S&DS 365 / 665

Intermediate Machine Learning

Nonparametric Bayes: Gaussian and Dirichlet Processes

(continued)

February 28

Yale

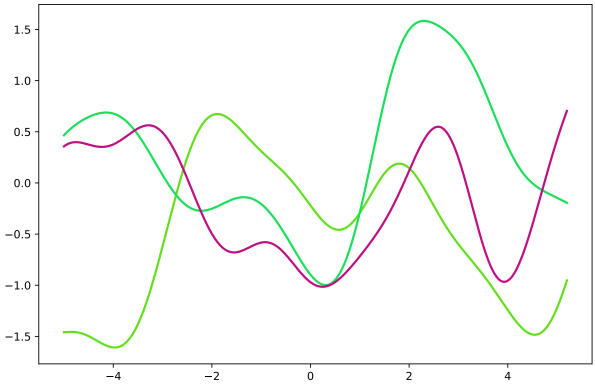
Reminders

- Assignment 2 is out
- Quiz 2 on Wednesday (CNN, GP, DP)
- Midterm on March 16 in class
  - practice exam next week
  - review week of March 14

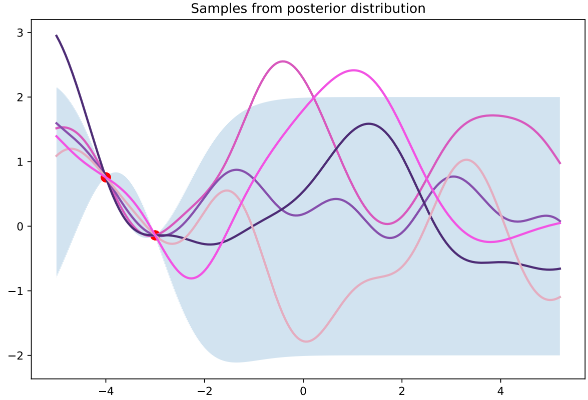
For Today

- Dirichlet process demos and definitions
- Next topic: Approximate inference

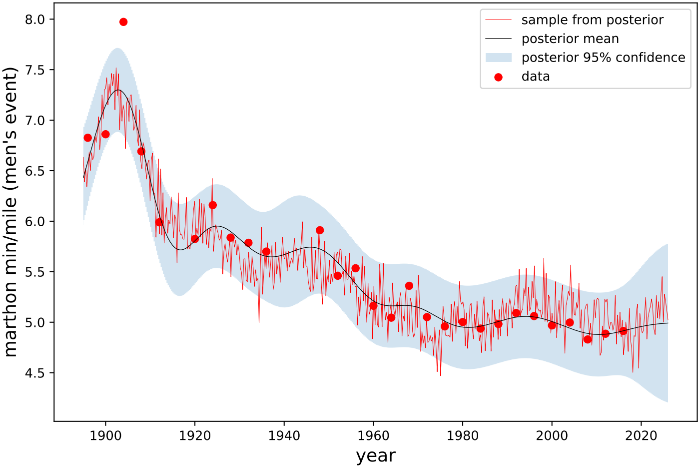
Last week's demo: GP samples



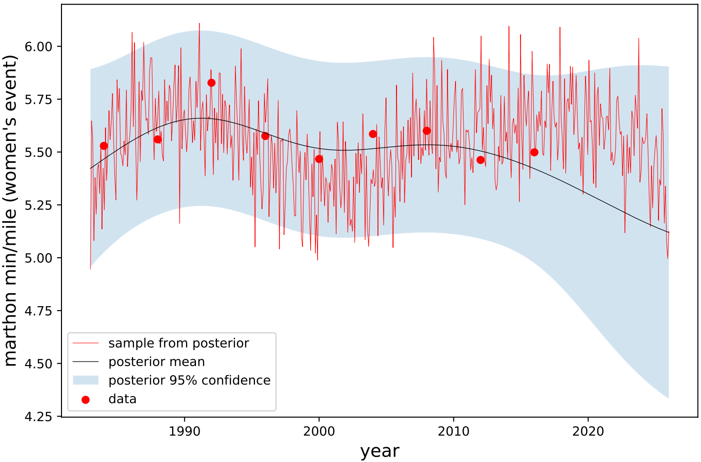
Last week's demo: GP samples



Olympic marathon times (men's race)



