S&DS 365 / 665

Intermediate Machine Learning

Nonparametric Bayes: Gaussian and Dirichlet Processes

February 23

Yale

Reminders

- Assignment 1 due tonight
- Assignment 2 posted tonight
- Quiz 2 next Wednesday (longer availability)
- Midterm on March 16 in class

For Today

- Gaussian processes
- Examples
- Dirichlet process prélude

Bayesian Inference

The parameter θ of a model is viewed as a random variable. Inference usually carried out as follows:

- Choose a *generative model* $p(x \mid \theta)$ for the data.
- Choose a *prior distribution* $\pi(\theta)$ that expresses beliefs about the parameter before seeing any data.
- After observing data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, update beliefs and calculate the *posterior distribution* $p(\theta \mid \mathcal{D}_n)$.

Nonparametric Bayes

- In nonparametric Bayesian inference, we replace a finite dimensional model θ with an infinite dimensional model
- This is usually a class of *functions* (regression functions, densities)
- Typically neither the prior nor the posterior have a density; but the posterior is still well defined.

Core questions

1. How do we construct a prior π on an infinite dimensional set $\mathcal{F}$?
2. How do we compute the posterior? How do we draw random samples from the posterior?
3. What are the properties of the posterior?

Essential methods

We'll explore these questions in three settings

Statistical problem	**Frequentist approach**	**Bayesian approach**
regression	kernel smoother	Gaussian process
CDF estimation	empirical cdf	Dirichlet process
density estimation	kernel density estimator	Dirichlet process mixture

Stochastic processes

A stochastic process is a collection of random variables indexed some set (such as time), all defined with respect to a common probability space.

We'll focus on two fundamental stochastic processes:

- Gaussian processes
- Dirichlet processes

More technically, a stochastic process $\{X(t)\}_{t \in T}$ is a collection of random variables indexed by a set T and defined on a common probability space $(\Omega, \mathcal{F}, P)$ where Ω is a sample space, $\mathcal{F}$ is a σ-algebra, and P is a probability measure.

Gaussian processes

The nonparametric regression model is

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where $\mathbb{E}(\epsilon_i) = 0$.

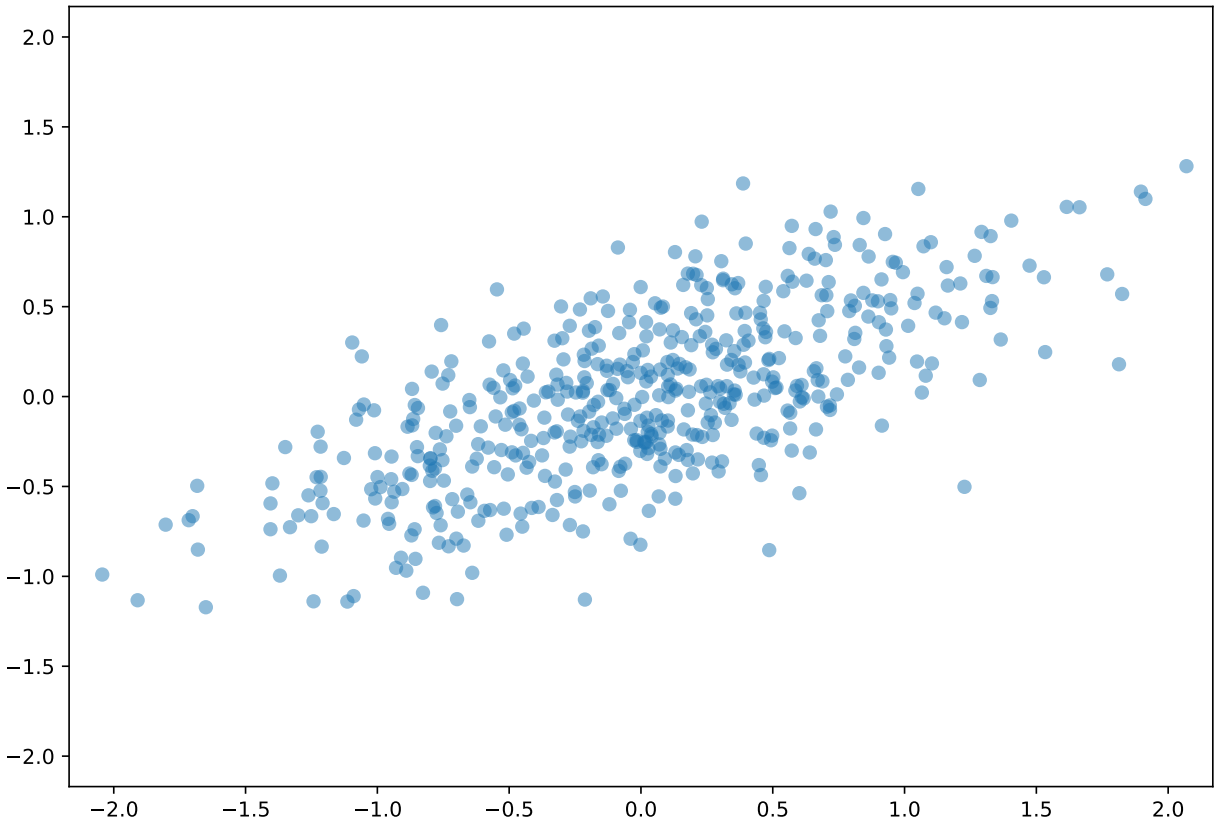
The frequentist kernel estimator for m is

