Bayesian Nonparametrics

Charlie Frogner

9.520 Class 11

March 14, 2012

About this class

Last time Bayesian formulation of RLS, for regression. (Basically, a normal distribution.)

This time a more complicated probability model: the Dirichlet Process.

And its application to clustering.

And also more Bayesian terminology.

Plan

- Dirichlet distribution + other basics
- The Dirichlet process
  - Abstract definition
  - Stick Breaking
  - Chinese restaurant process
- Clustering
  - Dirichlet process mixture model
  - Hierarchical Dirichlet process mixture model

Gamma Function and Beta Distribution

The Gamma function

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

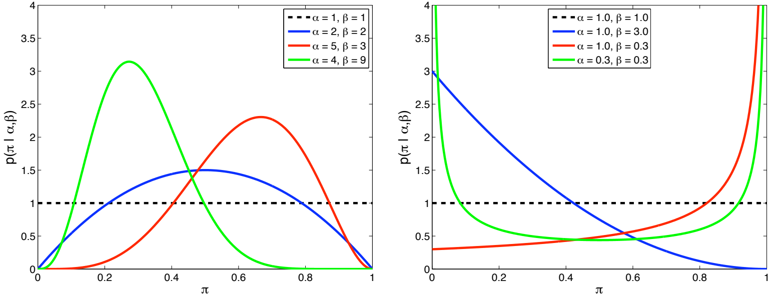
Extends factorial function to $\mathbb{R}^+$: $\Gamma(z+1) = z\Gamma(z)$.

Beta Distribution

$$P(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{(\alpha-1)}(1-x)^{(\beta-1)}$$

for $x \in [0,1]$, $\alpha > 0$, $\beta > 0$.
(Mean: $\frac{\alpha}{\alpha+\beta}$, variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.)

Beta Distribution



For large parameters the distribution is unimodal. For small parameters it favors biased binomial distributions.

Dirichlet Distribution

Generalizes Beta distribution to the K-dimensional simplex $\mathbb{S}^K$.

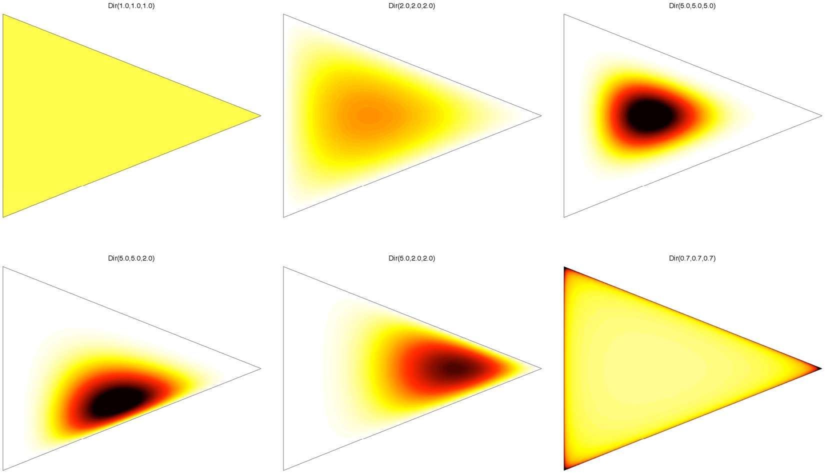
$$\mathbb{S}^K = \{x \in \mathbb{R}^K : \sum_{i=1}^{K} x_i = 1,\ x_i \geq 0\ \forall i\}$$

Dirichlet distribution

$$P(x|\alpha) = P(x_1, \dots, x_K) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\sum_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^{K} (x_i)^{\alpha_i - 1}$$

where $\alpha = (\alpha_1, \dots, \alpha_K)$, $\alpha_i > 0\ \forall i$, $x \in \mathbb{S}^K$.
We write $x \sim \mathrm{Dir}(\alpha)$, i.e. $x_1, \dots, x_K \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$.

Dirichlet Distribution



Dir(1.0,1.0,1.0)

Dir(2.0,2.0,2.0)

Dir(5.0,5.0,5.0)

Dir(5.0,5.0,2.0)

Dir(5.0,2.0,2.0)

Dir(0.7,0.7,0.7)

Properties of the Dirichlet Distribution

- Mean

$$\mathbf{E}[x_i] = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}.$$

- Variance

$$\mathbf{Var}[x_i] = \frac{\alpha_i(\sum_{i\neq j} \alpha_j)}{(\sum_{j=1}^{K} \alpha_j)^2(1 + \sum_{j=1}^{K} \alpha_j)}.$$

- Covariance

$$\mathbf{Cov}(x_i, x_j) = \frac{\alpha_i \alpha_j}{(\sum_{j=1}^{K} \alpha_j)^2(1 + \sum_{j=1}^{K} \alpha_j)}.$$

- Marginals: $x_i \sim \text{Beta}(\alpha_i, \sum_{j\neq i} \alpha_j)$
- Aggregation: $(x_1 + x_2, \ldots, x_k) \sim \text{Dir}(\alpha_1 + \alpha_2, \ldots, \alpha_K)$

Multinomial Distribution

If you throw n balls into k bins, the distribution of balls into bins is given by the multinomial distribution.

