probabilities, a DP is a measure on measures.
Basic Properties

• Let’s compare the finite case and the infinite case

• In the finite case we wrote \( g \sim \text{Dir}(\cdot | \alpha^*) \), in the infinite case we write

\[
G \sim \text{DP}(\cdot | G_0, \alpha_0)
\]

where \( G \) is a measure (Ferguson, 1973).

• Furthermore, in the finite case we wrote \( P(\Theta = \theta^k|g) = g_k \); in the infinite case we write

\[
\theta \sim G(\cdot)
\]

• \( G_0 \) is the base distribution (corresponds to the \( \alpha \)) and might be describes as a probability density, e.g., as a Gaussian \( G_0 \sim N(\cdot | 0, I) \)

• \( \alpha_0 \) again is a concentration parameter; the graphical structure is shown in the figure
Graphical Model of a Dirichlet Process

Plate representation
Processes and Measures

• In general: One speaks of a *process* (Gaussian process, Dirichlet process) when in some sense the degrees of freedom are infinite. Thus a Gaussian distribution is finite dimensional whereas a Gaussian process is infinite dimensional and is often used to define a prior distributions over functions. In the same sense, the sample of a Dirichlet distribution is a finite discrete probability distribution, whereas a sample from a Dirichlet process is a measure.

• A probability density assumes some continuity so in distributions including point masses it is more appropriate to talk about probability measures.

• In fact a sample from a DP can be written as an infinite sum of weighted delta distributions (see later).
Basic Properties: Posteriors

- In analogy to the finite case, the posterior is again a DP with
  
  \[ G|\theta_1 \ldots \theta_N \sim \text{DP}\left( \frac{1}{\alpha_0 + N} \left[ \alpha_0 G_0 + \sum_{k=1}^{N} \delta_{\theta_k} \right], \alpha_0 + N \right) \]

- \( \delta_{\theta_k} \) is a discrete measure concentrated at \( \theta_k \)

  compare to the finite case

  \[ g|\theta_1 \ldots \theta_N = \text{Dir}(\cdot | \alpha_1^* + N_1, \ldots, \alpha_r^* + N_r) \]
Generating Samples from $G$ and $\theta$
Sampling from $\theta$: Urn Representation

- Consider that $N$ samples $\theta_1, \ldots, \theta_N$ have been generated.

- In the Dirichlet distribution, we used $P(\Theta_{N+1} = \theta^k|D) = \frac{\alpha_0 \alpha_k + N_k}{\alpha_0 + N}$.

- This generalizes in an obvious way in the Dirichlet process to (Blackwell and MacQueen, 1973)

$$
\theta_{N+1}|\theta_1, \ldots, \theta_N \sim \frac{1}{\alpha_0 + N} \left( \alpha_0 G_0(\cdot) + \sum_{k=1}^{N} \delta_{\theta_k} \right)
$$

- This is associated with the Pólya urn representation: one draws balls with different colors out of a urn (with $G_0$); If a ball is drawn, one puts the ball back plus an additional ball with the same color ($\delta_{\theta_k}$); thus in subsequent draws balls with a color already encountered become more likely to be drawn again.

- Note, that there is no need to sample from $G$.
Sampling from $\theta$ (2)

- Note that the last equation can be interpreted as a mixture of distributions:
  - With prob. $\alpha_0/(\alpha_0 + N)$ a sample is generated from distribution $G_0$
  - With prob. $N/(\alpha_0 + N)$ a sample is generated uniformly from $\{\theta_1, \ldots, \theta_N\}$ (which are not necessarily distinct)

- Note, that in the urn process it is likely that identical parameters are repeatedly sampled
Chinese Restaurant Process (CRP)

• This is formalized as the Chinese restaurant process (Aldous, 1985); in the Chinese
  restaurant process it is assumed that customers sit down in a Chinese restaurant
  with an infinite number of tables; $Z_k = j$ means that customer $k$ sits at table $j$.
  Associated with each table $j$ is a parameter $\theta_j$

• The first customer sits at the first table $1$, $Z_1 = 1$; we generate a sample $\theta_1 \sim G_0$.

• With probability $1/(1 + \alpha_0)$, the second customer also sits at the first table $1$,
  $Z_2 = 1$, and inherits $\theta_1$; with probability $\alpha_0/(1 + \alpha_0)$ the customer sits at table
  $2$, $Z_2 = 2$, and a new sample is generated $\theta_1 \sim G_0$.