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

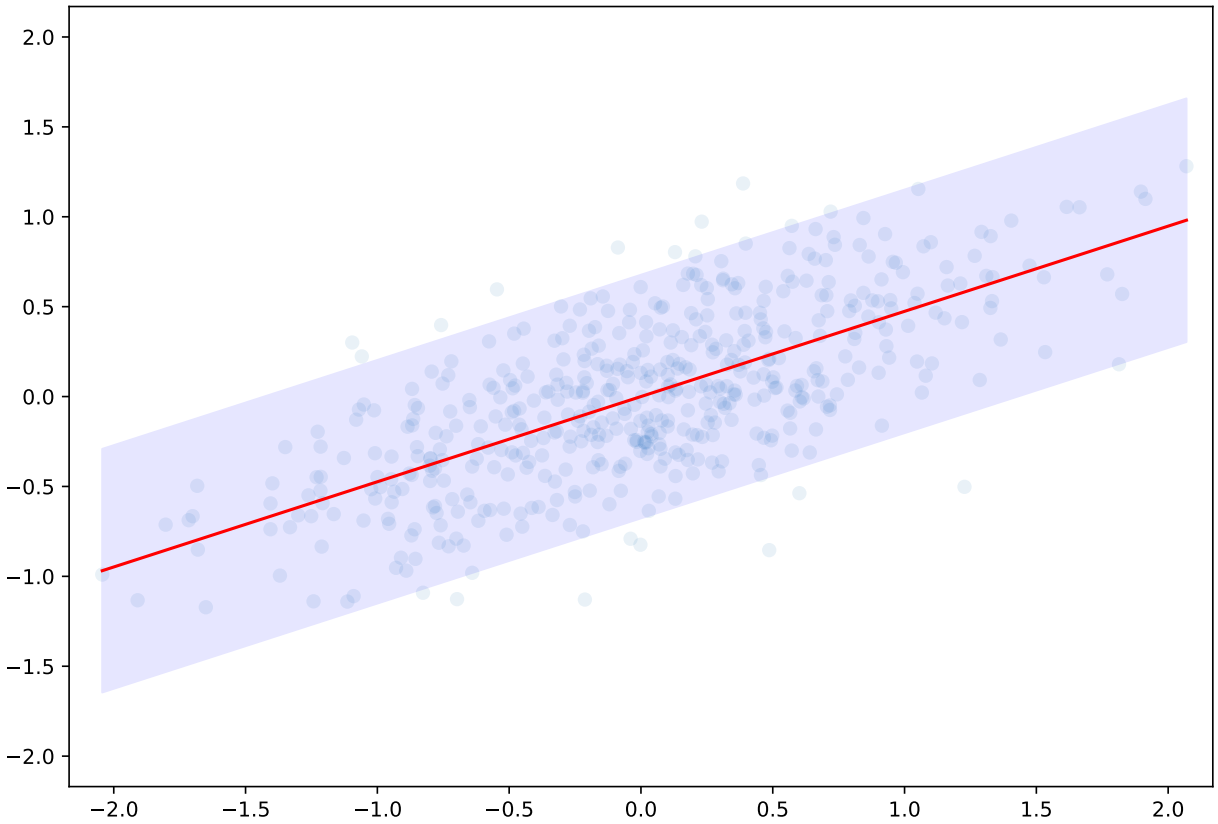
where K is a kernel and h is a bandwidth.

