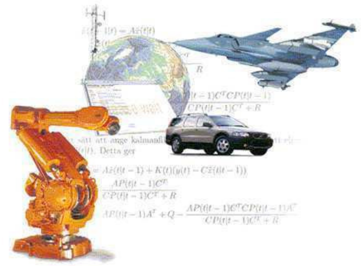
Machine Learning, Lecture 11
Bayesian nonparametrics



Fredrik Lindsten

Division of Automatic Control
Linköping University, Sweden

E-mail: lindsten@isy.liu.se

Outline

1. Summary of lecture 10
2. Parametric vs. nonparametric models
3. Mixture models – the standard example
4. The Dirichlet process
   - Stick-breaking
   - The Blackwell-MacQueen and Chinese restaurant processes
   - Dirichlet process mixture models
5. Beyond DP mixture models

Summary of lecture 10 (I/II)

The **idea** underlying Monte Carlo is to generate samples $\{z^i\}_{i=1}^M$ according to some proposal distribution $q(z)$ and possibly compute a weight for each sample, resulting in an **empirical estimate**

$$\widehat{\pi}(z) = \sum_{i=1}^{M} w^i \delta_{z^i}(z)$$

of the target distribution $\pi(z)$. This allows for approximations of general integrals according to

$$\mathrm{E}\left[g(z)\right] = \int g(z)\pi(z)dz \approx \sum_{i=1}^{M} w^i g(z^i)$$

Summary of lecture 10 (II/II)

Two “basic Monte Carlo samplers” were introduced; rejection sampling and importance sampling.

A **Markov chain Monte Carlo (MCMC)** method allows us to generate samples from an arbitrary target distribution $\pi(z)$ by simulating a Markov chain whose stationary distribution is $\pi(z)$.

A **Markov chain** $\{z^m\}_{m\geq 1}$ is a stochastic process specified by

1. Initial distribution: $z^1 \sim \mu_1(z^1)$
2. Transition kernel: $z^{m+1} \mid z^m \sim K(z^{m+1} \mid z^m)$

Two **constructive** ways of building Markov chains with a particular user-defined stationary distribution where introduced:

1. Metropolis Hastings (MH) sampler
2. Gibbs sampler

Parametric models

Parametric model,

$$Y \sim p(Y \mid \theta)$$

for some finite dimensional parameter θ.

1. Complexity/flexibility of model $\approx$ dimension of θ.
2. Can lead to over- or underfitting when there is a mismatch between the model complexity and the amount of available data!
3. Order selection is often a hard problem.

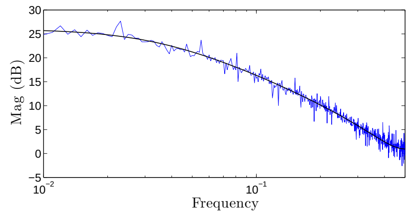
Nonparametric models

Nonparametric model – flexible model in which the complexity increases with the amount of data.

1. Attempts to avoid order selection.
2. The number of "parameters" increases with the number of data points.

ex) Empirical transfer function estimate (ETFE)

$$\widehat{G}_N(e^{i\omega}) = \frac{Y_N(\omega)}{U_N(\omega)}$$



Bayesian nonparametrics

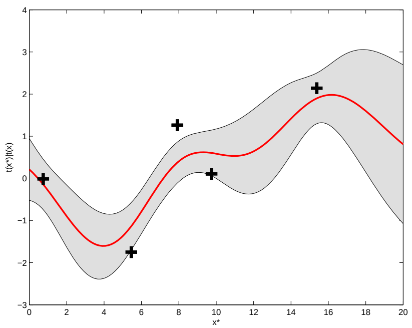
- Bayesian parametric model $=$ latent random variables (parameters).
- Bayesian nonparametric model $=$ **latent stochastic process**.

Gaussian processes

Recall Gaussian processes from lecture 5,

$$f(\cdot) \sim \mathrm{GP}(m(x), k(x, x')),$$
$$y_n = f(x_n) + e_n,$$

for $n = 1, \ldots, N$.



Stochastic processes in BNP

Many possibilities, depending on what we want to capture with the model,

1. Gaussian process
2. Dirichlet process, Chinese restaurant process
3. Pitman-Yor process
4. Beta process, Indian buffet process
5. ...

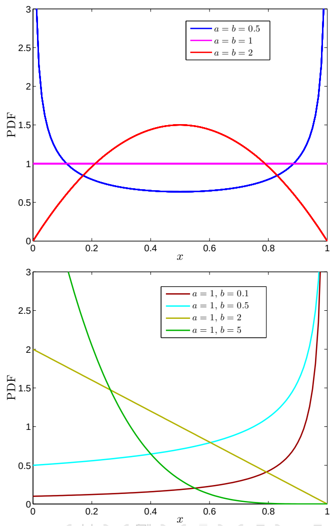
Beta distribution

Beta distribution:

$$x \sim \text{Be}(a, b)$$

Parameters: $a, b > 0$.