• The figure shows the situation after $N$ customers have entered the restaurant
Chinese Restaurant Process

\[ \theta_1 \quad \theta_2 \quad \cdots \quad \theta_M \]

Table 1  
Table 2  
Table M

N customers occupy M tables

\[ N_j \quad \text{customers sit at table } j \]

\[ 1 \leq j \leq M \]
Chinese Restaurant Process (CRP)(2)

- Customer $N + 1$ enters the restaurant

- Customer $N + 1$ sits with probability
  
  \[
  \frac{N_j}{N + \alpha_0}
  \]

  at a previously occupied table $j$ and inherits $\theta_j$. Thus: $Z_{N+1} = j$, $N_j \leftarrow N_j + 1$

- With probability
  
  \[
  \frac{\alpha_0}{N + \alpha_0}
  \]

  the customer sits at a new table $M + 1$. Thus: $Z_{N+1} = M + 1$, $N_{M+1} = 1$.

- For the new table a new parameter $\theta_{M+1} \sim G_0(\cdot)$ is generated. $M \leftarrow M + 1$. 
Chinese Restaurant Process (CRP) (3)

- Obviously, the generated samples exactly correspond to the ones generated in the urn representation
Discussion

- So really not much new if compared to the finite case
- In particular we observe the same clustering if $\alpha_0$ is chosen to be small
- The CRP makes the tendency to generate clusters even more apparent (see figure); again the tendency towards forming clusters can be controlled by $\alpha_0$
Samples from a distribution that was generated by a Dirichlet process with a Gaussian base distribution with $\alpha_0 = 10$

Same but with $\alpha_0 = 10000$
Sampling from $G$: Stick Breaking Representation

- After an infinite number of samples are generated the underlying $G(\cdot)$ can be recovered.

- Not surprisingly, the underlying measure can be written as (Sethuraman, 1994)

$$G(\cdot) \sim \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\cdot)$$

$$\pi_k \geq 0, \sum_{k=1}^{\infty} \pi_k = 1 \quad \theta_k \sim G_0(\cdot)$$

- Furthermore, the $\pi_k$ can be generated recursively with $\pi_k = \beta_1$ and

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \quad k \geq 2$$

- $\beta_1, \beta_2, \ldots$ are independent $\text{Be}(1, \alpha_0)$ random variables.

- One writes $\pi \sim \text{Stick}(\alpha_0)$.
Introducing an Auxiliary Variable

- Considering the particular form of the stick breaking prior, we can implement the DP model using an auxiliary variable $Z$ with an infinite number of states $z^1, z^2, \ldots$

- With the stick breaking probability,

  \[ \pi \sim \text{Stick}(\alpha_0) \]

  is generated

- Then, one generates independently for $k = 1, 2, \ldots$

  \[ Z_k \sim \pi \quad \theta_k \sim G_0 \]

- The CRP produces samples of $Z$ and $\theta$ in this model (integrating out $\pi$); compare the graphical model in the figure
A DP model derived from the stick breaking representation
Noisy Observations - The Dirichlet Process Mixture
The Dirichlet Process Mixture (DPM)

- Now we consider that the realizations of $\theta$ are unknown; furthermore we assume that derived quantities (e.g., noisy measurements) $X$ with some $P(X|\theta)$ are available. Let $D_k = \{x_{k,j}\}_j$ be the data available for $\theta_k$ and let $P(x_{k,j}|\theta_k)$ be the probability distribution.

- Note, that this also includes the case that the observer model is conditioned on some input $in_{k,j}: P(x_{k,j}|in_{k,j}, \theta_k)$.

- Recall, that this is exactly the situation encountered in the Dirichlet enhanced hierarchical Bayesian model.