Bayesian version requires prior π on set of regression functions

Starting point: Conditionals of Gaussian



Starting point: Conditionals of Gaussian



Gaussian conditionals

If (X_1, X_2) are jointly Gaussian with distribution

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \right)$$

then the conditional distributions are also Gaussian and given by

$$X_1 \mid x_2 \sim N\left(\frac{K_{12}}{K_{22}} x_2,\ K_{11} - \frac{K_{12}^2}{K_{22}} \right)$$

$$X_2 \mid x_1 \sim N\left(\frac{K_{12}}{K_{11}} x_1,\ K_{22} - \frac{K_{12}^2}{K_{11}} \right)$$

Gaussian process

A stochastic process $m(x)$ indexed by $x \in \mathbb{R}$ is a *Gaussian process* if for each set of points $x_1, \ldots, x_n$ the vector $(m(x_1), m(x_2), \ldots, m(x_n))^T$ is normally distributed:

$$(m(x_1), m(x_2), \ldots, m(x_n))^T \sim N(\mu(x), K(x))$$

where $K_{ij}(x) = K(x_i, x_j)$ is a Mercer kernel.

When $x_1, \ldots, x_n$ are fixed we will denote the $n \times n$ matrix with entries $K(x_i, x_j)$ by $\mathbb{K}$.

The definition makes sense when indexing by any set $\mathcal{X}$ for an appropriately defined Mercer kernel.

Gaussian process prior

Let's assume $\mu = 0$, so prior mean function is zero

Density of the Gaussian process prior of $m = (m(x_1), \ldots, m(x_n))$ is

$$\pi(m) = (2\pi)^{-n/2} |\mathbb{K}|^{-1/2} \exp\left(-\frac{1}{2} m^T \mathbb{K}^{-1} m\right).$$

Under change of variables $m = \mathbb{K}\alpha$, we have $\alpha \sim N(0, \mathbb{K}^{-1})$ and

$$\pi(\alpha) = (2\pi)^{-n/2} |\mathbb{K}|^{1/2} \exp\left(-\frac{1}{2} \alpha^T \mathbb{K} \alpha\right).$$

Gaussian processes prior

What functions have high probability according to the Gaussian process prior?

The prior favors $m^T \mathbb{K}^{-1} m$ being small. If v is an eigenvector of $\mathbb{K}$, with eigenvalue λ, then

$$\frac{1}{\lambda} = v^T \mathbb{K}^{-1} v$$

- Eigenfunctions of the Mercer kernel K with *large* eigenvalues are favored by the prior
- These correspond to smooth functions; the eigenfunctions that are very wiggly correspond to small eigenvalues

Using the likelihood

We observe $Y_i = m(x_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. So, log-likelihood is

$$\log p(Y \mid m) = -\frac{1}{2\sigma^2} \sum_i (Y_i - m(x_i))^2 + C$$

where $C = -\log(\sqrt{2\pi\sigma^2})$.

Log-posterior is

$$\begin{aligned} \log p(Y \mid m) + \log \pi(m) &= -\frac{1}{2\sigma^2} \| Y - \mathbb{K}\alpha \|_2^2 - \frac{1}{2} \alpha^T \mathbb{K} \alpha + C' \\ &= -\frac{1}{2\sigma^2} \| Y - \mathbb{K}\alpha \|_2^2 - \frac{1}{2} \| \alpha \|_K^2 + C' \end{aligned}$$

C' is just another constant.

Calculating the posterior

In Bayesian *maximum a posteriori (MAP)* inference, one estimates the mode of the posterior.