Olympic marathon times (women's race)



The Dirichlet Process

- The Dirichlet process is analogous to the Gaussian process
- Every partition of sample space has a Dirichlet distribution (more precise shortly)
- GPs are tools for regression functions; DPs are tools for distributions and densities
- DPs finesse the problem of choosing the number of components in a mixture model
  - Example: Don't need to specify the number of topics in a topic model

The Dirichlet Process

Dirichlet processes have some fun mnemonic metaphors, which help understand the concepts:

- Stick breaking
- Chinese restaurants

But it’s easy to get confused—we’re working with probability distributions over probability distributions

Starting point: CDF

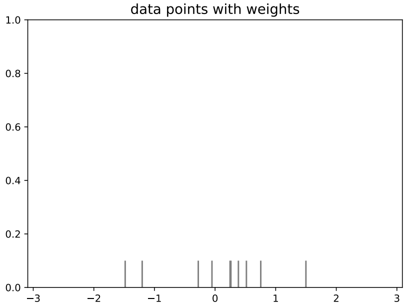
The *empirical distribution* of a set of data is the probability distribution that places probability mass $\frac{1}{n}$ on each data point $x_1, x_2, \ldots, x_n$.

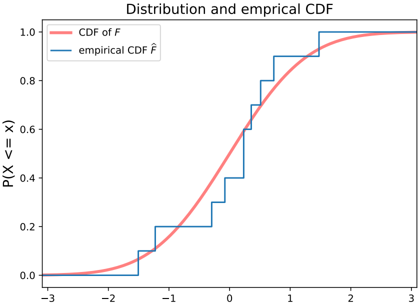
The *empirical CDF* is the function

$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i \leq x)$$

This is a step function with steps of size $\frac{1}{n}$ on each data point.

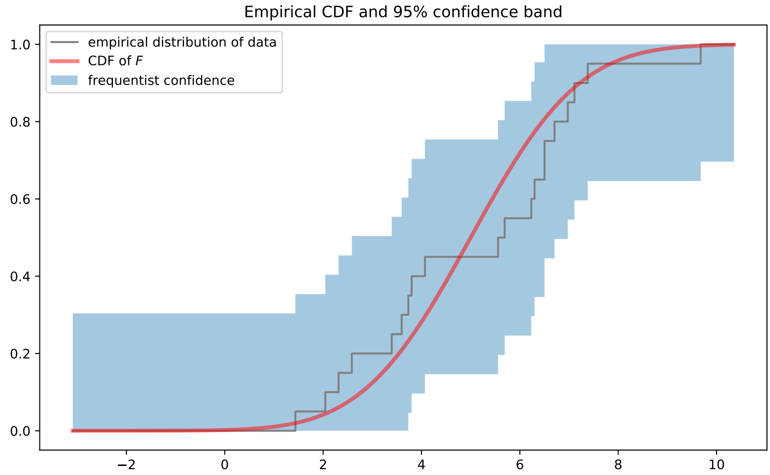
Empirical CDF





Empirical CDF

A frequentist 95% confidence band is given by $\widehat{F}(x) \pm \sqrt{\frac{1}{2n} \log\left(\frac{2}{.05}\right)}$