- The Dirichlet Process Mixture is also called: Bayesian nonparametric hierarchical model (Ishwaran), and, not quite accurately, Mixture of Dirichlet processes.
Dirichlet Process Mixture Models

- Gibbs sampling on the collapsed left model corresponds to urn-type sampling
- Gibbs sampling on the collapsed right model corresponds to CRP-type sampling
Gibbs Sampling from the DPM using the Urn Representation

- In analogy to the finite case, the crucial distribution is now

\[
\theta_k|\{\theta_i\}_{i\neq k}, D \sim \frac{1}{C} \left( \alpha_0 G_0(\cdot) + \sum_{l: l \neq k} \delta_{\theta_l} \right) P(D_k|\theta_k)
\]

- This can be re-written as

\[
\theta_k|\{\theta_i\}_{i\neq k}, D \sim \frac{1}{C} \left( \alpha_0 P(D_k) P(\theta_k|D_k) + \sum_{l: l \neq k} P(D_k|\theta_l) \delta_{\theta_l} \right)
\]

\[
C = \alpha_0 P(D_k) + \sum_{l: l \neq k} P(D_k|\theta_l) \delta_{\theta_l}
\]
Sampling from the DPM using the Urn Representation (2)

• Here,

\[ P(D_k) = \int P(D_k | \theta) dG_0(\theta) \]

\[ P(\theta_k | D_k) = \frac{P(D_k | \theta_k) G_0(\theta_k)}{P(D_k)} \]

• Both terms can be calculated in closed form, if \( G_0(\cdot) \) and the likelihood are conjugate. In this case, sampling from \( P(\theta_k | D_k) \) might also be simple.
Sampling from the DPM using the CRP Representation

- We can again use the CRP model for sampling from the DPM

- Folding in the likelihood, we obtain
  
  - We randomly select customer $k$; the customer sat at table $Z_k = i$; we remove him from his table; thus $N_i \leftarrow N_i - 1$; $N \leftarrow N - 1$; if the table $i$ is now unoccupied it is removed; assume, $M$ tables are occupied
  
  - Customer $i$ now sits with probability proportional to
    
    $$N_j P(D_k | \theta_j)$$
    
    at an already occupied table $j$ and inherits $\theta_j$. $Z_k = j$, $N_j \leftarrow N_j + 1$
  
  - With probability proportional to
    
    $$\alpha_0 P(D_k)$$
    
    the customer sits at a new table $M + 1$. $Z_k = M + 1$, $N_{M+1} = 1$. For the new table a new parameter $\theta_{M+1} \sim P(\theta | D_k)$ is generated
Sampling from the DPM using the CRP Representation (2)

- In the CRP representation, \( \theta_k, k = 1, \ldots \) are re-sampled occasionally from the posterior parameter distribution given all data assigned to table \( k \).
- Due to this re-estimation of all parameters assigned to the same table in one step, the Gibbs sampler mixes better than the sampler based on the urn representation.
The customer is removed from the table and then is placed at a table that explains the customer’s data well, giving a preference to highly occupied tables.

- The customer is also allowed to open a new table.
- Occasionally the parameters of the tables are re-estimated.
Example (what is all this good for (2))

• Let’s assume that $P(x_k|\theta_k)$ is a Gaussian distribution, i.e., $\theta_k$ corresponds to the center and the covariance of a Gaussian distribution.

• During CRP-sampling, all data points assigned to the same table $k$ inherit identical parameters, thus can be thought of to be generated from the same Gaussian.

• Thus, the number of occupied tables gives us an estimate of the true number of clusters in the data.

• Thus in contrast to a finite mixture model, we do not have to specify the number of clusters we are looking for in advance!

• $\alpha_0$ is a tuning parameter, tuning the tendency to generate a large number ($\alpha_0$: large) or a small number ($\alpha_0$: small) of clusters.
DPM with a Gaussian model corresponding to an infinite mixture of Gaussians

\[ \mathbf{x}_k \]

\[ \mathbf{\theta}_k \]

\[ \mathcal{N} \]

\[ \alpha_0 \]

\[ G_0 \]

\[ G \]
Snapshot of the sampling of a DPM representing an infinite mixture of Gaussians

- The data points are the customers
- A customer is assigned uniquely to a cluster (i.e., table)
- 6 tables are occupied
- The parameters of the Gaussians are sampled from the posterior distribution of the parameters given the data assigned to the particular table
Relationship to the Standard Mixture Model and Blocked Sampling
Sampling from the DPM using the CRP Representation (2)

- The DPM model with an auxiliary variable can be derived from the standard mixture model (see last section) if the Dirichlet prior for the mixing proportion is

\[ \pi \sim \text{Dir}(\cdot | \alpha_0/r, \ldots, \alpha_0/r) \]

and with

\[ \theta \sim G_0(\cdot) \]

when we let \( r \to \infty \).