The posterior mean (and mode) is

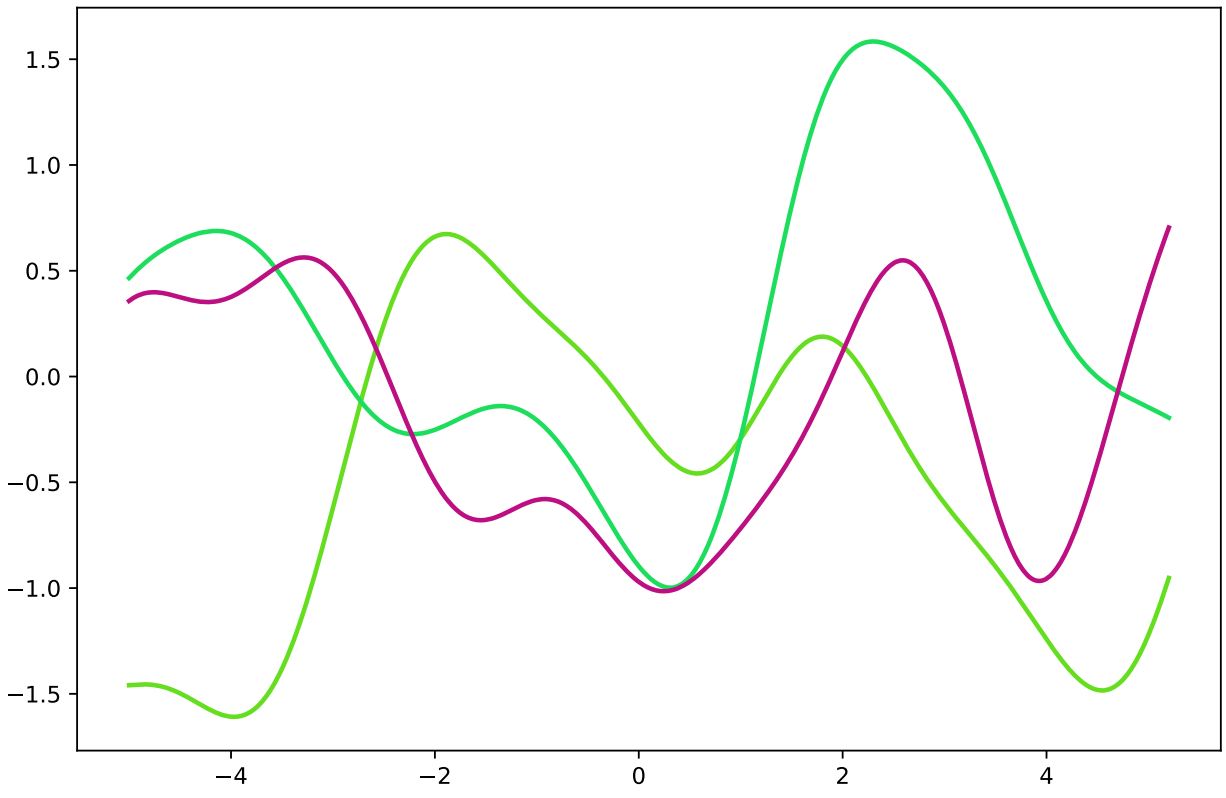
$$\mathbb{E}(\alpha \,|\, Y) = \left(\mathbb{K} + \sigma^2 I\right)^{-1} Y$$