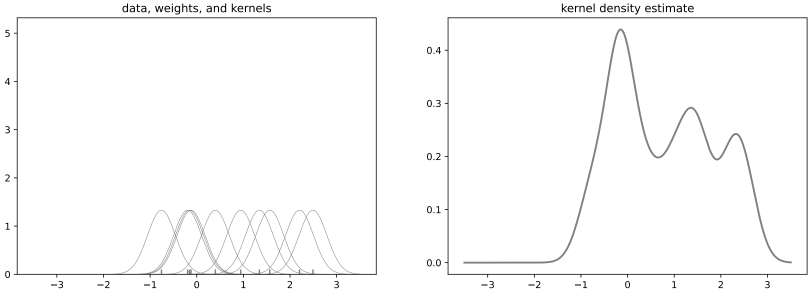


Recall: KDE

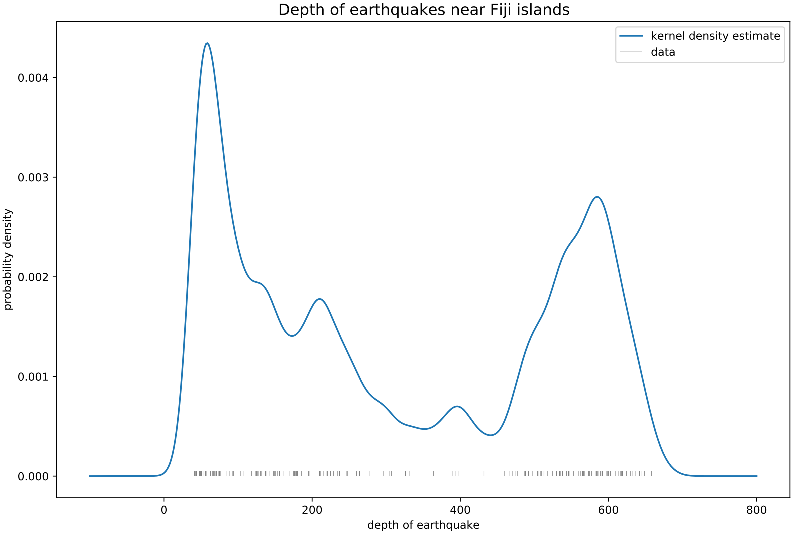
The *kernel density estimate* is the mixture model that places weight $\frac{1}{n}$ on the kernel bump function centered on each data point:

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

Recall: KDE


data, weights, and kernels
5
4
3
2
1
0
−3
−2
−1
0
1
2
3
kernel density estimate
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3

Recall: KDE



Depth of earthquakes near Fiji islands
kernel density estimate
data
probability density
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
depth of earthquake

Getting rid of the data

Both the empirical CDF and kernel density estimate involve the data

We want to construct a *prior* distribution over these objects, before we see any data

Solution: Use synthetic or "imaginary" data!

Think back to our interpretation of the Beta(α, β) prior.

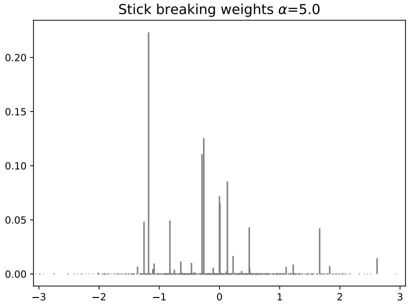
Dirichlet process

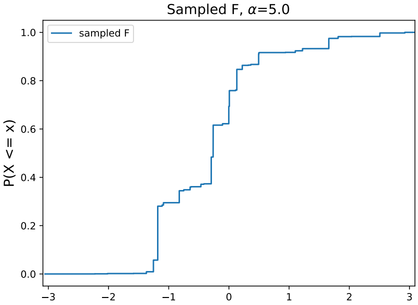
Each sample from a Dirichlet process prior has a *random collection of weights*, assigned to a *random selection of data*

Each sample from Dirichlet process mixture has a random collection of weights assigned to a random selection of *model parameters*

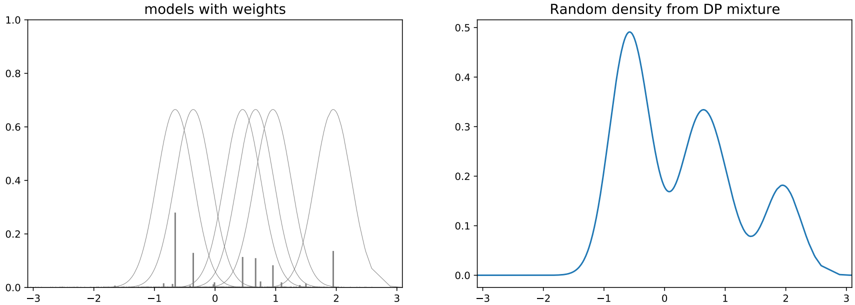
Recall our sticking breaking demo

Sample from DP prior





Sample from DP mixture



Stick breaking process

Stick breaking:

- At each step, break off a fraction $V \sim \text{Beta}(1, \alpha)$

"Imaginary data":

- At each step, sample $X \sim F_0$

Stick breaking process

To draw a single random distribution F from $\text{DP}(\alpha, F_0)$:

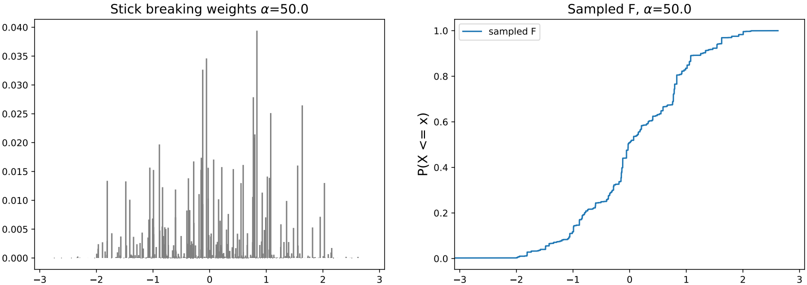
1. Draw $s_1, s_2, \ldots$ independently from F_0.
2. Draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$ and set $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$
3. Let F be the discrete distribution that puts mass w_j at s_j

Stick breaking process

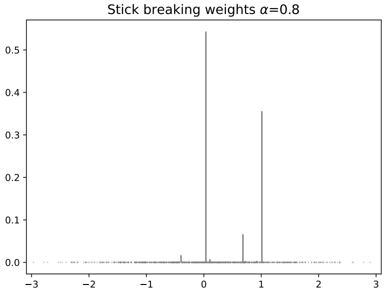
The mean of Beta$(1, \alpha)$ is $\frac{1}{1+\alpha}$.

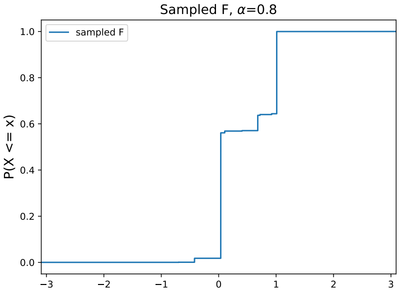
- As α gets larger, the weights get smaller
- Weights always sum to one

Different α



Different α





Clustering/repeats

Suppose we draw data F, drawn from a Dirichlet process, and then sample data from F:

$$F \sim DP(\alpha, F_0)$$

$$X_1, X_2, \ldots, X_n \mid F \sim F$$

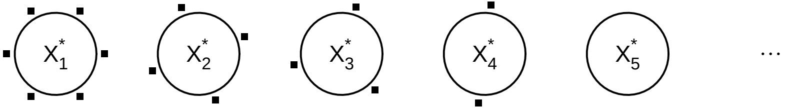
Since F is a mixture model, the samples X_i are clustered according to which mixture component they are sampled from.

The "Chinese restaurant process" captures this

Chinese restaurant mnemonic

Inspired by the large Chinese restaurants in San Francisco

Chinese restaurant mnemonic



A customer (data point) comes into the restaurant and either

1. sits at an empty table, with probability proportional to α, or
2. sits at an occupied table with probability proportional to number of customers already seated at that table

Chinese restaurant process

1. Draw $X_1 \sim F_0$.
2. For $i = 2, \ldots, n$: draw

$$X_i \mid X_1, \ldots X_{i-1} = \begin{cases} X \sim F_{i-1} & \text{with probability } \frac{i-1}{i+\alpha-1} \\ \\ X \sim F_0 & \text{with probability } \frac{\alpha}{i+\alpha-1} \end{cases}$$

   where F_{i-1} is the empirical distribution of $X_1, \ldots X_{i-1}$

This allows us to sample from the marginal distribution over X, without explicitly drawing a distribution F from the DP

