DirichletProcesses

Dirichlet Processes

overview

- useful for discrete set with infinite possibilities, but where fewer are prefered
- can be thought of as distribution over categories

background

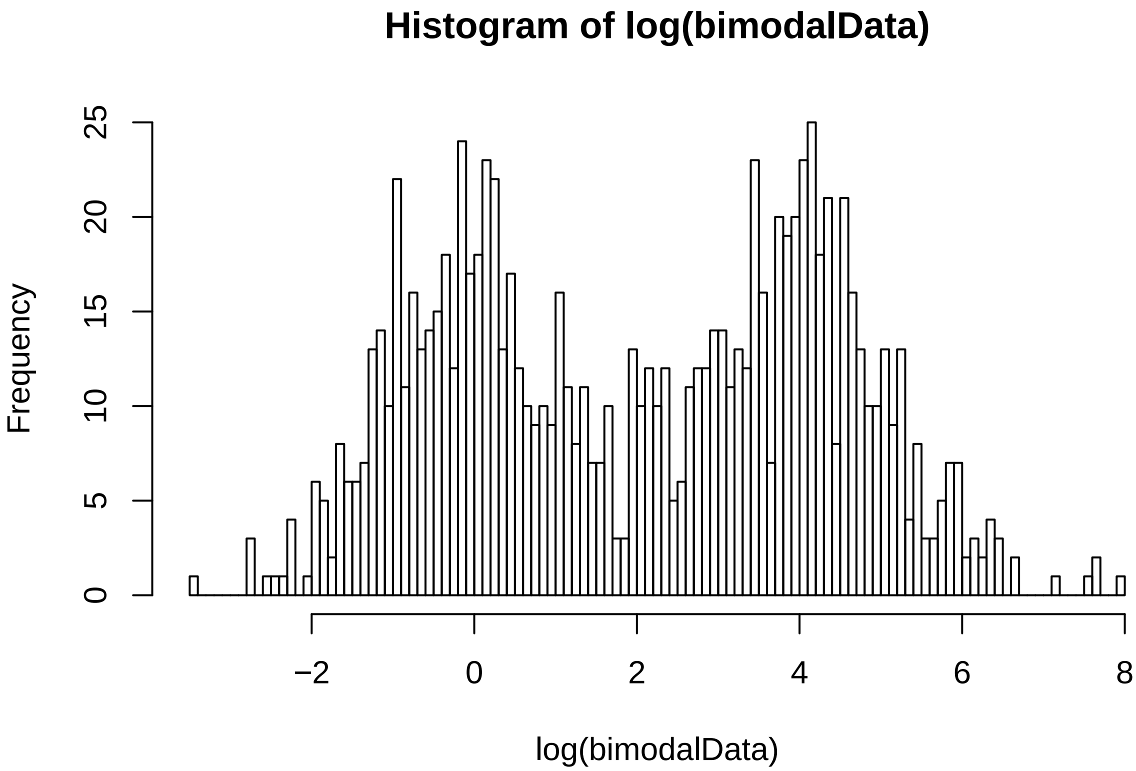
mixture model

- assume data is generated by k mixture components
- (compositions = categories)
  - each category is its own distribution
- generative model
  - $Y_i \sim Gauss(\mu, \sigma^2)$
  - each Y_i sampled from normal distribution
- $\exists$ prior distribution on μ, σ^2
  - let $\sigma = 1$, $\mu \sim Gauss(0, 1)$
  - so now there's a distribution on the parameters of Y_i
  - strongly resists means falling far from 0

graphical model

```
        h = (0, 1, sigma=1)
        |
        mu
/   /   |   \   \
Y1  Y2  Y3  Yi...Yn
```

- shows the dependencies
  - first you need hyperparameters in h
  - then you need to sample mu
  - then you can sample Y_i
- mixture model adds an extra step:
  - assume data came from 1 of k components
  - assume for now that each one was Gaussian
- assume the data looks something like this (density)



- how do we construct 1 distribution for the whole dataset?
  - need to combine 2 gaussians
- method for combining:
  - flip a coin, if it's heads, sample from one distribution, if it's tails, sample from the other
  - so the height of each distribution goes down, but they still add up to 1
  - this is true as long as the higher level distribution adds up to 1