and thus

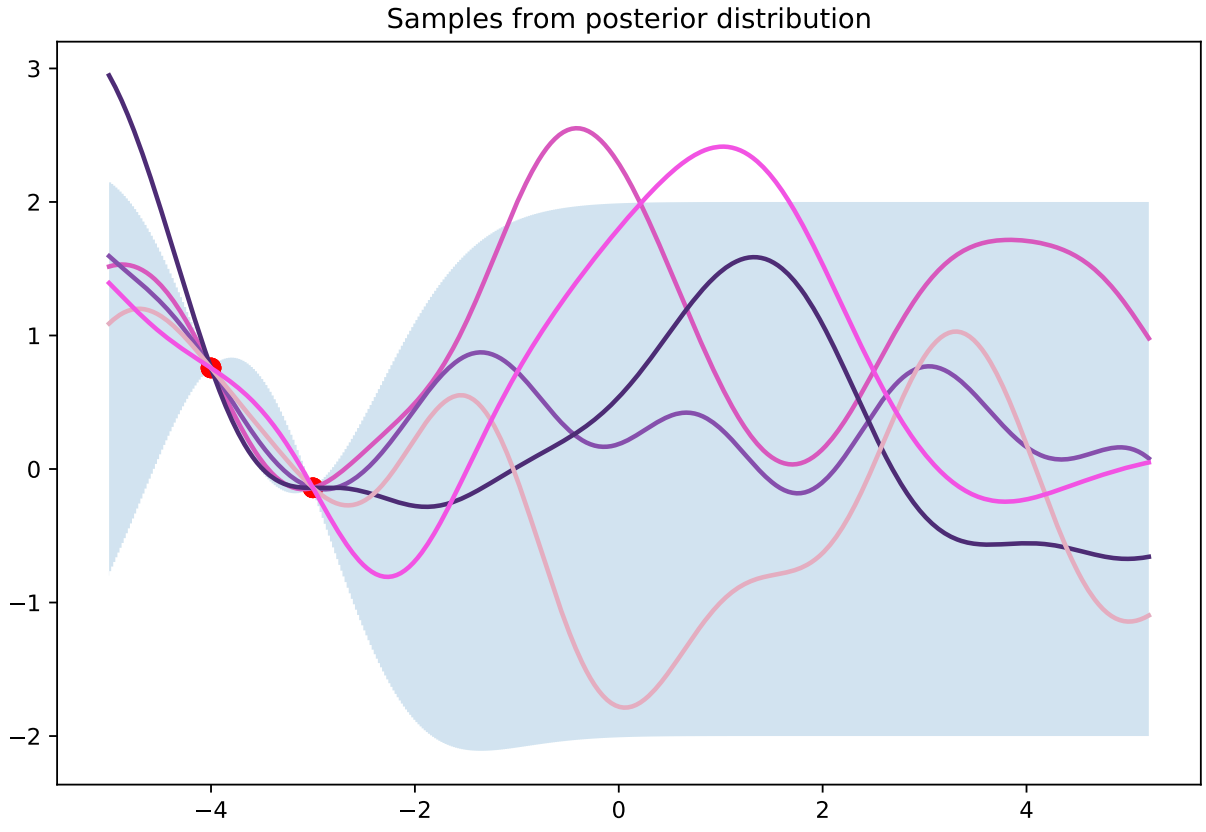
$$\widehat{m} = \mathbb{E}(m \,|\, Y) = \mathbb{K}\left(\mathbb{K} + \sigma^2 I\right)^{-1} Y.$$

Equivalent to Mercer kernel regression

Samples from prior and posterior



Samples from prior and posterior



Gaussian conditionals

If (X_1, X_2) are jointly Gaussian with distribution

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix} \right)$$

then the conditional distributions are also Gaussian and given by

$$\begin{aligned} X_1 \,|\, x_2 &\sim N\left(\mu_1 + CB^{-1}(x_2 - \mu_2),\ A - CB^{-1}C^T\right) \\ X_2 \,|\, x_1 &\sim N\left(\mu_2 + C^T A^{-1}(x_1 - \mu_1),\ B - C^T A^{-1} C\right) \end{aligned}$$

The matrix $A - CB^{-1}C^T$ is called the *Schur complement* of B.

Predicting at a new point

How do we predict $Y_{n+1} = m(x_{n+1}) + \epsilon_{n+1}$?

Let k be the vector

$$k = (K(x_1, x_{n+1}), \ldots, K(x_n, x_{n+1})).$$

Then $(Y_1, \ldots, Y_{n+1})$ are jointly Gaussian with covariance

$$\begin{pmatrix} \mathbb{K} + \sigma^2 I & k \\ k^T & K(x_{n+1}, x_{n+1}) + \sigma^2 \end{pmatrix}.$$

Predictive distribution

Using above expression for Gaussian conditionals:

The posterior mean and variance are

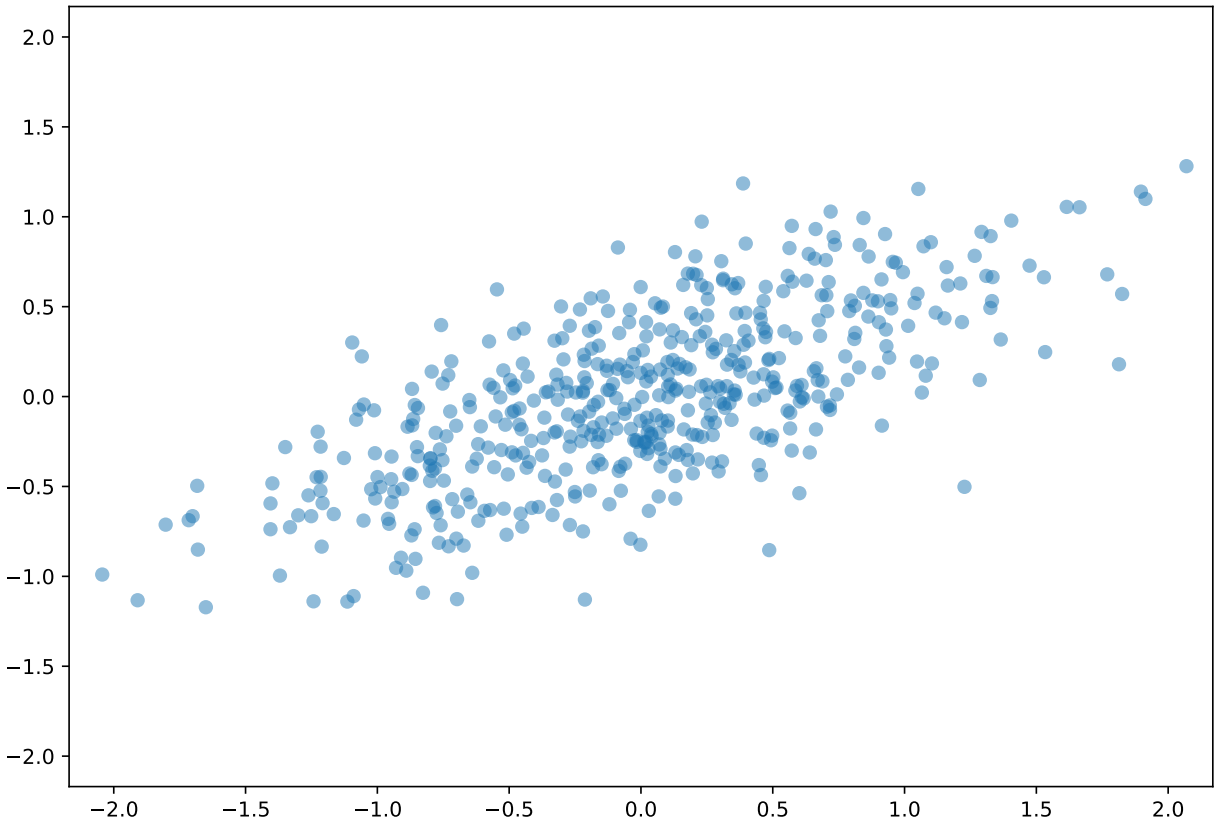
$$\mathbb{E}(Y_{n+1} \mid x_{1:n}, Y_{1:n}) = k^T(\mathbb{K} + \sigma^2 I)^{-1} Y$$
$$\text{Var}(Y_{n+1} \mid x_{1:n}, Y_{1:n}) = K(x_{n+1}, x_{n+1}) + \sigma^2 - k^T(\mathbb{K} + \sigma^2 I)^{-1} k$$

Predictive distribution

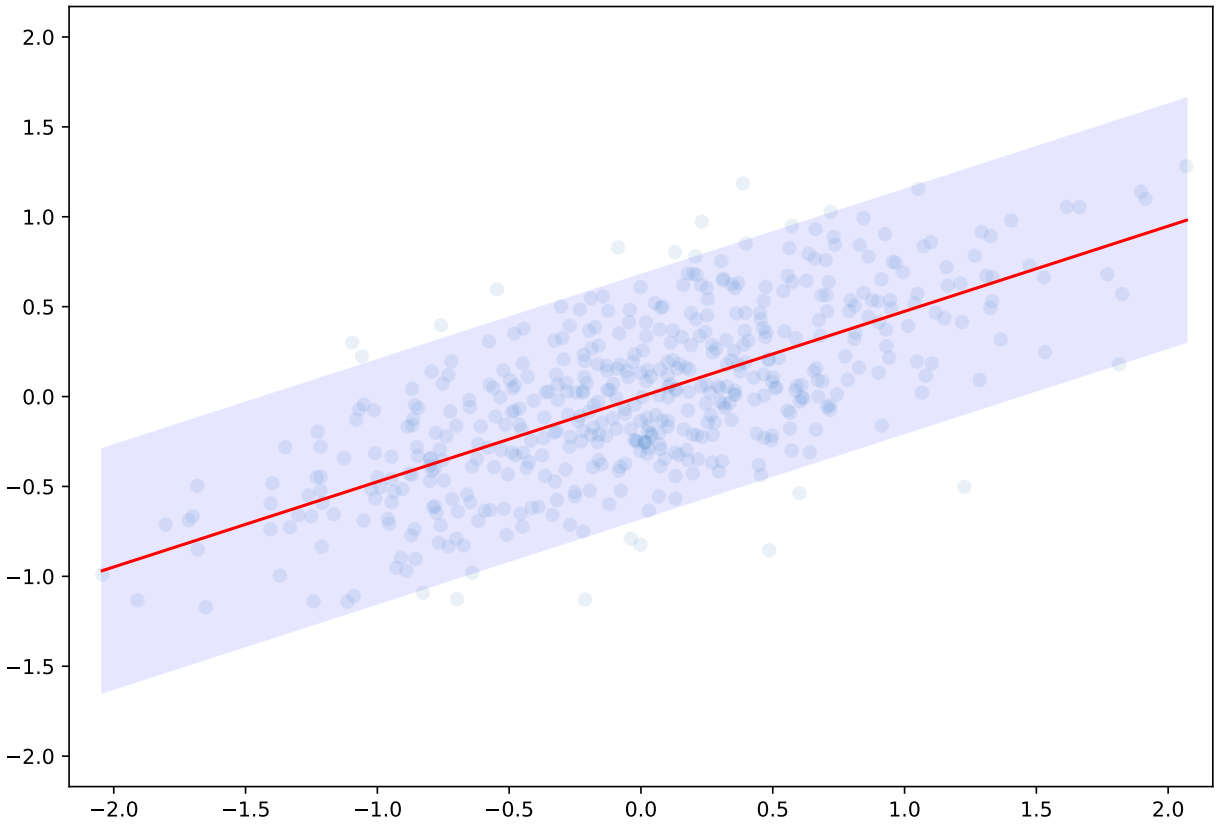
- Note that the mean is identical to what we saw for Mercer kernel regression
- But now we get a measure of uncertainty (the variance), which comes from the Gaussian process assumption