Multinomial distribution

Let $p = (p_1, \ldots, p_K)$ be probabilities over K categories and $C = (C_1, \ldots, C_K)$ be category counts. C_i is the number of samples in the ith category, from n independent draws of a categorical variable with category probabilities p. Then

$$P(C|n,p) = \frac{n!}{\prod_{i=1}^K C_i!} \prod_{i=1}^K p_i^{C_i}.$$

For $K = 2$ this is the binomial distribution.

An idea

Treat the Dirichlet distribution as a distribution on probabilities: each sample $\theta \sim \mathrm{Dir}(\alpha)$ defines a K-dimensional multinomial distribution.

$$x \sim \mathrm{Mult}(\theta), \theta \sim \mathrm{Dir}(\alpha)$$

An idea

Treat the Dirichlet distribution as a distribution on probabilities: each sample $\theta \sim \mathrm{Dir}(\alpha)$ defines a K-dimensional multinomial distribution.

$$x \sim \mathrm{Mult}(\theta), \theta \sim \mathrm{Dir}(\alpha)$$

Posterior on θ:

$$\theta | x \sim \mathrm{Dir}(\alpha + x)$$

Conjugate Priors

Say $x \sim F(\theta)$ (the **likelihood**) and $\theta \sim G(\alpha)$ (the **prior**).

Conjugate prior

G is a **conjugate prior** for F if the **posterior** $P(\theta|x, \alpha)$ is in the same family as G. (E.g. if F is Gaussian then $P(\theta|x, \alpha)$ should also be Gaussian.)

So the Dirichlet distribution is a conjugate prior for the multinomial.

Plan

- Dirichlet distribution + other basics
- The Dirichlet process
    - Abstract definition
    - Stick Breaking
    - Chinese restaurant process
- Clustering
    - Dirichlet process mixture model
    - Hierarchical Dirichlet process mixture model

Parametric vs. nonparametric

- **Parametric**: fix parameters independent of data.
- **Nonparametric**: effective number of parameters can grow with the data.

E.g. density estimation: fitting Gaussian vs. parzen windows.
E.g. Kernel methods are nonparametric.

Dirichlet Process

Want: distribution on all K-dimensional simplices (for all K).

Informal Description

X is a space, F is a probability distribution on X and $\mathcal{F}(X)$ is the set of all possible distributions on X.
A **Dirichlet Process** gives a distribution over $\mathcal{F}(X)$. A sample path from a DP is an element $F \in \mathcal{F}(X)$. F can be seen as a (random) probability distribution on X.

Dirichlet Process

Want: distribution on all K-dimensional simplices (for all K).

Formal Definition

Let X be a space and H be the base measure on X. F is a sample from the Dirichlet Process DP(α, H) on X if its finite-dimensional marginals have the Dirichlet distribution:

$$(F(B_1), \ldots, F(B_K)) \sim \mathrm{Dir}(\alpha H(B_1), \ldots, \alpha H(B_2))$$

for all partitions $B_1, \ldots, B_K$ of X (for any K).

Stick Breaking Construction

Explicit construction of a DP.

Let $\alpha > 0$, $(\pi_i)_{i=1}^{\infty}$ such that

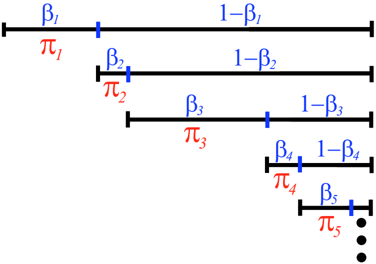
$$p_i = \beta_i \prod_{j=1}^{i-1} (1 - \beta_j) = \beta_i (1 - \sum_{j=1}^{i-1} p_j)$$

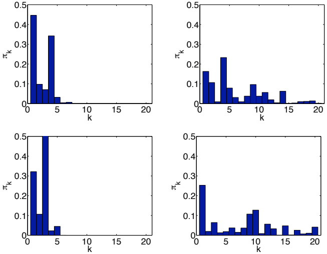
where $\beta_i \sim \mathit{Beta}(1, \alpha)$, for all i.
Let H be a distribution on X and define

$$F = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$$

where $\theta_i \sim H$, for all i.

