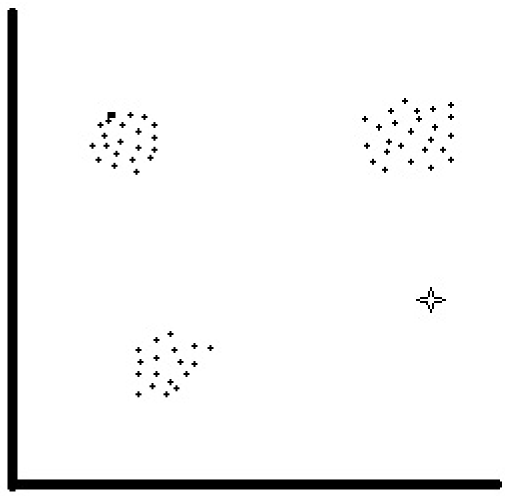
Baysesian NonParametrics

Scribe: Alex Tank

January 13, 2011

1 Introduction

Two Perspectives:
Machine Learning: Priors on structures that can accommodate infinite sizes/infinite cardinality. When combined with data the posterior using these methods gives a distribution that can grow with new observations.



Nonparametric methods do not need to know there are three clusters in advance. If a new data point is observed that has low probability under current cluster parameters the methods can assign a new cluster to that data point.

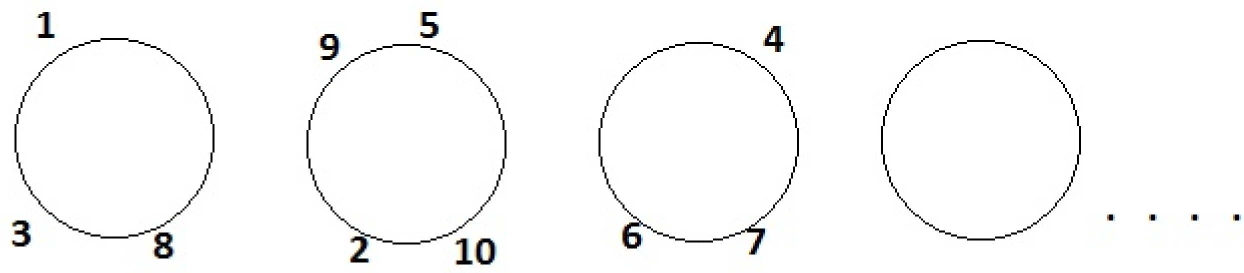
Statistics: Nonparametric methods contain priors on densities or distributions over an arbitrary space. When these methods are used in a model the posterior distribution is over densities.

2 The Chinese Restaurant Process

- Distribution over space of partitions of all integers - Any partition implies a sub partition Eg. (1 3 8)(2 5 9 10)(4 6 7)

2.1 CRP idea

Imagine a Chinese Restaurant with an infinite number of tables....



Customers walk in sequentially and sit down.

P(next unoccupied table | current seating plan) =

$$\frac{\alpha}{n-1+\alpha}$$

P(a previously occupied table| current seating plan) =

$$\frac{c_i}{n-1+\alpha}$$

Where c_i is the number of people sitting at table i.
1st customer sits at T_i with p = 1.

Using the seating locations for customers 1...10 above where $z_1 ... z_{10}$ are the table assignments for each customer, the probability of that seating assignment is:

$$p(z_1...z_{10}) = p(z_1)p(z_2|z_1)...p(z_{10}|z_{1:9}) = (\frac{\alpha}{\alpha})(\frac{\alpha}{1+\alpha})(\frac{1}{2+\alpha})(\frac{\alpha}{3+\alpha})(\frac{1}{4+\alpha})(\frac{1}{5+\alpha})(\frac{2}{6+\alpha})...$$

Notice that if the ordering of customers is switched, the probability of the partition is the same. If we instead use the ordering [5, 2, 7, 1, 3, 6, 4]:

$$p(z'_1....z'_7) = (\frac{\alpha}{\alpha})(\frac{1}{1+\alpha})(\frac{\alpha}{2+\alpha})(\frac{\alpha}{3+\alpha})(\frac{1}{4+\alpha})(\frac{1}{5+\alpha})(\frac{2}{6+\alpha})$$

Notice that the two orderings share the same collection of numbers in numerators and same collection of numbers in denominators

CRP properties

1) The CRP is exchangeable
2) The partition can always expand with the next customer
3) The number of partitions is random(controlled by α).

2.2 Using CRPs

Generative process for finite mixture of Gaussians:

1)Choose $\pi \sim$ Dirichlet(α)
2) Choose K means: $\mu_k \sim N(0, \sigma^2)$
3) for each data point a) choose $z_n \sim$ Mult(π) b) choose $x_n \sim\ N(\mu_z, \lambda^2)$

MCMC lets us reverse the process.

Generative Process for CRP:

1)Choose ∞ means: $\mu_k \sim$ N$(0, \sigma_2)(G_o)$ for k = 1,2,3...
2)For each data point: a) $z_n|z_{1:n-1} \sim$ CRP(α) b) $x_n \sim$N(μ_{z_n}, λ^2)

The expected number of clusters for N samples is: $\mathbf{E}[k_N] = \alpha$log N

The reverse, $P(Z_{1:N}, \mu_{1:\infty}|x_{1:N})$, cannot be computed directly. Instead we can use Gibbs sampling.

In Gibbs sampling, we can compute $P(Z_i|z_{-i}, x_{1:N})$ in a finite mixture by integrating out the μs.

For CRP: Pretend z_i is the last customer(because of exchangeability). Now look at posterior distribution of different μs, then sample μ_k from current assignments.

$$P(z_i|z_{-i}, x_{i:N}) = \sum_{k=1}^{k_n^{-i}} p(z_k)p(x_i|z_i = k, x_{1:N}, z_{-i}) + p(newtable)p(x_i)$$

For more information: Radford Neal (2000)

2.3 Dirichlet Process

Ferguson(1973)

The Dirichlet Process defines a distribution over distributions. It can be thought of as an infinite-dimensional Dirichlet distribution.
Choose random distribution from reals:

$$G \sim DP(\alpha, G_o)$$

where α is a scaling parameters and G_o is the base distribution.
Choose random parameters from reals:

$$\theta_n \sim G$$

Now use Variational Inference Algorithm. For more information see Erik Sudderth background chapter.

3 An Overview of the Field

$*$ indicates topics not covered in class

1. Graphical Models
   (a) Directed
   (b) Undirected
   (c) Factor Graphs $*$
   (d) Factor Graphs $*$
   (e) Independence and Bayes Ball
2. Exact Inference
   (a) Elimination
   (b) Propagation on trees
   (c) Junction Tree $*$
3. Predictive, Fully Observable Models
   (a) Linear Regression
   (b) Regularized Linear Regression
   (c) Exponential Family
   (d) GLM
   (e) Additive Models $*$
4. Latent Variable Models
   (a) Mixture models
   (b) PCA/factor analysis
   (c) HMM

   (d) Mixed - membership models - model of group data $*$
   (e) Hierarchical Modeling - *the future*

5. Approximate Inference
   (a) MCMC - gibbs samplng
   (b) Mean-field Variational Inference

   (a) Causality
   (b) Applications
   (c) Model Checking