Non-parametric Bayesian Models

Jie Tang

Department of Computer Science and Technology

Tsinghua University

Contents

Contents

Introduction

We'll talk about *nonparametric Bayesian methods* in following slides

- Regression, classification – Gaussian Process
- Clustering – Dirichlet Process

What is *nonparametric* ?

What is *Bayesian* ?

What is Nonparametric Model?

Nonparametric doesn't mean there are no parameters.

Parametric model assumes a finite set of parameters.

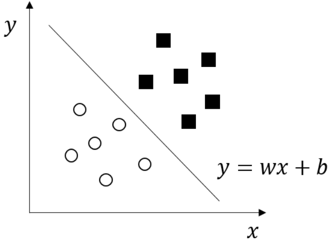
- Finite model parameters captures everything about the data
- Predictions for future data are made merely through parameters, independent from observed data.
- Model complexity are bounded, even with infinite data

Nonparametric model assumes the model can't be specified by a finite set of parameters. But often a infinite-dimensional parameter space will work.

- Model complexity can grow with data

Example: Classification

Consider a linear model

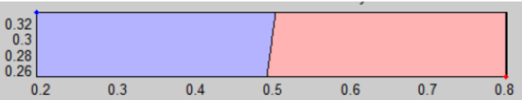


No matter how many samples are trained, number of parameters are fixed (w, b)

Example: Classification

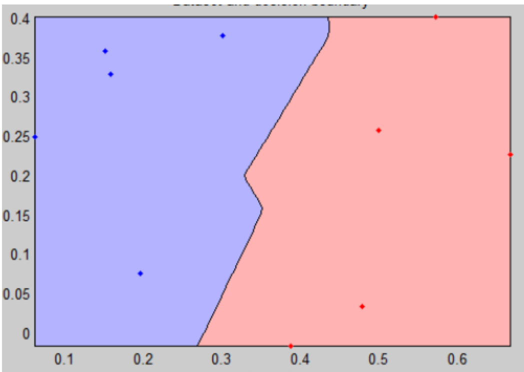
How about k-NN?

2 samples



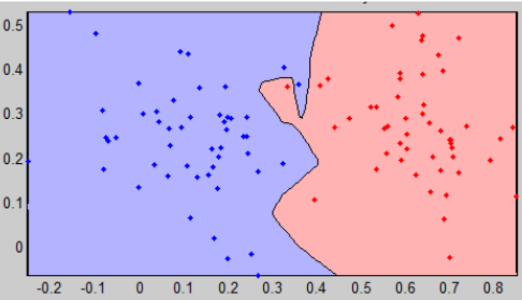
Example: Classification

10 samples



Example: Classification

100 samples



Parameters are unbounded as dataset grow larger

Nonparametric!

What about SVM?

Linear SVM

$$y = \text{sign}(w^T x)$$

What about SVM?

Linear SVM

$$y = \text{sign}(w^T x)$$

Kernel SVM

$$y = \text{sign}(\sum_i \alpha_i y_i \kappa(x_i, x)))$$

Parametric vs. Nonparametric

Parametric Models:

- Inflexible. Simple and easy to interpret.
- Performs well and fast, if model captures the data correctly
- Otherwise will give inconsistent result

Nonparametric Models:

- Flexible.
- Complex and hard to interpret, thus giving inaccurate estimation

Question: which of the following algorithm belongs to parametric models?

(A) k-NN
(B) SVM
(C) Logistic Regression
(D) Decision Tree

What is Bayesian?

Consider a regression problem

Given training data D, we aim to find a model f. For each sample x, return a prediction y

How to give an optimal y?

What is Bayesian?

Consider a regression problem

Given training data D, we aim to find a model f. For each sample x, return a prediction y

How to give an optimal y?

- MLE
- MAP
- Bayesian

Maximum Likelihood Estimation

In training, we maximize a likelihood

$$f^* = \arg\max_f L(f|D) = \arg\max_f p(D|f)$$

In prediction, just feed f with x

$$p(y|x, D) = p(y|f^*, x)$$

Maximum A Posteriori

To avoid overfitting, we place a prior on the model

$$f^* = \arg\max_f \frac{p(D|f)p(f)}{p(D)}$$

Prediction process remains the same

$$p(y|x, D) = p(y|f^*, x)$$

Bayesian Model

In training, we just estimate the posterior distribution of f (same as MAP)

$$p(f|D) = \frac{p(D|f)p(f)}{p(D)}$$

In prediction, we no longer want a point estimation. Instead, we consider all possible models and take the expectation

$$p(y|x, D) = \int_f p(y|f, x)p(f|D)\,\mathrm{d}f$$

Why Bayesian?