Stick Breaking Construction: Interpretation





$\alpha = 1$ $\alpha = 5$

The weights π partition a unit-length *stick* in an infinite set: the i-th weight is a random proportion β_i of the stick remaining after sampling the first $i - 1$ weights.

Stick Breaking Construction (cont.)

It is possible to prove (Sethuraman ’94) that the previous construction returns a DP and conversely a Dirichlet process is discrete almost surely.

Chinese Restaurant Process

There is an infinite (countable) set of tables.

- First customer sits at the first table.
- Customer i sits at table j with probability

$$\frac{n_j}{\alpha + i + 1},$$

where n_j is the number of customers at table j, and i sits at the first open table with probability

$$\frac{\alpha}{\alpha + i + 1}$$

The Role of the Strength Parameter

Note that $\mathbb{E}[\beta_i] = 1/(1 + \alpha)$.

- for small α, the first few components will have all the mass.
- for large α, F approaches the distribution H assigning uniform weights to the samples θ_i.

Number of Clusters and Strength Parameter

It is possible to prove (Antoniak '77??) that the number of components with positive count grows as

$$\alpha \log n$$

as we increase the number of samples n.

Another idea

Clustering with the K-dimensional Dirichlet: take each sample $\theta \sim \text{Dir}(\alpha)$ to define a K-dimensional categorical (instead of multinomial) distribution.

$$x \sim G(\phi), \phi \sim \text{Cat}(\theta), \theta \sim \text{Dir}(\alpha)$$

(G is a a distribution on observation space X, say, Gaussian.)

θ_i is the probability of x coming from the ith cluster.

Another idea

Clustering with the K-dimensional Dirichlet: take each sample $\theta \sim \text{Dir}(\alpha)$ to define a K-dimensional categorical (instead of multinomial) distribution.

$$x \sim G(\phi), \phi \sim \text{Cat}(\theta), \theta \sim \text{Dir}(\alpha)$$

(G is a a distribution on observation space X, say, Gaussian.)

θ_i is the probability of x coming from the ith cluster.

Another idea

Clustering with the K-dimensional Dirichlet: take each sample $\theta \sim \text{Dir}(\alpha)$ to define a K-dimensional categorical (instead of multinomial) distribution.

$$x \sim G(\phi), \phi \sim \text{Cat}(\theta), \theta \sim \text{Dir}(\alpha)$$

(G is a a distribution on observation space X, say, Gaussian.)

θ_i is the probability of x coming from the ith cluster.

Another idea

Clustering with the Dirichlet Process: take each sample $\theta \sim \text{DP}(\alpha, H)$ to define a K-dimensional categorical (instead of multinomial) distribution.

$$x \sim G(\phi), \phi \sim \text{Cat}(\theta), \theta \sim \text{DP}(\alpha, H)$$

(G is a a distribution on observation space X, say, Gaussian. H can be uniform on $\{1, \ldots, K\}$.)

Another idea

Clustering with the Dirichlet Process: take each sample $\theta \sim \mathrm{DP}(\alpha, H)$ to define a K-dimensional categorical (instead of multinomial) distribution.

$$x \sim G(\phi), \phi \sim \mathrm{Cat}(\theta), \theta \sim \mathrm{DP}(\alpha, H)$$

(G is a a distribution on observation space X, say, Gaussian. H can be uniform on $\{1, \ldots, K\}$.)

Another idea

Clustering with the Dirichlet Process:

$$x \sim G(\phi), \phi \sim \text{Cat}(\theta), \theta \sim \text{DP}(\alpha, H)$$

This is the **Dirichlet Process mixture model.**

Hierarchical Dirichlet Process

What if we want to model **grouped data**, each group corresponding to a different DP mixture model?

Hierarchical Dirichlet Process

For each $i \in \{1, \ldots, n\}$, draw x_i according to

$$x_i \sim G(\phi), \phi \sim \mathrm{Cat}(\theta), \theta \sim \mathrm{DP}(\alpha, H_0), \alpha \sim \mathrm{DP}(\gamma, H).$$

Hierarchical Dirichlet Process

What if we want to model **grouped data**, each group corresponding to a different DP mixture model?

Hierarchical Dirichlet Process

For each $i \in \{1, \ldots, n\}$, draw x_i according to

$$x_i \sim G(\phi), \phi \sim \mathrm{Cat}(\theta), \theta \sim \mathrm{DP}(\alpha, H_0), \alpha \sim \mathrm{DP}(\gamma, H).$$

Conclusions

- Dirichlet distribution gives a distribution over the K-simplex.
- Dirichlet is conjugate to the multinomial, which makes inference in the Dirichlet/multinomial model easy.
- Dirichlet process generalizes the Dirichlet distribution to countably infinitely many components.
    - Every finite marginal of the DP is Dirichlet distributed.
- Complexity of the DP is controlled by the strength parameter α.
- The posterior distribution cannot be found analytically. Approximate inference is needed.