Let’s look at the notebook demo

Starting point: Conditionals of Gaussian



Starting point: Conditionals of Gaussian



Gaussian conditionals

If (X_1, X_2) are jointly Gaussian with distribution

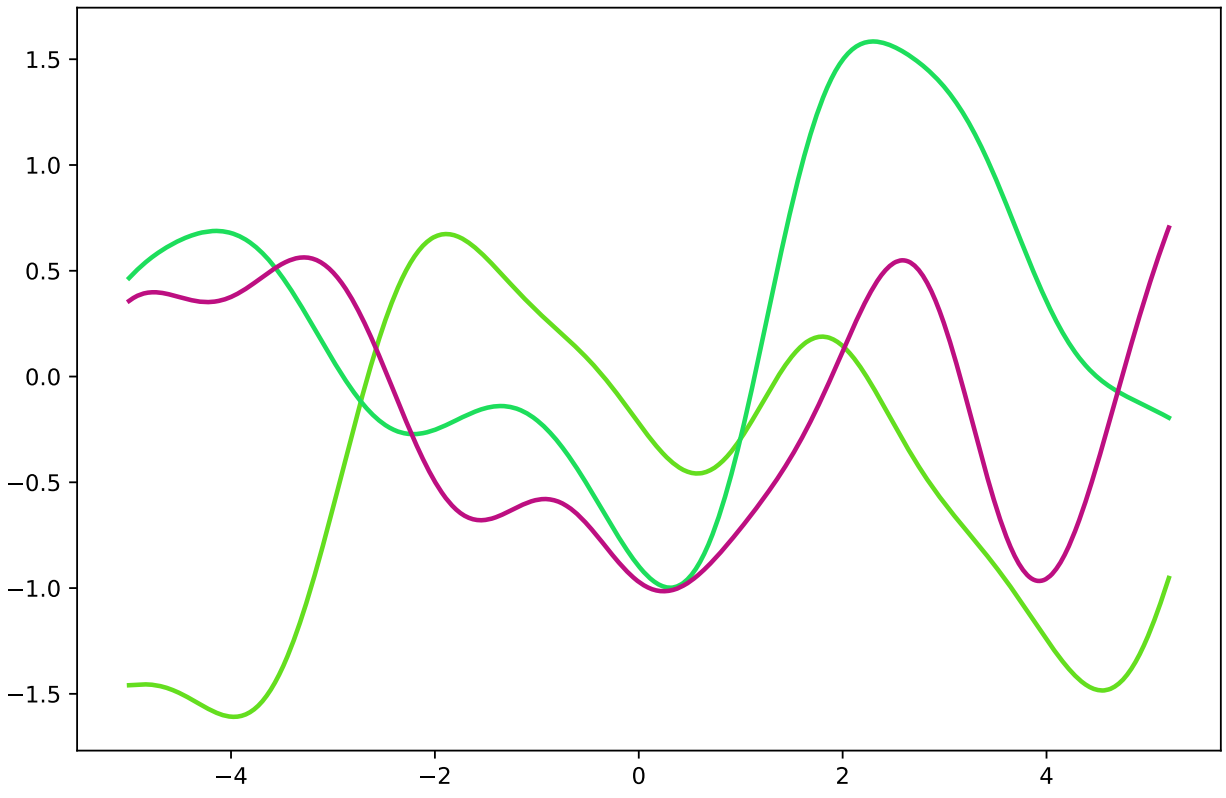
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \right)$$