- Support: $0 \leq x \leq 1$.
- Often used as prior for a probability.
- Conjugate prior for Bernoulli, binomial and geometric distr.



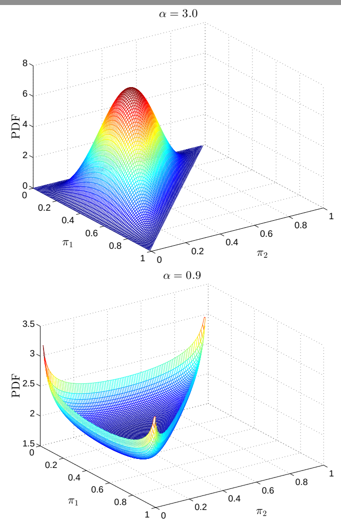
Dirichlet distribution

Dirichlet distribution:

$$\boldsymbol{\pi} \sim \text{Dir}(\alpha_1, \dots, \alpha_K)$$

Parameters: $\alpha_k > 0$.

- Support: $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$.
- A draw $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ can be interpreted as a discrete probability distribution.
- The Dirichlet distribution is a "distribution over distributions"!
- Conjugate prior for discrete and multinomial distributions.



Dirichlet process

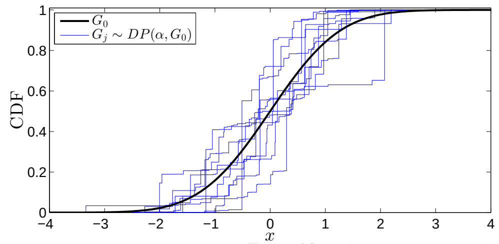
Dirichlet process:

$$G \sim \text{DP}(\alpha, G_0),$$

with base distribution G_0 and concentration parameter α.

A draw from the DP is a discrete probability distribution!

- $\mathbb{E}[G] = G_0$
- $\mathbb{V}[G] \propto (1 + \alpha)^{-1}$



Going nonparametric

Can we make this nonparametric? Define an **infinite mixture model** by letting $K \to \infty$, i.e. we get $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$, where,

$$\phi_k \overset{\text{i.i.d.}}{\sim} G_0,$$
$$\pi \sim \text{Dir}(\alpha/K, \dots, \alpha/K), \qquad K \to \infty.$$

- Will π have a proper distribution as $K \to \infty$?
- Will $\sum_{k=1}^{\infty} \pi_k = 1$?
- Will the model have clustering properties?

A better way – use a constructive definition.

Posterior updates

- Assume that $G \sim \text{DP}(\alpha, G_0)$ and $\theta_1 \sim G$.
- What can be said about the *posterior* distribution "$G \mid \theta_1$"?
- Discrete-Dirichlet-conjugacy carries over to DP!

$$G \mid \theta_1 \sim \text{DP}\left(\alpha + 1, \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha + 1}\right).$$

- Iterating the posterior update we get,

$$G \mid \theta_1, \dots, \theta_n \sim \text{DP}\left(\alpha + n, \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right).$$

DP mixture model with CRP

Note that

$$p(z_{1:n}) = \prod_{i=1}^{n} p(z_i \mid z_{1:i-1}).$$

We can thus write the DP mixture model in terms of the CRP,

$$z_{n+1} \mid z_1, \dots, z_n = \begin{cases} k & \text{w.p.} \frac{m_k}{\alpha+n}, \\ K+1 & \text{w.p.} \frac{\alpha}{\alpha+n}, \end{cases}$$
$$\phi_k \overset{\text{i.i.d.}}{\sim} G_0, \qquad k = 1, 2, \dots$$
$$y_n \mid \{z_n = k\}, \phi_k \sim p(y_n \mid \phi_k).$$

Similar to a finite mixture model with latent variables, but now the latent variables are given by the CRP.

Inference

Inference for DP mixture models

- Different representations (stick-breaking, Blackwell-MacQueen, CRP) give rise to different algorithms.
- Different inference tools – MCMC (Gibbs/split-merge), VB, Particle filters, Maximization-Expectation, ...

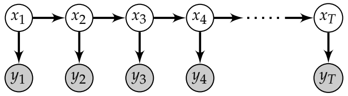
ex) Gibbs sampler using CRP (Neal, 2000).

Given $\{y_n\}_{n=1}^{N}$, iterate:

- For $n = 1, \dots, N$ draw: $z_n \mid z_{-n}, y_n, \{\phi_k\}_{k=1}^{K}$;
- For $k \in \{z_1, \dots, z_n\}$ draw: $\phi_k \mid \{\text{all } y_n \text{ s.t. } z_n = k\}$.

