Dirichlet Process

or, K-means without K

I. Holmes

Department of Bioengineering
University of California, Berkeley

Spring semester

Outline

K-means (yet) again

We presented K-means as a special case of the EM algorithm

- Model: mixture of Gaussians
- Parameters: Gaussian parameters
- Missing data: cluster = mixture component
- Counts: post.prob. of each component

Note that it's not Bayesian:

$$\hat{\theta} = \text{argmax}_\theta \sum_X P(X, Y|\theta)$$

For example, there's no explicit prior on θ...

K-means made (a bit) Bayesian

Let's put a prior on θ:

$$\begin{aligned}\hat{\theta} &= \text{argmax}_\theta P(\theta|Y) \\ &= \text{argmax}_\theta \sum_X \frac{P(X,Y|\theta)P(\theta)}{P(Y)} \\ &= \text{argmax}_\theta \sum_X P(X,Y|\theta)P(\theta)\end{aligned}$$

OK so far, but a good Bayesian might also be interested in more details about the posterior distribution e.g. the moments

$$\langle \theta^n \rangle_{P(\theta|Y)}$$

K-means made (a bit) Bayesian

What should the prior be?
The "easy" choice is a conjugate Normal-gamma prior.
Suppose each mixture component has mean μ and precision τ, and generates samples y, then

$$\begin{aligned} \tau &\sim \text{Gamma}(\alpha, \beta) \\ \mu \mid \tau &\sim \text{Normal}(\epsilon, (\lambda\tau)^{-1/2}) \\ y \mid \tau, \mu &\sim \text{Normal}(\mu, \tau^{-1/2}) \end{aligned}$$

This leads to straightforward modifications of our EM algorithm (α, β, λ and ϵ play the role of pseudocounts).

K-means made (more) Bayesian

The Good Bayesian is not content with $\hat{\theta}$.
The Good Bayesian wants $P(\theta|Y)$, and maybe $P(X|Y)$ too.

- We can use Gibbs sampling for this
- Note that, given X (cluster assignments), we can compute

$$P(Y|X) = \int P(\theta)P(Y|X,\theta)d\theta$$

  and hence $P(\theta|X,Y)$.
- Therefore we only need to sample cluster assignments

Complete Gibbs sampler for K-component mixture

Joint likelihood of data and cluster assignments (assuming uniform mixture weights)

$$P(X,Y) = P(X)P(Y|X,\theta) = \left(\frac{1}{K}\right)^N \prod_{k=1}^{K} \int P(\theta_k) \left(\prod_{i:X_i=k} P(Y_i|\theta_k) \right) d\theta$$

Gibbs-sampling algorithm:

- Remove a random datapoint from its cluster
- Calculate the likelihood for placing the datapoint in any existing cluster
- Randomly sample the datapoint's new cluster from these likelihoods

Why K-means can never be truly Bayesian

- We Still Need To Know K (the number of mixture components)
- Of course we can introduce a prior $P(K)$
- However, doing MCMC over K is tricky, because changing K changes the dimensionality of θ
- There are ways around this, e.g. reversible-jump MCMC
- The Dirichlet Process is (arguably) a more elegant solution

K-multinomials: motivation

- Generative model for clustering is a mixture distribution
  - First, randomly choose a component
  - Next, use that component to generate your data
- Before going further, it is useful to have a physical analogy for each component distribution
  - Physical analogy for a Gaussian is possible, but...
  - Physical analogy for a multinomial is much easier:

K-multinomials vs K-Gaussians

- Mixture of K Gaussians $\rightarrow$ Mixture of K multinomials
- Physical analogy:
  - Bag of (loaded) dice, each D-sided
  - Pick a die from the bag
  - Roll it N times; count outcomes
  - This corresponds to one D-dimensional datapoint
- EM algorithm is actually quite similar (even simpler)
  - Each casino punter picks a die from the bag
    - Rolls it N times, then puts it back in the bag
  - Missing data (X) specifies which die each punter chose
  - Observed data (Y) are the vectors of D counts for each punter
  - Parameters (θ) are the probabilities for each of the dice
  - Likelihood $P(X, Y|\theta)$ is a bit different from the Gaussian...
    - Exercise: derive the EM algorithm for K-multinomials