Infinite Exchangeability:

$$\forall n, \forall \sigma, p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

De Finetti's Theorem: If $(x_1, x_2, \ldots)$ are infinite exchangeable, then there exists a random variable f, $\forall n$,

$$p(x_1, \ldots, x_n) = \int_f \prod_{i=1}^{n} p(x_i|f)p(f)\,\mathrm{d}f$$

Why Bayesian?

Infinite Exchangeability:

$$\forall n, \forall \sigma, p(x_1, \dots, x_n) = p(x_{\sigma(1)}, \dots, x_{\sigma(n)})$$

De Finetti's Theorem: If $(x_1, x_2, \dots)$ are infinite exchangeable, then there exists a random variable f, $\forall n$,

$$p(x_1, \dots, x_n) = \int_f \prod_{i=1}^{n} p(x_i|f) p(f) \, \mathrm{d}f$$

How to define the prior $p(f)$?

Parametric Approach

In previous courses, we focused on parametric representation of f, e.g.

$$f(x) = w^T x$$

Instead of defining prior on function itself, we define prior on the parameters

$$p(w) = \mathcal{N}(0, \sigma^2 I)$$

Now, we define the prior on the function f directly

Contents

Warmup: Multivariate Gaussian

Before introducing Gaussian Process, let's recall multivariate Gaussian distribution first.

We say a n-dimensional random variable $\mathbf{x}$ follows multivariate Gaussian

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

if

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\}$$

where

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu} \\ \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] &= \Sigma_{ij} \end{aligned}$$

Marginal Distribution

Suppose

$$\begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix} \right)$$

Integrating $\boldsymbol{x_2}$ out,

$$\boldsymbol{x_1} \sim \mathcal{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_{11}})$$

Just drop irrelevant components

Conditional Distribution

Suppose

$$\begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix} \right)$$

Then

$$\begin{aligned} p(\mathbf{x_1}|\mathbf{x_2}) &= \mathcal{N}(\mathbf{x_1}|\boldsymbol{\mu_{1|2}}, \boldsymbol{\Sigma_{1|2}}) \\ \boldsymbol{\mu_{1|2}} &= \boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}(\mathbf{x_2} - \boldsymbol{\mu_2}) \\ \boldsymbol{\Sigma_{1|2}} &= \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}} \end{aligned}$$

We will use it later

How to get that?

Some Details

First, we diagonalize the variance matrix to ease the inverse matrix calculation

$$\begin{bmatrix} \Sigma_{11}/\Sigma_{22} & O \\ O & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ O & I_2 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_1 & O \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_2 \end{bmatrix}$$

Define $\Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ here. Then

$$\begin{aligned} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = & \begin{bmatrix} I_1 & O \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_2 \end{bmatrix} \\ & \begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & O \\ O & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I_1 & -\Sigma_{12}\Sigma_{22}^{-1} \\ O & I_2 \end{bmatrix} \end{aligned}$$

More Details

Check the exponent of $p(\boldsymbol{x_1}, \boldsymbol{x_2})$:

$$
\begin{aligned}
&(x_1-\mu_1, x_2-\mu_2)\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}^{-1}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}\\
=&\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}^T\begin{bmatrix}I_1 & O\\ -\Sigma_{22}^{-1}\Sigma_{21} & I_2\end{bmatrix}\begin{bmatrix}(\Sigma/\Sigma_{22})^{-1} & O\\ O & \Sigma_{22}^{-1}\end{bmatrix}\\
&\begin{bmatrix}I_1 & -\Sigma_{12}\Sigma_{22}^{-1}\\ O & I_2\end{bmatrix}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}\\
=&(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}\\
&(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))+C
\end{aligned}
$$

Note that $p(\boldsymbol{x_1}|\boldsymbol{x_2}) = \frac{p(\boldsymbol{x_1},\boldsymbol{x_2})}{p(\boldsymbol{x_2})}$, and C is just the exponent of $p(\boldsymbol{x_2})$. From the equation we can get the mean and variance of the conditional distribution.

Contents

Gaussian Process

Gaussian Process is a generalization of multivariate Gaussian to infinite variables – *stochastic process*

$\forall \mathbf{x_1}, \ldots, \mathbf{x_n}, (f(\mathbf{x_1}), \ldots, f(\mathbf{x_n}))$ is jointly Gaussian

A Gaussian Process is fully specified by mean function $m(x)$ and covariance function $\kappa(\mathbf{x}, \mathbf{x}')$

$$
\begin{aligned}
m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\
\kappa(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]
\end{aligned}
$$

Then,

$$(f(\mathbf{x_1}), f(\mathbf{x_2}), \ldots, f(\mathbf{x_n}))^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$$

where

$$\mu_i = m(\mathbf{x_i}), K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$$

We require κ to be a positive definite kernel.