Chinese restaurant process

Let $X_1^*, X_2^*, \ldots$ denote unique values of $X_1, \ldots, X_n$

Define cluster assignment variables $c_1, \ldots, c_n$ where $c_i = j$ means that X_i takes the value X_j^*

Let $n_j = |\{i : \; c_j = j\}|$. Then

$$X_n = \begin{cases} X_j^* & \text{with probability } \frac{n_j}{n+\alpha-1} \\ X \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha-1} \end{cases}$$

This allows us to sample from the marginal distribution over X, without explicitly drawing a distribution F from the DP

The posterior distribution

Let $X_1, \ldots, X_n \sim F$ and let F have prior $\pi = \text{Dir}(\alpha, F_0)$

Then the posterior π for F given $X_1, \ldots, X_n$ is

$$\text{Dir}\left(\alpha + n, \overline{F}_n\right)$$

where

$$\overline{F}_n = \frac{n}{n+\alpha} F_n + \frac{\alpha}{n+\alpha} F_0.$$

Here F_n is the empirical distribution of $X_1, \ldots, X_n$

This says that the Dirichlet process is conjugate to sampling from the distribution—the posterior is another DP

DP Demo

+ Code + Text Cannot save changes

RAM
Disk

Dirichlet process demo

In this notebook we demonstrate the Dirchlet process, using the stick breaking construction. First we sample from the prior distribution $\pi = DP(\alpha, F_0)$, then we sample from the posterior distribution $p(F\,|, x)$ given a data set $x_1, \ldots, x_n$ that is drawn from a distribution that is different from the base distribution F_0.

[1]
```
import numpy as np
import pandas as pd
from scipy.stats import norm
from IPython.display import clear_output
from time import sleep
import matplotlib.pyplot as plt
%matplotlib inline
```

Below we define three "helper" functions.

Given a parameter α, the `stick_break` function returns a set of weights $w_1, w_2, \ldots, w_N$ given by

$$w_i = \begin{cases} V_0 & \text{if } i = 0 \\ V_i(1 - V_{i-1}) \cdots (1 - V_0) & \text{if } i > 0 \end{cases}$$

where the random variables $V_0, V_1, \ldots, V_N$ are independent draws from a Beta$(1, \alpha)$ distribution. This gives a set of weights that sums to one (if N is large).

[2]
```
def stick_break(alpha, N):
    V = np.random.beta(1, alpha, size=N)
```

But what actually is a DP?

Recall:

A random function m is distributed according to a Gaussian process if for every $x_1, x_2, \ldots, x_n$ the random vector $m(x_1), \ldots, m(x_n)$ has a multivariate Gaussian distribution

$$N(\mu(x), K(x))$$

But what actually is a DP?

A random distribution F is distributed according to a Dirichlet process $DP(\alpha, F_0)$ if for every partition $A_1, \ldots, A_n$ of the sample space the random vector $F(A_1), \ldots, F(A_n)$ has a Dirichlet distribution

$$\mathrm{Dir}\left(\alpha F_0(A_1), \alpha F_0(A_2), \ldots, \alpha F_0(A_n)\right)$$

But what actually is a DP?