References

This lecture heavily draws (sometimes literally) from the list of references below, which we suggest as further readings.
Figures are taken either from Sudderth PhD thesis or Teh Tutorial.

Main references/sources:

- Yee Whye Teh, *Tutorial in the Machine Learning Summer School*, and his notes *Dirichlet Processes*.
- Erik Sudderth, PhD Thesis.
- Gosh and Ramamoorthi, *Bayesian Nonparametrics*, (book).

See also:

- Zoubin Ghahramani, Tutorial ICML.
- Michael Jordan, Nips Tutorial.
- Rasmussen, Williams, *Gaussian Processes for Machine Learning*, (book).
- Ferguson, paper in Annals of Statistics.
- Sethuraman, paper in *Statistica Sinica*.
- Berlinet, Thomas-Agnan, *RKHS in Probability and Statistics*, (book).

APPENDIX

Dirichlet Process (cont.)

A partition of X is a collection of subsets $B_1, \ldots, B_N$ is such that, if $B_i \cap B_j = \emptyset$, $\forall i \neq j$ and $\cup_{i=1}^{N} B_i = X$.

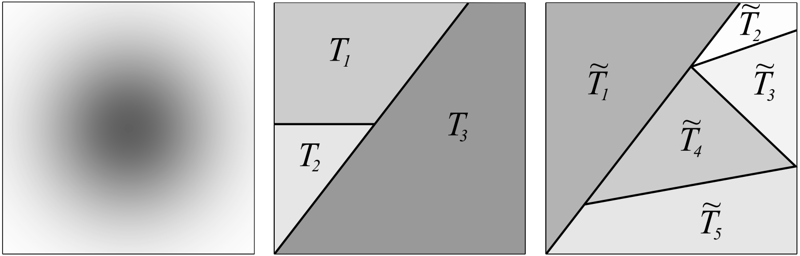
Definition (Existence Theorem)

Let $\alpha > 0$ and H a probability distribution on X.
One can prove that there exists a unique distribution $DP(\alpha, H)$ on $\mathcal{F}(X)$ such that, if $F \sim DP(\alpha, H)$ and $B_1, \ldots, B_N$ is a partition of X then

$$(F(B_1), \ldots, F(B_N)) \sim \mathrm{Dir}(\alpha H(B_1), \ldots, \alpha H(B_N)).$$

The above result is proved (Ferguson '73) using Kolmogorov's Consistency theorem (Kolmogorov '33).

Dirichlet Processes Illustrated



Properties of Dirichlet Processes

Hereafter $F \sim DP(\alpha, H)$ and A is a measurable set in X.

- Expectation: $\mathbb{E}[F(A)] = \alpha H(A)$.
- Variance: $\mathbb{V}[F(A)] = \frac{H(A)(1-H(A))}{\alpha+1}$

Properties of Dirichlet Processes (cont.)

- Posterior and Conjugacy: let $x \sim F$ and consider a fixed partition $B_1, \dots, B_N$, then

$$P(F(B_1), \dots, F(B_N) | x \in B_k) =$$
$$\mathrm{Dir}(\alpha H(B_1), \dots, \alpha H(B_k) + 1, \dots, \alpha H(B_N)).$$

It is possible to prove that if $S = (x_1, \dots, x_n) \sim F$, and $F \sim DP(\alpha, H)$, then

$$P(F|S, \alpha, H) = DP\left(\alpha + n, \frac{1}{n+\alpha}\left(\alpha H + \sum_{i=1}^{n} \delta_{x_i}\right)\right)$$

A Qualitative Reasoning

From the form of the posterior we have that

$$\mathbb{E}(F(A)|S,\alpha,H) = \frac{1}{n+\alpha}\left(\alpha H(A) + \sum_{i=1}^{n} \delta_{x_i}(A)\right).$$

If $\alpha < \infty$ and $n \to \infty$ one can argue that

$$\mathbb{E}(F(A)|S,\alpha,H) = \sum_{i=1}^{\infty} \pi_i \delta_{x_i}(A)$$

where $(\pi_i)_{i=1}^{\infty}$ is the sequence corresponding to the limit $\lim_{n\to\infty} C_i/n$ of the empirical frequencies of the observations $(x_i)_{i=1}^{\infty}$.

If the posterior concentrates about its mean the above reasoning suggests that the obtained distribution is discrete.