Contents

GP Regression with noise-free observations

For simplicity, let's start with predictions using noise free observations

Training set $D = \{(x_i, f_i)\}$, where $f_i = f(x_i)$

Noise free! – No uncertainty.

Test set $\mathbf{X}_*$. We want to predict outputs f_*

From definition of GP,

$$\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{\boldsymbol{T}} & \boldsymbol{K}_{**} \end{bmatrix} \right)$$

GP Regression with noise-free observations

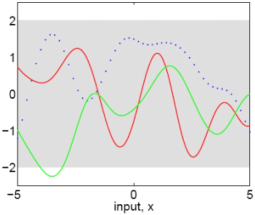
Using results of marginal multivariate Gaussian,

$$
\begin{aligned}
p(\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f}) = \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)& \\
\boldsymbol{\mu}_* &= \boldsymbol{\mu}(\boldsymbol{X}_*) + \boldsymbol{K}_*^T \boldsymbol{K}^{-1} (\boldsymbol{f} - \boldsymbol{\mu}(\boldsymbol{X})) \\
\boldsymbol{\Sigma}_* &= \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}^{-1} \boldsymbol{K}_*
\end{aligned}
$$

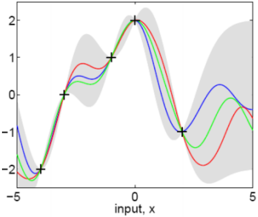
We usually set $\mu(\boldsymbol{x}) = 0$, because GP can model the mean arbitrarily well.

Visualization

Let's see some samples from GP prior and posterior.



(a) prior



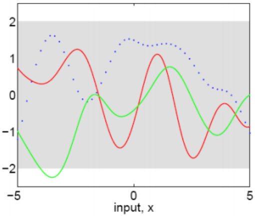
(b) posterior

Training data points are marked with '$+$'.

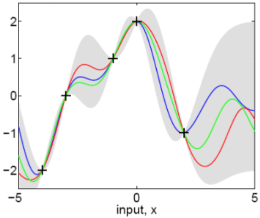
No uncertainty at training data. Why?

Visualization

Let's see some samples from GP prior and posterior.



(c) prior



(d) posterior

Training data points are marked with '$+$'.

No uncertainty at training data. Why?

$\boldsymbol{\Sigma}_* = \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}^{-1} \boldsymbol{K}_* = 0$ when $\boldsymbol{K}_*$ and $\boldsymbol{K}_{**}$ are sub-matrices of $\boldsymbol{K}$ – Noise-free

GP Regression with noisy observations

When observations are noisy, the true f is hidden.

Instead, we observe f plus a Gaussian noise

$$y = f(\boldsymbol{x}) + \epsilon$$

where

$$\epsilon \sim \mathcal{N}(0, \sigma_y^2)$$

Now

$$\mathrm{cov}[y_p, y_q] = \kappa(\boldsymbol{x_p}, \boldsymbol{x_q}) + \sigma_y^2 \delta_{pq}$$

in other words

$$\mathrm{cov}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{K} + \sigma_y^2 \boldsymbol{I} \triangleq \boldsymbol{K}_y$$

GP Regression with noisy observations

Take $m(\boldsymbol{x}) = 0$ here

Similar to noise-free result

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{K_y} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^{\boldsymbol{T}} & \boldsymbol{K}_{**} \end{bmatrix}\right)$$

$$\begin{aligned} p(\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{y}) &= \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \boldsymbol{K}_*^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y} \\ \boldsymbol{\Sigma}_* &= \boldsymbol{K}_{**} - \boldsymbol{K}_*^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{K}_* \end{aligned}$$

Deeper Insight

Feed the result with a single sample $\boldsymbol{x_*}$

$$p(f_*|\boldsymbol{x_*}, \boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{k}_*^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y}, k_{**} - \boldsymbol{k}_*^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{k}_*)$$

$$\bar{f}_* = \sum_i \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x_*}) \qquad \boldsymbol{\alpha} = \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y}$$

What can we find?

Deeper Insight

Feed the result with a single sample $\boldsymbol{x_*}$

$$p(f_*|\boldsymbol{x_*}, \boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{k}_*^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y}, k_{**} - \boldsymbol{k}_*^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{k}_*)$$

$$\bar{f}_* = \sum_i \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x_*}) \qquad \boldsymbol{\alpha} = \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y}$$

What can we find?