K-multinomials: the Bayesian version

- Again, we need a prior for each model component θ
- Again, an easy choice is the conjugate to $P(y|\theta)$
  - conjugate to the multinomial = the Dirichlet distribution
- Analogy: the Dirichlet represents a dice-making machine
  - For $D = 2$, the Beta distribution, a coin-minting machine
- Note that Dirichlet distribution $\neq$ Dirichlet process
  - We could use (Gaussian,Normal-Gamma) instead of (Multinomial,Dirichlet)
  - DP and DD *are* related, but in a somewhat technical sense
- Key point is that it's useful to be able to directly compute

$$P(Y) = \int_{\theta} P(Y|\theta)P(\theta)d\theta$$

for an individual component θ generating datapoints Y.

Infinite-component mixtures

The general idea is that, instead of trying to estimate the number of mixture components K, we will assume that there are an infinite number of components.

What happens if we set $K = \infty$?

- With $K = \infty$, there are an infinite number of datapoint→component assignments
  - Infinite number of dice in the bag (or "on the bench"?)
  - So, infinite number of possible explanations
- Suppose $X = (x_1, x_2, x_3, x_4, x_5)$ are the dice used by the first five consecutive punters
- In some sense, $(5, 2, 5, 10, 2)$ is the same as $(104, 300, 104, 2095, 300)$
- All we really care about is the partition — not the exact labeling. The number of partitions is always finite.
  - Both the above cases have the partition $\{(1\ 3)(2\ 5)(4)\}$
- Note that the problem of identifiability arises even with finite K. Only the partition matters, really...

The Dirichlet process

- The DP can be conceived of as an infinite mixture, $K = \infty$
  - But not with uniform component weights!
  - Some dice are more likely to be drawn from the bench
  - Otherwise no two punters would never use the same die
- There are several (equivalent) ways to define the DP:
  - The stick-breaking construction
    - This is the "infinite-component mixture" view
  - The Chinese Restaurant process
    - This emphasizes the partitions, rather than the components
  - The formal definition
    - This is where Dirichlet distributions come in
- A common application is to use MCMC to sample partitions

Parameters of the Dirichlet process

The DP has two parameters:

- A concentration parameter, α, that drives the number of clusters, via the Chinese Restaurant Process (CRP)
  - Not fixed like K, but analogous
  - High α means lots of clusters
  - Canonical value is $\alpha = 1$
- A base distribution, $G_0(\theta)$, that is auxiliary to the CRP
  - This represents the prior on parameters for an individual component (cluster)
  - Typical choice: conjugate to the component likelihood
    - So, in our casino example, it is a Dirichlet distribution (conjugate to multinomial)
    - For the DP version of K-means, it would be the Normal-Gamma (conjugate to Gaussian)

Chinese Restaurant Process

The usual formulation:

- Chinese restaurant with infinite tables, each infinitely large
- Customer #1 enters and sits at a table
- Customer #2 sits at the same table with probability $1/(1+\alpha)$, or a new table with $P = \alpha/(1+\alpha)$
- This continues: customer #$(n+1)$ sits at a k-person table with $P = k/(n+\alpha)$ and a new table with $P = \alpha/(n+\alpha)$

With n customers in the restaurant, let b denote the set of customers at a particular table and $B_n = \{b\}$ the set of all occupied tables. (B_n is a partition of the n customers.) Then

$$P(B_n = B|\alpha) = \frac{\Gamma(\alpha)\alpha^{|B|}}{\Gamma(\alpha+n)} \prod_{b\in B} \Gamma(|b|)$$

Can think of α as the number of "imaginary friends"...