- Here, it is even more clear that the DPM can be interpreted as a mixture model with an infinite number of components, where the prior distribution for the parameters is given by the base distribution.
Blocked Sampling Derived from a Finite Stick Representation

- As discussed in the part on the finite models, the sampler mixes better if one could generate samples from $G$, since then, the parameters can be sampled independently; thus it allows blocked updates.

- The stick breaking process allows us to sample from $G$ but the representation is infinite; a finite representation with $K$ terms derived from the stick breaking representation would be the obvious solution.

\[
G(\cdot) \sim \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\cdot)
\]
Truncated Stick Breaking and Dirichlet-multinomial Allocation

- In the *truncated* approach one simply terminates the stick breaking procedure at \( K \) terms; one sets \( \beta_K = 1 \) so that the probabilities sum to one.

- In the *Dirichlet-multinomial allocation* one sets

\[
\pi \sim \text{Dir}(\cdot | \alpha_0/r, \ldots, \alpha_0/r)
\]

- The latter case is identical to a finite mixture model with a large number of components \( r \).
Gibbs sampling for the Dirichlet mixture model

<table>
<thead>
<tr>
<th></th>
<th>Original Model</th>
<th>With auxiliary variable</th>
</tr>
</thead>
</table>
| Blocked Gibbs sampling | Finite approximation:  
                         | • truncated stick breaking  
                         | • Dirichlet multin. alloc.  |
| Collapsed Gibbs sampling| Urn based       | CRP based                |
Formal Definition of a DP

Let $\Theta$ be a measurable space, $G_0$ be a probability measure on $\Theta$, and $\alpha_0$ a positive real number.

For every partition $B_1, B_2, ..., B_k$ of $\Theta$

$$G \sim \text{DP}(\cdot | G_0, \alpha_0)$$

means that

$$(G(B_1), G(B_2), ..., G(B_k)) \sim \text{Dir}(\alpha_0 G_0(B_1), \alpha_0 G_0(B_2), ..., \alpha_0 G_0(B_k))$$

(Ferguson, 1973, Ghahramani, 2005)
Even More Formal Definition of a DP

The theorem asserts the existence of a Dirichlet process and also serves as a definition. Let \((\mathbb{R}, \mathcal{B})\) be the real line with the Borel \(\sigma\)-algebra \(\mathcal{B}\) and let \(M(\mathbb{R})\) bet the set of probability measures on \(\mathbb{R}\), equipped with the \(\sigma\)-algebra \(\mathcal{B}_M\).

**Theorem 1** Let \(\alpha\) be a finite measure on \((\mathbb{R}, \mathcal{B})\). Then there exists a unique probability measure \(D_\alpha\) on \(M(\mathbb{R})\) called the Dirichlet process with parameters \(\alpha\) satisfying:

For every partition \(B_1, B_2, \ldots, B_k\) of \(\mathbb{R}\) by Borel sets
\[
(P(B_1), P(B_2), \ldots, P(B_k)) \sim \text{Dir}(\alpha(B_1), \alpha((B_2), \ldots, \alpha((B_k)))
\]
V: Applications and Extensions
Sharing Statistical Strength: A Recommendation System
Recommendation Systems

- Let’s assume that the task is to build recommendation systems for different users based on the features of the items.

- For a particular user there might not be sufficient data for obtaining a reasonable model.

- Thus it is sensible to build a nonparametric hierarchical model to share statistical strength.
Does Belle like book 3?

 Predictive Models:
 SVM, NN, GLM, ...

\[ P(y | x, \theta) \]
\[ P(y | x, D) \]
It seems that Belle agrees more with Beast!
Different Modeling Assumptions

Different assumptions lead to different models:

- All users are the same: use one model trained on all data
- Each user is a complete individual: train a separate model for each user
- Learn from one another: each user has her/his own model generated from a common prior distribution, which is learned from data and shared between user models
A DPM Recommendation System
WASHINGTON, Dec 5 (Reuter) - The Federal Reserve must be wary when "irrational exuberance" infects stock and other asset markets because that could end up doing damage to the economy, Fed Chairman Alan Greenspan said.

Greenspan, in a speech yesterday to the America's Federal Reserve Bank, said he was encouraged by the conservative American economic policies of commenting on the economy from Wall Street to Washington at the same time.

Greenspan said clear that the Fed needed to cool the financial Street into its deliberations.