Although we define GP on infinite dimension, it's suffices to consider only $n+1$ dimension.

In terms of prediction mean, GP regression can be considered as linear predictor when features are projected to high dimension space using feature map $\phi(\cdot)$ implied by κ. (Recall $\kappa(x_i, x_j) = \phi(x_i)^T \phi(x_j)$)

Effect of Kernel Parameters

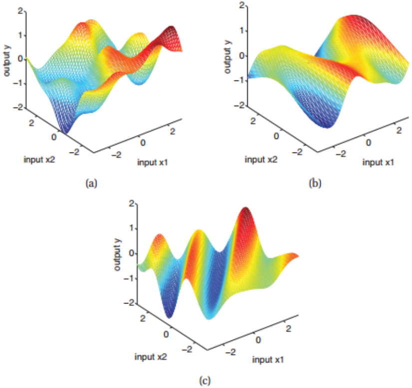
Kernel parameters are crucial in GP

RBF kernel: $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)\} + \sigma_y^2 \delta_{ij}$

$M = I$

$M = \text{diag}(1, 3)^{-2}$

$M = (1, -1; -1, 1) + \text{diag}(6, 6)^{-2}$

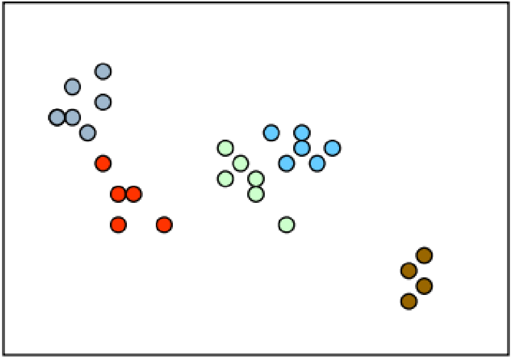


Contents

Example: Clustering

We're give a data set which are generated by a mixture of Gaussian.

How to model the data and cluster them?

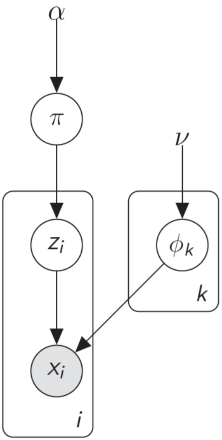


Gaussian Mixture Clustering

Suppose the dataset is generated by K Gaussian components

Generating process:

$$\begin{aligned}\phi_k = (\mu_k, \Sigma_k) &\sim \mathcal{NIW}(\nu)\\ \pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)\\ z_i &\sim \text{Multinomial}(\pi)\\ x_i &\sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})\end{aligned}$$

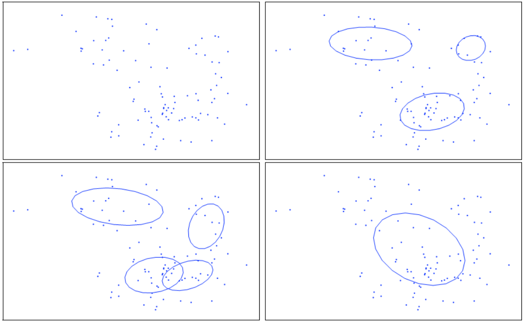


where

$$\begin{aligned}\mathcal{NIW}(\nu) &: \text{conjugate prior of Gaussian}\\ \textit{Dirichlet}(\ldots) &: \text{conjugate prior of multinomial}\\ z_i &: \text{mixture indicator of } x_i\end{aligned}$$

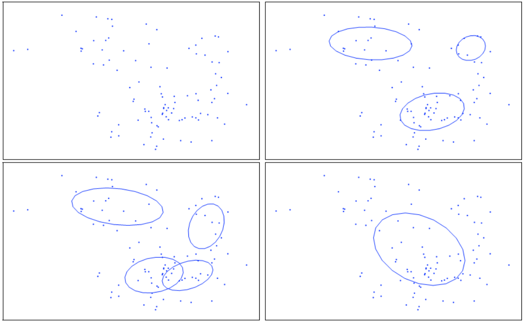
How to choose K?

How many clusters?



How to choose K?

How many clusters?



What if let $K \to \infty$?