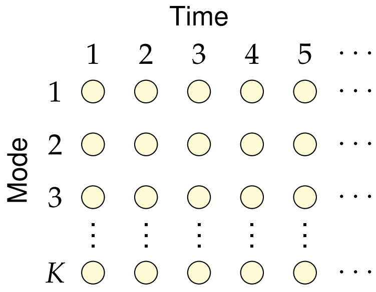
Hidden Markov model

Hidden Markov model (HMM), $x_t \in \{1, \dots, K\}$,



$$P(x_{t+1} = \ell \mid x_t = k) = \pi_{k\ell},$$
$$y_t \mid \{x_t = k\} \sim p(y_t \mid \phi_k).$$

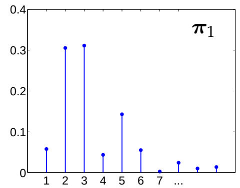
$$\Pi = \begin{bmatrix} - \pi_1 - \\ - \pi_2 - \\ \vdots \\ - \pi_K - \end{bmatrix}$$



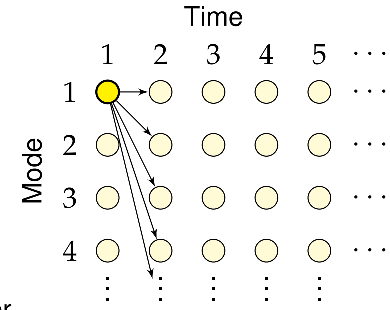
Infinite hidden Markov model

Infinite hidden Markov model, $x_t \in \mathbb{N}$,

- Stick-breaking for $\boldsymbol{\pi}_\ell$ for $\ell = 1, 2, \dots$

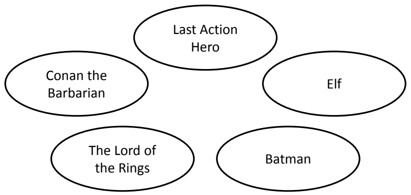


- Hierarchical DP – tie mode transition distributions together.
- Share sparsity patterns.



Feature models

Can we cluster movies?



We might get a more accurate model by using features, e.g.,

- Fantasy • Arnold Schwarzenegger • Elves • Action • Comedy.

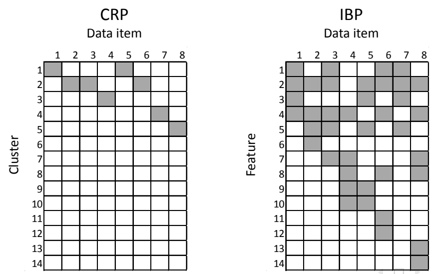
Feature models

Binary latent variables,

$$z_{nk} = \begin{cases} 1 & \text{if item } n \text{ has feature } k, \\ 0 & \text{otherwise,} \end{cases}$$

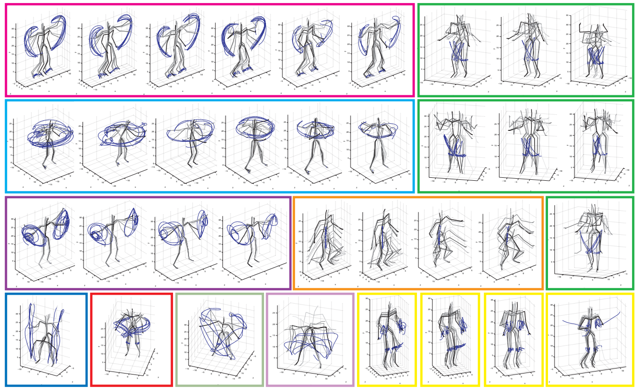
for $n = 1, \dots, N$ and $k = 1, \dots, K$.

Going nonparametric ("$K \to \infty$") $\Rightarrow$ the indian buffet process (IBP)



Motion capture with the IBP

Learning from multiple time series using the IBP



E. B. Fox, E. B. Sudderth, M. I. Jordan, A. S. Willsky. **Sharing Features among Dynamical Systems with Beta Processes**, Proceeding of Neural Information Processing Systems (NIPS), Vancouver, Canada December 2009.

A few concepts to summarize lecture 11

- Nonparametric models allow the complexity to increase with the amount of data
- Bayesian nonparametrics = latent stochastic processes
- Dirichlet process,
  - A draw from the Dirichlet process is a (random) discrete probability distribution
  - Dirichlet process mixture model for clustering
  - Hierarchical DPs can be used to construct an “infinite” HMM
- Indian buffet process for feature models

Further reading

M. I. Jordan.
Bayesian nonparametric learning: Expressive priors for intelligent systems.
In R. Dechter, H. Geffner, and J. Halpern, editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*. College Publications, 2010.

R. M. Neal.
Markov chain sampling methods for Dirichlet process mixture models.
Journal of Computational and Graphical Statistics, 9(2):249–265, 2000.

E. B. Sudderth.
Graphical Models for Visual Object Recognition and Tracking.
PhD thesis, Massachusetts Institute of Technology, 2006.