As a special case, if the sample space is the real line we can take the partition to be

$$A_1 = \{z \ : \ z \leq x\}$$

$$A_2 = \{z \ : \ z > x\}$$

and then

$$F(x) \sim \text{Beta}\Big(\alpha F_0(x), \alpha(1 - F_0(x))\Big)$$

Big picture

The definition tells us the precise sense in which a DP is an infinite Dirichlet distribution

But this is not concrete

The sticking breaking and Chinese restaurant processes give us *algorithms* for working with a DP

Big picture

Historically:

DP definition $\longrightarrow$ CRP $\longrightarrow$ Stick breaking

Big picture

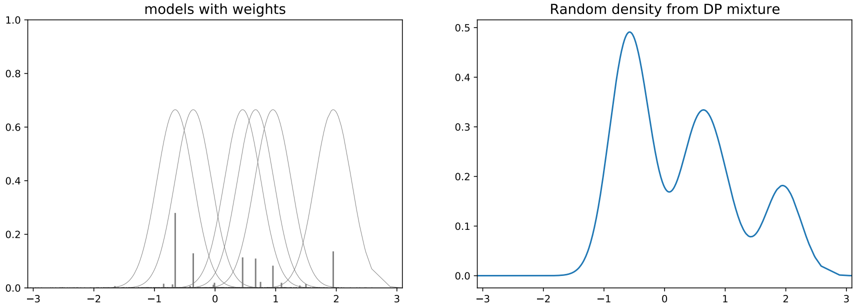
Conceptually, algorithmically:

DP definition $\longleftarrow$ CRP $\longleftarrow$ Stick breaking

From DP to DPM

- A DP is a distribution over distributions
- A Dirichlet process mixture is a distribution over mixture models
- DPMs are Bayesian versions of kernel density estimation
- Subject to the curse of dimensionality!
- In stick breaking we replace X_i by θ_i
- In Chinese restaurant process we replace X_i^* by θ_i^*

Sample from DP mixture



Nonparametric Bayesian mixture model

$$
\begin{aligned}
F &\sim \mathrm{DP}(\alpha, F_0) \\
\theta_1, \ldots, \theta_n \,|\, F &\sim F \\
X_i \,|\, \theta_i &\sim f(x \,|\, \theta_i), \quad i = 1, \ldots, n.
\end{aligned}
$$

Stick breaking process for DPM

Stick breaking:

- At each step, break off a fraction $V \sim \text{Beta}(1, \alpha)$

Sample model parameters:

- At each step, sample $\theta \sim F_0$

Stick breaking process for DPM

To draw a single random mixture from DPM(α, F_0):

1. Draw $\theta_1, \theta_2, \ldots$ independently from F_0.
2. Draw $V_1, V_2, \ldots \sim \text{Beta}(1, \alpha)$ and set $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$
3. Let f be the (infinite) mixture model

$$f(x) = \sum_{j=1}^{\infty} w_j f(x \mid \theta_j)$$

Chinese restaurant process for a DPM

1. Draw $\theta_1 \sim F_0$.
2. For $i = 2, \ldots, n$: draw

$$\theta_i \,|\, \theta_1, \ldots \theta_{i-1} = \begin{cases} \theta \sim F_{i-1} & \text{with probability } \frac{i-1}{i+\alpha-1} \\ \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{i+\alpha-1} \end{cases}$$

where F_{i-1} is the empirical distribution of $\theta_1, \ldots \theta_{i-1}$

Chinese restaurant process for a DPM

Let $\theta_1^*, \theta_2^*, \ldots$ denote unique values of $\theta_1, \ldots, \theta_n$

Define cluster assignment variables $c_1, \ldots, c_n$ where $c_i = j$ means that θ_i takes the value θ_j^*

Let $n_j = |\{i : \; c_j = j\}|$. Then

$$\theta_n = \begin{cases} \theta_j^* & \text{with probability } \frac{n_j}{n+\alpha-1} \\ \theta \sim F_0 & \text{with probability } \frac{\alpha}{n+\alpha-1} \end{cases}$$

The posterior for a DPM

- The posterior distribution does not have a closed form — need to approximate it algorithmically
- Two forms of approximations: Gibbs sampling and variational methods — next topic

Summary

- A Dirichlet process is a prior over distribution functions
- The stick breaking process tells us how to sample F
- The Chinese restaurant proccess tells us how to sample X
- A Dirichlet process is a Bayesian version of the empirical CDF
- A Dirichlet process mixture is a Bayesian version of kernel density estimation
- Bayesian nonparametric methods require a lot of conceptual machinery and computation