Infinity Mixture Models

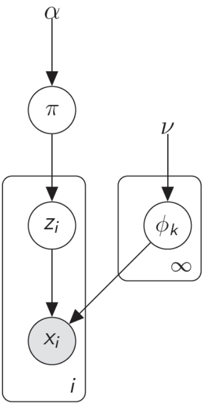
Imagine $K \gg 0$,

In Bayesian inference, we integrate ϕ_k and π out. Number of latent variables would not grow with K – No overfitting

At most *n* active components are associated with data

This is an *infinite mixture model*

Issue: can we take $K \to \infty$? What's the limiting model?



An Alternative View of Mixture Model

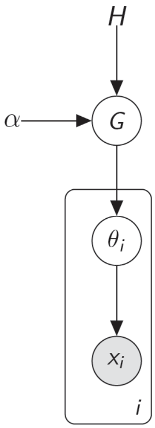
Mixture parameter for each data point can be sampled directly

z_i is absorbed into a *discrete probability measure* G

δ_{ϕ_k} is an *atom* positioned at ϕ_k

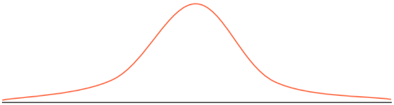
θ_i is the sampled Gaussian parameter to generate x_i

$$\phi_k \sim H$$
$$\pi_k \sim Dirichlet(\alpha/K, \ldots, \alpha/K)$$
$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$
$$\theta_i \sim G$$
$$x_i \sim p(x_i|\theta_i)$$

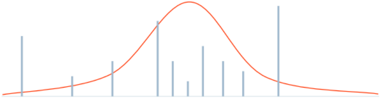


An Alternative View of Mixture Model

H: continuous



G: discrete, with mass at ϕ_k



Recall

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

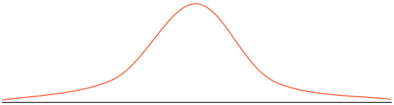
Take $K \to \infty$,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

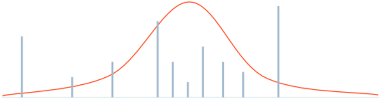
To execute Bayesian inference, we need to find a prior for G

An Alternative View of Mixture Model

H: continuous



G: discrete, with mass at ϕ_k



Recall

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

Take $K \to \infty$,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

To execute Bayesian inference, we need to find a prior for G

Dirichlet Process

Warmup: Dirichlet Distribution

Dirichlet distribution is a distribution over the K-dim probability simplex

$$\Omega_K = \{(\pi_1, \ldots, \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\}$$

We say $(\pi_1, \ldots, \pi_K)$ is Dirichlet distributed with parameters $(\alpha_1, \ldots, \alpha_K)$, if

$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

Warmup: Properties of Dirichlet Distribution

Aggregation Property:

If

$$(\pi_1, \ldots, \pi_i, \pi_{i+1}, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1}, \ldots, \pi_k)$$

then

$$(\pi_1, \ldots, \pi_i + \pi_{i+1}, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i + \alpha_{i+1}, \ldots, \pi_k)$$

Inverse of Aggregation Property:

If

$$\begin{aligned}(\pi_1, \ldots, \pi_k) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ t &\sim \text{Beta}(\alpha_i \beta, \alpha_i(1-\beta)) \\ (aka(t, 1-t) &\sim \text{Dirichlet}(\beta\alpha_i, (1-\beta)\alpha_i))\end{aligned}$$

then

$$\begin{aligned}&(\pi_1, \ldots, t\pi_i, (1-t)\pi_{i+1}, \ldots, \pi_K) \\ &\quad \sim \text{Dirichlet}(\alpha_1, \ldots, \beta\alpha_i, (1-\beta)\alpha_{i+1}, \ldots, \alpha_K)\end{aligned}$$

Warmup: Dirichlet-Multinomial conjugacy

Dirichlet-Multinomial conjugacy:

If

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\
z &\sim \textit{Multinomial}(\pi_1, \ldots, \pi_K)
\end{aligned}
$$

then

$$(\pi_1, \ldots, \pi_k)|z \sim \text{Dirichlet}(\alpha_1 + \delta_z(1), \ldots, \pi_K + \delta_z(K))$$

Contents

Dirichlet Process

Starting from π, slice the space finer and finer, what do we get?

$$(\pi) \sim \mathit{Dirichlet}(\alpha)$$
$$(\pi_1, \pi_2) \sim \mathit{Dirichlet}(\alpha_1, \alpha_2)$$
$$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) \sim \mathit{Dirichlet}(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}) \dots$$

Because the dividing schema is arbitrary, finally, any finite partition of π should follow Dirichlet distribution.