Derivation of the partition probability

Let B_{n-1} be the partition when customer n enters the restaurant.
Let b_n be the random table joined by customer n.

$$P(b_n = b | b_1, \ldots, b_{n-1}) = \begin{cases} \frac{\alpha}{\alpha+n-1} & \text{if } b_n \notin B_{n-1} \text{ (new table)} \\ \frac{|b|}{\alpha+n-1} & \text{if } b_n \in B_{n-1} \text{ (existing table)} \end{cases}$$

This leads to the stated form of $P(B_n)$

$$P(B_n = B) = \frac{\alpha^{|B|}}{(\alpha+n-1) \times (\alpha+n-2) \times \ldots \times (\alpha+1) \times \alpha} \prod_{b \in B} (|b|-1)!$$

The CRP and the Dirichlet Process

Now think of dice-making machines:

- Restaurant is also a casino, owns a dice-making machine
- There is a die on every table (all from the same machine)
- Each punter chooses a table sociably (just like the CRP) and then rolls the die at that table
- Thus, punter #2 re-uses first die with $P = 1/(1+\alpha)$, or a new die with $P = \alpha/(1+\alpha)$
- This continues: the $(n+1)$'th punter uses a new die with probability $P = \alpha/(n+\alpha)$; otherwise, they pick one of the previous punters at random and re-use that punter's die

Thus, we use the CRP (with parameter α) to define a prior on the partitioning of the data to components, and G_0 (which here is a Dirichlet distribution) to sample the parameter used by each component.

The partition-conditioned likelihood

Given a particular partition, B, we can write down the likelihood. Let the data be $Y = \{Y_1 \ldots Y_n\}$ and let $Y^{(b)} = \{Y_n : n \in b\}$ be the subset associated with a particular block of the partition.

$$P(Y|B, G_0) = \prod_{b \in B} \int G_0(\theta) P(Y^{(b)}|\theta) d\theta$$

This is simplified if G_0 is conjugate to the likelihood $P(Y|\theta)$.

Properties of the CRP

With n customers in the restaurant, let b denote the set of customers at a particular table and $B_n = \{b\}$ the set of all occupied tables.

$$P(B_n = B|\alpha) = \frac{\Gamma(\alpha)\alpha^{|B|}}{\Gamma(\alpha + n)} \prod_{b \in B} \Gamma(|b|)$$

In the special case of $\alpha = 1$

$$P(B_n = B|\alpha = 1) = \frac{\prod_{b \in B}(|b| - 1)!}{n!}$$

Note that $P(B_n)$ is exchangeable, i.e. invariant to permutations of the customers. This is a useful property because it means we can remove any customer, and reintroduce them as if they were the last one to enter the restaurant. This is effectively how we do Gibbs-sampling under the DP.

Gibbs sampling reviewed

Recall the core idea of Gibbs sampling:

- We want to sample from a multivariate probability distribution $P(\mathbf{x}) = P(x_1, x_2, x_3, \ldots x_N)$
- We seek a Markov chain with transition probability $Q(\mathbf{x}'|\mathbf{x})$ and equilibrium $P(\mathbf{x})$, satisfying detailed balance

$$P(\mathbf{x})Q(\mathbf{x}'|\mathbf{x}) = P(\mathbf{x}')Q(\mathbf{x}|\mathbf{x}')$$

- We achieve this by sampling exactly from the marginal distribution of one of the N variables, given all the others

$$Q_k(\mathbf{x}'|\mathbf{x}) = P(x'_k|\{x_n : n \neq k\}) \prod_{n \neq k} \delta(x'_n = x_n)$$

$$Q(\mathbf{x}'|\mathbf{x}) = \sum_{k=1}^{N} \frac{1}{N} Q_k(\mathbf{x}'|\mathbf{x})$$

Complete Gibbs sampler for the DP

Joint likelihood of data and partition

$$P(Y, B|\alpha, G_0) = P(B|\alpha)P(Y|B, G_0)$$

Gibbs-sampling algorithm:

- Remove a random datapoint from its cluster (block, table)
- Calculate the likelihood for placing the datapoint in any existing cluster, as well as in its own new cluster
- Randomly sample the datapoint's new cluster from these likelihoods

Stick-breaking construction of the CRP

- Coin-making machine produces coins $\sim$ Beta$(1, \alpha)$
- There is a coin on each table
- The tables are numbered $(1, 2, 3 \ldots)$
- Visit the tables in order, flipping coins until you get a "Heads"
- Sit at that table

Notes:

- This gives the same marginal distribution of partitions as the CRP, if you integrate out the coin probabilities and sum out the table numbers

Why "stick-breaking"?