  - component weights: odds on the coin
  - this is a mixture model

clustering/categories

- model with mixing distributions
  - mixture distribution is a discrete R.V.
  - formalized as vector of weights
- how to learn mixture distributions
  - let Z_i be r.v.$\sim Categorical_k(\Theta)$

  - indexed over datapoints
- **categorical distribution**
  - biased die with k outcomes
- so $Y_i \sim Gauss(\mu_{z_i}, \sigma^2_{z_i})$
- now what if we don't know k
  - don't know how many categories there will be
  - so we make k infinite
- NB: when we see infinite in this context, go from thinking about distributions to thinking about processes
  - we need an algorithmic process
  - we can always get more precision out of the algorithm if needed

stick breaking process

- how do we ensure that $\sum_{\forall\Theta} \Theta = 1$
  - choose first theta from some distribution between [0,1]
    - call this π_1
  - then choose π_2 from $1 - \pi_1$
  - repeat k times
- in this process, probability of stopping at component k is :
  - $(1-\pi_1)(1-\pi_2)...(1-\pi_{k-1})\pi_k$
  - $P(\Theta_j) = \pi_j \prod_{k=1}^{j-1}(1-\pi_k)$
- **this is the Dirichlet process**
- favors stopping earlier (product of fewer fractions)
- we use the Beta distribution to sample each π.
  - Beta takes two parameters (pseudocounts)
  - let them be ψ_h, ψ_t (pretend counts of heads (h) and tails (t))
  - pseudocounts may be <1
  - the mean is always $\frac{\psi_h}{\psi_h+\psi_t}$
- lets you sample coin weights π_i
- usually written as $DP(G_0, \alpha)$ where $\pi_i \sim Beta(1, \alpha)$
- this is just a distribution over infinite vectors of probabilities
  - a distribution over distributions
- G_0 is the prior distribution on parameters of each component
  - how you sample μ (in our case) if you get to a component that you've never reached before
  - called the **base distribution** or **base measure**
  - it is a distribution over kinds of things you want back

different approach

- assume this graphical model (biased coin):

```
        Theta ~ Beta(1,1)
/   /    |   \   \
flip_1 f_2 f_3 f_4... f_n
```

- suppose you don't know Θ (the weight of the coin)
  - first sample Θ from $Beta(1,1)$
  - then sample f from $Bernoulli(\Theta)$
  - flips are independent
- **BUT** if you don't know Θ then the flips are **not** independent
  - informationally entangled when you observe data
  - e.g. HHHHHTHHTHHH would make you think H is more likely than tails (i.e. heads-biased coin)
- so:

$$P(\Theta|f,\psi_h,\psi_t) = \frac{P(f|\Theta)P(\Theta|\psi_h,\psi_t)}{\int_{\forall\Theta} P(f|\Theta)P(\Theta|\psi_h,\psi_t)d\Theta}$$

sequential update scheme

- Polya urn representation

- assume a fair coin, flip it, add 1 to the resulting side and renormalize
- i.e. start with .5,.5 heads and tails
  - say you get heads, now .67, .33
  - say you get tails next, back to .5,.5
  - etc. etc.
- if you follow this scheme, P(next outcome) has the same distribution as the Bayesian formula generative model above
- this coin scheme is for the Beta-Binomial distribution

Chinese restaurant process

- for Dirichlet, imagine a restaurant with infinite tables
  - each table has a dish that is served at that table
  - first customer sits at any table
  - second customer sits at new table with probability $\frac{\alpha}{1+\alpha}$ and at the same table as the first with probability $\frac{1}{1+\alpha}$
  - after that, the $n+1$th customer sits at a new table with probability $\frac{\alpha}{n+\alpha}$ and at occupied table k with probability $\frac{n_k}{n+\alpha}$ where n_k is the number of people currently at table k
- label the tables with observations from the base distribution
  - then this is also the same distribution (Dirichlet process)
- going back to the mixture model:
  - each flip resulting from a Θ partition is analogous to each μ sampled
  - each customer is analogous to the z_i's
  - observations drawn from a table are parametrized by μ_i