Dirichlet Process

Starting from π, slice the space finer and finer, what do we get?

$$(\pi) \sim \mathit{Dirichlet}(\alpha)$$
$$(\pi_1, \pi_2) \sim \mathit{Dirichlet}(\alpha_1, \alpha_2)$$
$$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) \sim \mathit{Dirichlet}(\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}) \ldots$$

Because the dividing schema is arbitrary, finally, any finite partition of π should follow Dirichlet distribution.

Dirichlet Process is a generalization of Dirichlet distribution to infinite many components – Recall Gaussian Process and multivariate Gaussian distribution.

Dirichlet Process

Let G be a probability measure over Ω. For any measurable subset $A \subset \Omega$, $G(A) = p(x \in A), x \in \Omega$

Dirichlet process (DP) is a distribution over G.

We write

$$G \sim \mathrm{DP}(\alpha, H)$$

if for any finite partition $(A_1, \ldots, A_n)$ of Ω,

$$(G(A_1), \ldots, G(A_n)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_n))$$

H is the base distribution

$$\mathrm{E}(G(A)) = H(A)$$

α is the concentration parameter

$$\mathrm{Var}(G(A)) = \frac{H(A)(1 - H(A))}{\alpha + 1}$$

Dirichlet Process

Our definition is all about properties of DP. How to construct a concrete sample from DP?

Why is DP useful for clustering?

We'll see later

Posterior Dirichlet Process

We'll need the posterior Dirichlet Process later

Suppose $G \sim DP(\alpha, H)$, $\theta \sim G$

What is $p(\theta)$ and $G|\theta$?

For any $\theta \in A \subset \Omega$,

$$p(A) = \int_G G(A)p(G)\,\mathrm{d}G = H(A)$$

So $\theta \sim H$

Posterior Dirichlet Process

Using Dirichlet-multinomial conjugacy, for any partition $(A_1, \ldots, A_n)$,

$$(G(A_1), \ldots, G(A_n))|\theta \\ \sim \text{Dirichlet}(\alpha H(A_1) + \delta_\theta(A_1), \ldots, \alpha H(A_n) + \delta_\theta(A_n))$$

where $\delta_\theta(A) = 1$ iff $\theta \in A$

So, the posterior is also a DP

$$G|\theta \sim DP\left(\alpha + 1, \frac{\alpha H + \delta_\theta}{\alpha + 1}\right)$$

Generalization for n observations:

$$G|\theta_1, \ldots, \theta_n \sim DP\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right)$$

We will construct a sample from DP based on the posterior

Contents

Stick Breaking

Define an infinite sequence of Beta random variables:

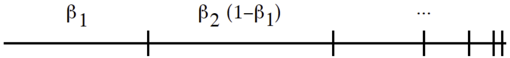
$$\beta_k \sim Beta(1, \alpha_0) \qquad k = 1, 2, \ldots$$

And then define an infinite sequence of mixing proportions as:

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \qquad k = 2, 3, \ldots$$

This can be viewed as breaking off portions of a stick:



Stick Breaking Construction with Dirichlet Process

Sample θ_1 from H

Consider partition $(\theta_1, \Omega - \{\theta_1\})$

$$(G(\theta_1), G(\Omega - \{\theta_1\}))|\theta_1 \sim \mathrm{Dir}(\alpha H(\theta_1) + 1, \alpha H(\Omega - \{\theta_1\}))$$

Noting that $G(\Omega - \{\theta_1\}) = 1 - G(A_1)$, $H(\theta) = 0$, $H(\Omega) = 1$, we get

$$G(\theta_1) \sim \mathrm{Beta}(1, \alpha)$$

Let $\beta_1 = G(\theta_1)$

$$G = \beta_1 \delta_{\theta_1} + (1 - \beta_1) G_1$$

where G_1 is a renormalized probability measure on $\Omega - \{\theta_1\}$

What is G_1?

Stick Breaking Construction with Dirichlet Process

Consider arbitrary partition $(\theta_1, A_1, \dots, A_n)$, we have

$$\begin{aligned}
&(G(\theta_1), G(A_1), \dots, G(A_n)) \\
&\quad = (\beta_1, (1-\beta_1)G_1(A_1), \dots, (1-\beta_1)G_1(A_n)) \\
&\quad \sim (1, H(A_1), \dots, H(A_n))
\end{aligned}$$

Using conditional distribution of Dirichlet distribution,

$$\frac{1}{1-\beta_1}((1-\beta_1)G_1(A_1), \dots, (1-\beta_1)G_1(A_n)) \sim (H(A_1), \dots, H(A_n))$$

So, we construct a new DP:

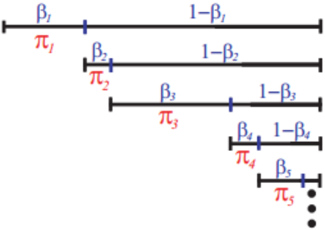
$$G_1 \sim DP(\alpha, H)$$

We can continue the breaking process likewise

Stick Breaking Construction with Dirichlet Process

Continue the process

$$G_2 = \beta_1 \delta_{\theta_1} + (1 - \beta_1) G_1$$
$$G_3 = \beta_1 \delta_{\theta_1} + (1 - \beta_1)(\beta_2 + (1 - \beta_2) G_2)$$
$$\vdots$$
$$G_k = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$



where

$$\beta_k \sim \text{Beta}(1, \alpha)$$
$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i)$$
$$\theta_k \sim H$$

Stick Breaking Construction

The above process is called the *stick breaking construction*

What can we find about DP from this construction?

- A sample G from DP is discrete with probability 1
- If we keep sampling from G, we'll get more and more repeated values
- That's reasonable for clustering (why?)

Contents

Blackwell-MacQueen Urn Scheme

But for clustering, we're actually interested in sampling mixture parameters $\theta_1, \theta_2, \ldots$ from Ω, integrating G out.

Blackwell-MacQueen Urn Scheme

But for clustering, we're actually interested in sampling mixture parameters $\theta_1, \theta_2, \ldots$ from Ω, integrating G out.

It seems easy. If $\theta_1, \theta_2, \ldots, \theta_n$ are independent given G, then we can simply sample each θ from H independently. Right?

Blackwell-MacQueen Urn Scheme

But for clustering, we're actually interested in sampling mixture parameters $\theta_1, \theta_2, \ldots$ from Ω, integrating G out.

It seems easy. If $\theta_1, \theta_2, \ldots, \theta_n$ are independent given G, then we can simply sample each θ from H independently. Right?

Unfortunately, the answer is No! θs are no longer independent if we integrate G out

$$p(\theta_1, \theta_2, \ldots, \theta_n) = \int_G \prod_{i=1}^{n} p(\theta_i|G)p(G)\,\mathrm{d}G$$

Blackwell-MacQueen Urn Scheme

Consider two samples θ_1, θ_2.

$$\begin{aligned} p(\theta_2|\theta_1) &= \int_G p(\theta_2, G|\theta_1)\,\mathrm{d}G \\ &= \int_G p(\theta_2|G)p(G|\theta_1)\,\mathrm{d}G \end{aligned}$$

We have already known the posterior distribution of G given θ_1,

$$G|\theta_1 \sim \mathrm{DP}(\alpha+1, \frac{\alpha H + \delta_{\theta_1}}{\alpha+1})$$

Then, given θ_1, the second sample θ_2 can be sampled from the posterior base distribution.

The sampling process becomes quite simple

Blackwell-MacQueen Urn Scheme

First sample:

$$\theta_1 \sim H, \quad G|\theta_1 \sim \mathrm{DP}(\alpha + 1, \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1})$$

Second sample:

$$\theta_2|\theta_1 \sim \frac{\alpha H + \delta_{\theta_1}}{\alpha + 1}, \quad G|\theta_1, \theta_2 \sim \mathrm{DP}(\alpha + 2, \frac{\alpha H + \delta_{\theta_1} + \delta_{\theta_2}}{\alpha + 2})$$

n^{th} sample:

$$\theta_n|\theta_{1,\ldots,n-1} \sim \frac{\alpha H + \sum_{k=1}^{n-1} \delta_{\theta_k}}{\alpha + n - 1}, \quad G|\theta_{1,\ldots,n} \sim \mathrm{DP}(\alpha + n, \frac{\alpha H + \sum_{k=1}^{n} \delta_{\theta_k}}{\alpha + n})$$

Blackwell-MacQueen Urn Scheme

One issue, how to sample from $\frac{\alpha H + \sum_{k=1}^{n-1} \delta_{\theta_k}}{\alpha + n - 1}$?

With probability $\propto \alpha$, sample from H

With probability $\propto n-1$, randomly pick one from $\theta_1, \ldots, \theta_{n-1}$

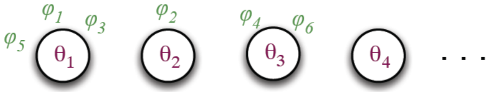
The above sample process is called *Blackwell-MacQueen Urn Scheme*

Samples $\theta_1, \ldots, \theta_n$ induces a partition over $1, \ldots, n$, if we cluster repeated values together

Working with float value is annoying. We can postpone sampling from H – Chinese Restaurant Process

Contents

Chinese Restaurant Process



Generating process of CRP:

- customer 1 sits at the table 1.
- the n^{th} customer
  - With probability $\frac{n_k}{\alpha+n-1}$, sits at existing table k, where n_k is number of customers at table k.
  - With probability $\frac{\alpha}{\alpha+n-1}$, sits at a new table $K+1$, where K is number of existing tables.