The more usual formulation is that instead of a series of coins with probabilities $p_1, p_2, p_3 \ldots$ we break a unit-length stick into successively shorter pieces, with the n'th piece having length

$$\ell_n = p_n \prod_{k=1}^{n-1} (1 - p_k)$$

The probability of sitting at the n'th table is ℓ_n.

Stick-breaking construction of the DP

- Dice-making machine produces dice $\sim G_0$
- Coin-making machine produces coins $\sim \text{Beta}(1, \alpha)$
- There is a coin *and* a die on each table
- The tables are numbered $(1, 2, 3 \ldots)$
- Visit the tables in order, flipping coins until you get a "Heads"
- Pick the die on that table

Notes:

- The resulting distribution over dice (θ) shares many properties with the underlying machine (G_0), but unlike G_0, it is a discrete distribution (G_0 is continuous)
- Equivalently: there is a finite probability of getting the exact same die twice (whereas the machine will **never** produce two identical dice)

Formal definition of the DP

For completeness...

Let G_0 be a probability measure over some measurable space $(\Theta, \mathcal{B})$ and let α be a positive real number.

A *Dirichlet Process*, $\mathrm{DP}(\alpha, G_0)$, is the distribution of a random measure G over $(\Theta, \mathcal{B})$ such that, for any partition $(A_1, A_2, \ldots, A_r)$ of Θ, the finite-dimensional distribution of that partition according to G is Dirichlet-distributed:

$$(G(A_1), G(A_2) \ldots G(A_r)) \sim \mathrm{Dirichlet}(\alpha G_0(A_1), \alpha G_0(A_2) \ldots \alpha G_0(A_r))$$

CRP and DP in biology

- Pólya Urn formulations of the CRP
- Ewens' Sampling Formula
- Machine learning (clustering, regression, . . .)

Pólya Urn formulation

The CRP itself (Hoppe, J of Math. Biol. 1984)

- Start with an urn containing
  - one black ball of mass α
  - n_k balls of color k and unit mass
  - Initially all the n_k are zero (just one black ball)
- Draw a ball from the urn, proportionally to its mass
  - If it's black, replace it in the urn, together with a unit-mass ball of a new color
  - If it's not black, replace it in the urn, together with a unit-mass ball of the same color
- Note that the black ball is *"ignored in describing the urn configuration since it is always present and merely a device for generating new labels"* (Hoppe)

Pólya Urn formulation

The Gibbs sampler for the CRP

- The urn contains
  - one black ball of mass α
  - n_k balls of color k and unit mass
- Remove a (colored) ball from the urn and discard it
- Draw a ball from the urn, proportionally to its mass
  - If it's black, replace it in the urn, together with a unit-mass ball of a new color
  - If it's not black, replace it in the urn, together with a unit-mass ball of the same color

Ewens's Sampling Formula

Here (again) is the partition probability for n customers

$$P(B_n = B|\alpha) = \frac{\Gamma(\alpha)\alpha^{|B|}}{\Gamma(\alpha+n)} \prod_{b\in B} \Gamma(|b|)$$

Let a_k be the number of tables having occupancy k. Then

$$P(a_1, \ldots, a_n) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha+n)} \prod_{k=1}^{n} \frac{\alpha^{a_k}}{k^{a_k}\Gamma(a_k+1)}$$

This is Ewens' Sampling Formula, a.k.a. the multivariate Ewens distribution.

Derivation of Ewens's Sampling Formula

From Hoppe (J.Math.Biol., 1984)

- Arrange the K tables in decreasing order of occupancy and let n_k be the occupancy of table k
- Let a_i denote the number of tables with occupancy i; let $A = \{i : a_i > 0\}$
- There are $K!/\prod_{i \in A} a_i!$ ways of distributing the counts $(n_1, n_2, \ldots, n_K)$ amongst the K tables, and for each such way there are $n!/\prod_{k=1}^{K} n_k!$ equivalent permutations of the customers; multiply these to get

$$\mathcal{C}(\mathbf{a}) = \frac{K! n!}{\prod_{i \in A} a_i! \prod_{k=1}^{K} n_k!}$$

