Dirichlet Distribution, Dirichlet Process and Dirichlet Process Mixture

Leon Gu

CSD, CMU

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.

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

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.)

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

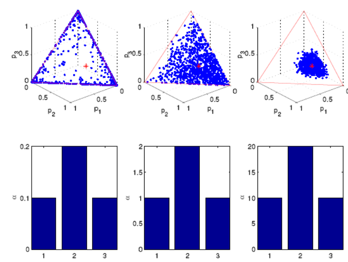
- Beta is the conjugate prior of Binomial.

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.

Mean and Variance



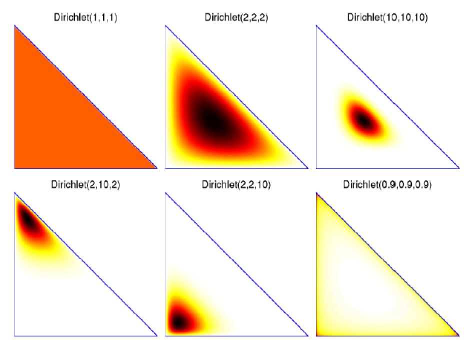
- The base measure determines the mean distribution;
- Altering the scale affects the variance.

$$E(p_i) = \frac{\alpha_i}{\sigma} = \alpha'_i \quad (1)$$

$$Var(p_i) = \frac{\alpha_i(\sigma - \alpha)}{\sigma^2(\sigma + 1)} = \frac{\alpha'_i(1 - \alpha'_i)}{(\sigma + 1)} \quad (2)$$

$$Cov(p_i, p_j) = \frac{-\alpha_i \alpha_j}{\sigma^2(\sigma + 1)} \quad (3)$$

Another Example



- A Dirichlet with small concentration σ favors extreme distributions, but this prior belief is very weak and is easily overwritten by data.
- As $\sigma \rightarrow \infty$, the covariance $\rightarrow 0$ and the samples $\rightarrow$ base measure.

- posterior is also a Dirichlet

$$p(P = \{p_i\} \mid \alpha_i) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \prod_i p_i^{\alpha_i - 1} \tag{4}$$

$$P(x_1, \ldots, x_k \mid n, p_1, \ldots, p_k) = \frac{n!}{\prod_{i=1}^k x_i!} p_i^{x_i} \tag{5}$$

$$p(\{p_i\} | x_1, \ldots, x_k) = \frac{\prod_i \Gamma(\alpha_i + x_i)}{\Gamma(N + \sum_i \alpha_i)} \prod_i p_i^{\alpha_i + x_i - 1} \tag{6}$$

- marginalizing over parameters (condition on hyper-parameters only)

$$p(x_1, \ldots, x_k | \alpha_1, \ldots, \alpha_k) = \frac{\prod_i \alpha_i^{x_i})}{\sigma^N}$$

- prediction (conditional density of new data given previous data)

$$p(new\_result = j | x_1, \ldots, x_k, alpha_1, \ldots, \alpha_k) = \frac{\alpha_j + x_j}{\sigma + N}$$

Dirichlet Process

Suppose that we are interested in a simple generative model (monogram) for English words. If asked "what is the next word in a newly-discovered work of Shakespeare?", our model must surely assign non-zero probability for *words that Shakespeare never used before*. Our model should also satisfy a consistency rule called exchangeability: the probability of finding a particular word at a given location in the stream of text should be the same everywhere in thee stream.

Dirichlet process is a model for a stream of symbols that 1) satisfies the exchangeability rule and that 2) allows the vocabulary of symbols to grow without limit. Suppose that the mode has seen a stream of length F symbols. We identify each symbol by an unique integer $w \in [0, \infty)$ and F_w is the counts if the symbol. Dirichlet process models

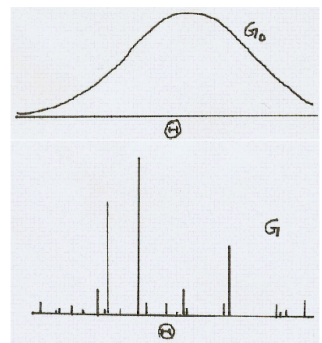
- the probability that the next symbol is symbol w is

$$\frac{F_w}{F + \alpha}$$

- the probability that the next symbol is never seen before is

$$\frac{\alpha}{F + \alpha}$$

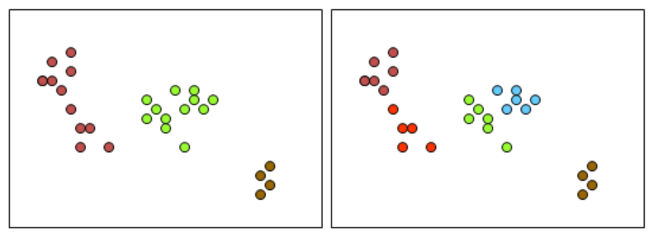
$$G \sim DP(\cdot | G_0, \alpha)$$



- Dirichlet process generalizes Dirichlet distribution.
- G is a distribution function in a space of infinite but countable number of elements.
- G_0: base measure; α: concentration

- pretty much the same as Dirichlet distribution
    - expectation and variance
    - the posterior is also a Dirichlet process $DP(|G_0, \alpha)$
    - prediction
    - integration over G (data conditional on G_0 and αonly)
    -
- equivalent constructions
    - Polya Urn Scheme
    - Chinese Restaurant Process
    - Stick Breaking Scheme
    - Gamma Process Construction

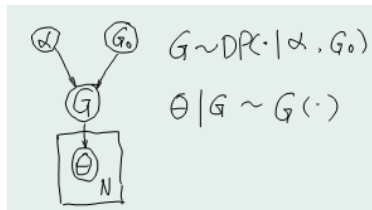
Dirichlet Process Mixture



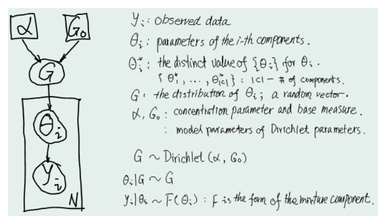
- How many clusters?
- Which is better?

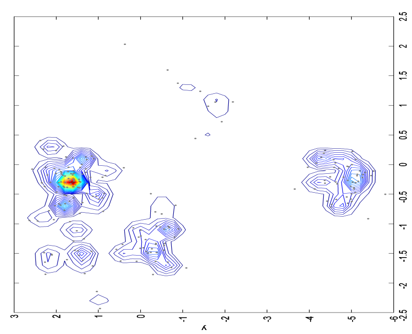
Graphical Illustrations

Multinomial-Dirichlet Process



Dirichlet Process Mixture




x
y
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
3
-3
-4
-5
-6