CRP teases apart clustering property of DP from base distribution H

To get back to Blackwell-MacQueen Urn Scheme, sample θ_k for each table k

CRP and DP

We can prove that the order of customer coming in doesn't affect partition probability

Infinite Exchangeability:

$$\forall n, \forall \sigma, p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

De Finetti's Theorem: If $(x_1, x_2, \ldots)$ are infinite exchangeable, then there exists a random variable θ, $\forall n$,

$$p(x_1, \ldots, x_n) = \int_\theta \prod_{i=1}^{n} p(x_i|\theta) p(\theta) \, \mathrm{d}\theta$$

Here, DP is the distribution of θ

CRP and Clustering

We can view *data points* as customers, *clusters* as tables.

CRP defines a prior distribution on partition (cluster) of data. Number of clusters can be induced

Combined with a parameterized probability distribution for each cluster, we can get a posterior

That's all we need in cluster settings

Contents

Gibbs Sampling DP Mixtures

Go back to original Gaussian Mixture clustering model. Let's place $DP(\alpha, H)$ as prior of G – DP mixture

Introduce z_i as cluster indicator for each x_i

We use collapsed Gibbs sampling for inference

$$p(z_i = k | \boldsymbol{z}_{-i}, \boldsymbol{x}, \boldsymbol{\nu}, \alpha) \propto p(z_i = k | \boldsymbol{z}_{-i}, \alpha) p(x_i | \boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\nu})$$

where

$$p(z_i = k | \boldsymbol{z}_{-i}, \alpha) = \frac{n_k}{\alpha + n - 1}$$

for existing component k among $\boldsymbol{z}_{-i}$, and

$$p(z_i = k | \boldsymbol{z}_{-i}, \alpha) = \frac{\alpha}{\alpha + n - 1}$$

for new component $K + 1$, $\forall j \neq i, z_j \neq K + 1$

Gibbs Sampling for DP Mixtures

What is $p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\nu})$?

For new component $K+1$, x_i is conditional independent from other data points

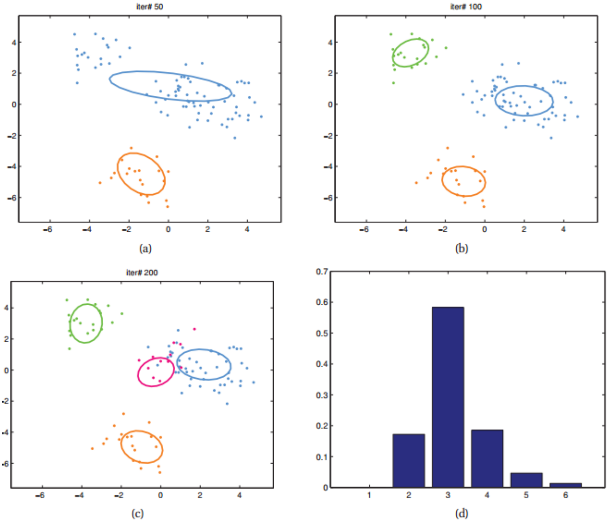
$$p(x_i|\boldsymbol{x}_{-i}, z_i = K+1, \boldsymbol{z}_{-i}, \boldsymbol{\nu}) = \int_\phi p(x_i|\phi)p(\phi|\nu)\,\mathrm{d}\phi$$

For existing component k, x_i is conditional independent from data points in different cluster

$$p(x_i|\boldsymbol{x}_{-i}, z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\nu}) = \frac{p(x_i, \boldsymbol{x}_{-i}^{(k)}|\nu)}{p(\boldsymbol{x}_{-i}^{(k)}|\nu)}$$

It's model specific.

Visualization



Other Methods

Hierarchical Dirichlet Processes

Infinite Hidden Markov Models

Polya Trees

Dirichlet (Diffusion) Trees

Dirichlet Forest

India Buffet Processes

(Zoubin Ghahramani, Non-parametric Bayesian Methods. 2005)

Outline

Thanks.

HP: http://keg.cs.tsinghua.edu.cn/jietang/
Email: jietang@tsinghua.edu.cn