Derivation of Ewens's Sampling Formula

- The partition probability is

$$P(B) = \frac{\Gamma(\alpha)\alpha^{|B|}}{\Gamma(\alpha+n)} \prod_{b \in B} \Gamma(|b|)$$

- The number of table-customer permutations is

$$\mathcal{C}(\mathbf{a}) = \frac{K!n!}{\prod_{i \in A} a_i! \prod_{k=1}^{K} n_k!}$$

- Only a fraction $1/K!$ of these permutations will have the K tables appearing in the correct order
- Multiplying these together gives Ewens' distribution

$$P(\mathbf{a}) = \frac{P(B)\mathcal{C}(\mathbf{a})}{K!} = \frac{\Gamma(\alpha)n!\alpha^{|B|}}{\Gamma(\alpha+n)} \prod_{i \in A} \frac{a_i!}{i^{a_i}}$$

Ewens's Sampling Formula in ecology

Unified Neutral Theory of Biodiversity

- An ecological niche has room only for a fixed number of individuals, n
- When an individual reproduces, it displaces another individual
- With probability α/n, the progeny mutates to create a new species

Ewens's Sampling Formula in ecology

Urn-based Neutral Theory of Ball-diversity

- An urn has room only for a fixed number of balls, n
- When an individual ball reproduces, it displaces another
- With probability α/n, the progeny mutates to create a new color

This is precisely our Gibbs-sampling scheme for the CRP—therefore the equilibrium distribution over partitions is that of the CRP, and Ewens' Sampling Formula gives the frequency distribution of species.

Ewens's Sampling Formula in population genetics

Pretty similar to the ecological version

- A population has fixed size, n (Wright-Fisher model)
- When an individual reproduces, it displaces another individual (Moran model)
- With probability α/n, the progeny mutates to create a new allele (Infinite-Alleles model)

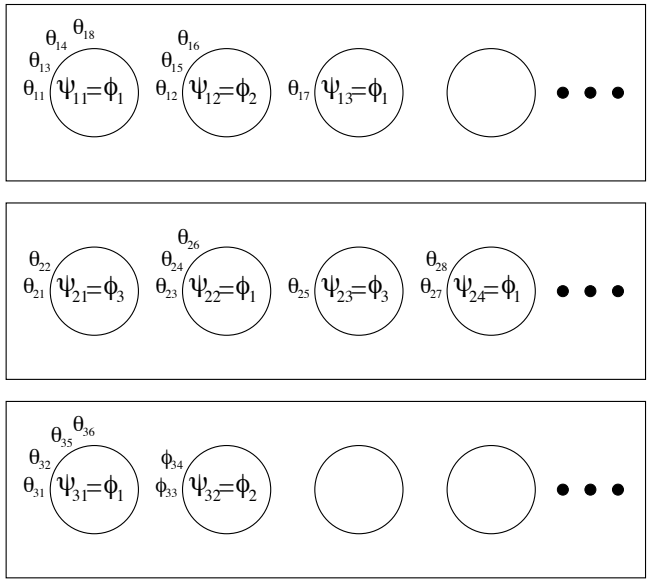
Ewens' Sampling Formula gives the frequency distribution of alleles. It can also be derived using coalescent theory (understanding Ewens' result was, reputedly, Kingman's motivation for developing the coalescent).

The hierarchical Dirichlet process

- Extension of the DP due to Jordan *et al*
- Suitable for clustering tasks where the data are *already* subdivided into groups
- You want to cluster within each group, but you also want to re-use clusters across groups
- Solution: allow the base distribution G_0 to itself be a draw from a DP
- Analogy: the Chinese Restaurant Franchise

The Chinese Restaurant Franchise

From Teh, Jordan, Beal & Blei (2005)



- Within each restaurant, you have the usual CRP
- At each table, a single dish is served
- Dishes are drawn from a global menu, via a *separate* CRP
- Tables in different restaurants can share the same dish

Summary

- Chinese Restaurant Process
  - Induces a prior over partitions of customers to tables
- Dirichlet Process
  - Associates a mixture component with each table
  - Numerous extensions, e.g. hierarchical DP
- Ewens' Sampling Formula
  - Distribution of table occupancies
  - Occurs in population genetics & ecology