Greenspan said, "I am concerned that the Fed needs to cool the financial Street into its deliberations.

British Telecommunications Plc (BT.L) is aiming to offer a package of cash and stock for MCI Communications Corp to bring it 20 percent stake up to 100 percent, a source close to the deal told Reuters on Friday.

MCI earlier confirmed it was talking to BT about a business combination, but declined to give details.

----The BT/MCI Offer----

IT (BT.L) confirms interest in MCI (MCIC.O) merger [nO111302]

IT said to offer cash and stock for MCI (MCIC.O), [nO111302]

Russia marks Eurobond mandate due in Feb-dealers

St Petersburg plans Eurobond late Jan/early Feb

CSFB, Salomon said leading

MGT.S<MGTS.RTS> Eurobond

Focus: Russian shares hit record despite Yeltsin scare

Sidonko plans convertible bond, raises capital

Gazprom<GAZP.RTS> says plans Eurobond in...
News

- We used 36 categories covering totally 10,034 news articles from the Reuters text data set (1,152 articles belong to more than one category)

- For the experiments, we assume that each user is interested in exactly one of these categories

- We generate example data for 360 (artificial) users by choosing at random a set of 30 (positive and negative) example items

- The goal is to predict the probability that a user likes an article

- Rank all unseen articles and select top N ranked articles;

- Plot: how many of the top N articles are truly positive
Welcome to Our Painting Survey Page!

Please click the Save button after rating complete!

Save Ratings and Exit
Paintings

- Task: Predict image rating (642 images)
- 190 users with 89 rated images on average
- Image features: 256 correlogram features (colour/texture) 10 features based on wavelet texture 9 features on colour moments giving a 275-dimensional feature vector for each image.
- Weak indicators of high-level information about an image
- The model predicts N highest ranked pictures; among those, how many were rated positively (in comparison to unrated or negatively rated)
- Classifier: probabilistic SVM with Gaussian kernel http://honolulu.dbs.informatik.uni-muenchen.de:8080/paintings/index.jsp
Active user: given 5 known ratings
Implementation Details

- We used a deterministic variational EM approximation which leads to an approximation (Yu, Schwaighofer, Tresp, Ma, and Zhang, 2003).

\[ G|D \propto \alpha_0 G_0(\cdot) + \sum_{k=1}^{N} \xi_k \delta_{\theta_k^{ML}} \]

- Here, \( \theta_k^{ML} \) is the ML (or MAP) -estimate of each user model trained on its own data
- \( \xi_k \) is optimized in the variational EM approximation
- After convergence, the prediction of an active model \( a \in 1, \ldots, N \) becomes

\[
P(Y_a = y|x, \{D_k\}_{k=1}^{N}) \approx \frac{1}{C} \left( \alpha_0 P(D_a) P(Y_a = y|x, D_a) + \sum_{k=1}^{N} \xi_k P(D_a|\theta_k^{ML}) P(Y_a = y|x, \theta_k^{ML}) \right)
\]
Implementation Details (2)

- Let’s take a look at the second term.
- Essentially it says that any user should make a prediction and that this prediction is weighted by the likelihood that a user can explain the past data of the active user.
- Thus, initially, we obtain a simple average of all user’s predictions.
- If the active user has seen some data, those users get higher weight, which agree well with the past ratings of the active user.
- Eventually, with many data for the active user, only the active user’s own model will contribute to the prediction.
Cluster Analysis Using Textual Data: Dirichlet Enhanced Latent Semantic Analysis
The Goal

- In latent semantic analysis (LSA), we aim at modeling a large corpus of *high-dimensional discrete data* from a *probabilistic* perspective.

- **The assumption:** one data point can be modeled by *latent factors*, which account for the co-occurrence of items within the data.

- We are also interested in the *clustering* structure of the data, which may benefit from the latent factors of the items.

- For example:
  - In document modelling, the data are document-word pairs.
    - **Latent factors:** topics for words
    - **Data clustering:** categories of documents
  - In collaborative filtering, the data are user ratings (for, e.g., movies).
    - **Latent factors:** categories or structures of movies
    - **Data clustering:** user interest groups
Latent Dirichlet Allocation (LDA)

- **Latent Dirichlet Allocation (LDA)** Assign a discrete latent model to words and let each document maintain a random variable $\theta$, indicating its probabilities of belonging to each topic
LDA

\[ \lambda \]

\[ \text{Dir}(\cdot | \lambda) \]

\[ M \]

\[ \theta_d \]

\[ N_d \]

\[ z_{d,n} \]

\[ \beta \]

\[ W_{d,n} \]
A Potential Problem with LDA

- Assumption: The prior for $\theta$ is a Dirichlet distribution (which is learned from data)
- Limitation: A single Dirichlet distribution is not flexible enough to represent interesting dependencies such as a clustering structure in documents

The true distribution of $\theta$ in a toy problem

The learned Dirichlet distribution in LDA
The key point of the DELSA model (Yu, K., Yu, S., Tresp, V., 2005) is to replace the single Dirichlet distribution in LDA with a nonparametric Dirichlet process prior.

We select Dirichlet-multinomial allocation (DMA), as a finite approximation to DP denoted as $\text{DP}_N$.

Inference is based on a mean field approximation.
\[ \lambda \xrightarrow{\text{Dir}(. | \lambda)} M \xrightarrow{N_d} Z_{d,n} \xrightarrow{\theta_d} W_{d,n} \xrightarrow{\beta} \text{LDA} \]

\[ \alpha_0 \xrightarrow{G} G \xrightarrow{G_0} \sim \text{Dir}(\cdot | \lambda) \]

\[ \xrightarrow{\text{DELSA}} \]

Evaluation on Toy Data

A dictionary of 200 words are associated with 5 latent topics. 100 documents are generated with 6 document clusters. $N = 100$ before learning.

Random initialization  After 1 EM step  After 5 EM step (final)
Evaluation on Toy Data: Clustering

We then vary the number of clusters from 5 to 12 and randomize the data for 20 trials. We record the detected number of clusters.

- We can correctly detect number of clusters
- The calculation is fast without overfitting
- The recovered parameter $\beta$ is very good
Document Modelling

We compare DELSA with PLSI and LDA on Reuters-21578 and 20-Newsgroup in terms of perplexity: 

$$\text{Perp}(D_t) = \exp \left( -\ln p(D_t) / \sum_d |w_d| \right).$$

- DELSA is consistently better than PLSI and LDA without overfitting
- Better for data set with strong clustering structure (like 20-Newsgroup)
Document Clustering

- We test DELSA on 20-Newsgroup data with 4 categories *autos*, *motorcycles*, *baseball* and *hockey*, each taking 446 documents. 6 clusters are found.

- Documents in one category show similar behavior.

- A very similar model to DELSEA was developed based on truncated stick breaking (Blei and Jordan, 2005)
More Models
Automatically Determining the Right Number of Clusters

The following papers employ DPs to form infinite mixture models:

The true number of mixture components is determined by the clustering effect in the Gibbs sampler:

- Infinite Mixtures of Gaussians (Rasmussen, C.E., 2000)
- Infinite Mixtures of Gaussian Processes
- Infinite Markov Networks (Beal, Ghahramani, and Rasmussen, 2002) (Teh et al. 2004)
Relational Modeling Using DPs

• Probabilistic relational models form truthful statistical representations of relational data, e.g., data stored in a relational data base.

• Effective learning can be realized using hierarchical Bayesian modeling.

• DPs have been applied to relational modeling to obtain a sharing strengths effect (Xu, Tresp, Yu and Yu, 2005) and for clustering, exploiting the relational information (Tenenbaum et al., 2005).
Conclusions

- Nonparametric Bayesian models allow for much flexibility in hierarchical modeling and other applications such as clustering.

- The most important setting is the Dirichlet mixture model.

- There is a large literature on nonparametric Bayesian modeling in the statistics fields and more needs to be explored.

- In Machine Learning one would often prefer to work with efficient approximations than with inference based on Gibbs sampling; here is an open field for more research.

- As it is obvious by the peaky structure of $G$, this is not a very nice probability density; to achieve a probability density, one might introduce another hierarchical smoothing level (Tomlinson and Escobar, 2005).

- More processes to be explored, Hierarchical DP, Dirichlet Diffusion trees, Indian buffet processes, ...

- For more on DP, look at the related tutorials by Ghahramani (2005) and Jordan (2005) and the introductory paper by Tresp and Yu (2004).
Acknowledgements

• This is joint work with Kai Yu and Shipeng Yu