then the conditional distributions are also Gaussian and given by

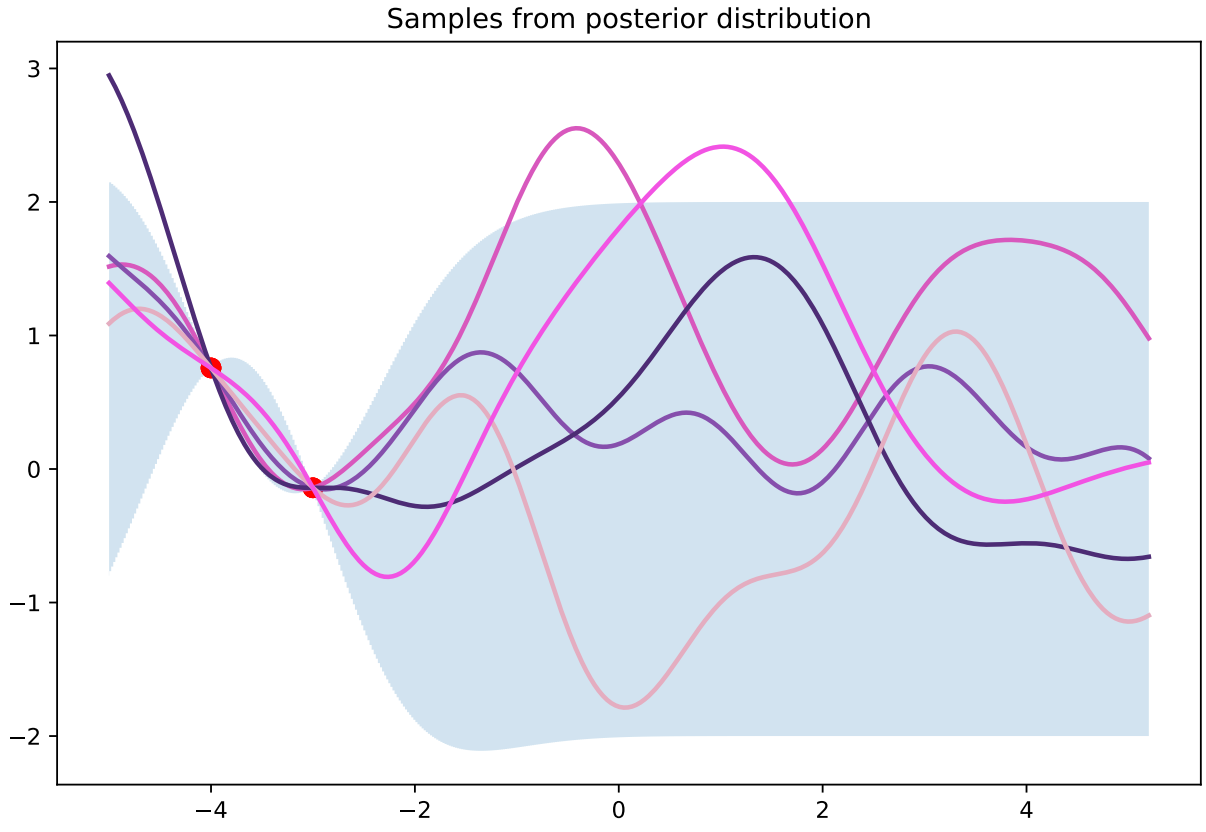
$$X_1 \mid x_2 \sim N\left(\frac{K_{12}}{K_{22}} x_2,\ K_{11} - \frac{K_{12}^2}{K_{22}} \right)$$

$$X_2 \mid x_1 \sim N\left(\frac{K_{12}}{K_{11}} x_1,\ K_{22} - \frac{K_{12}^2}{K_{11}} \right)$$

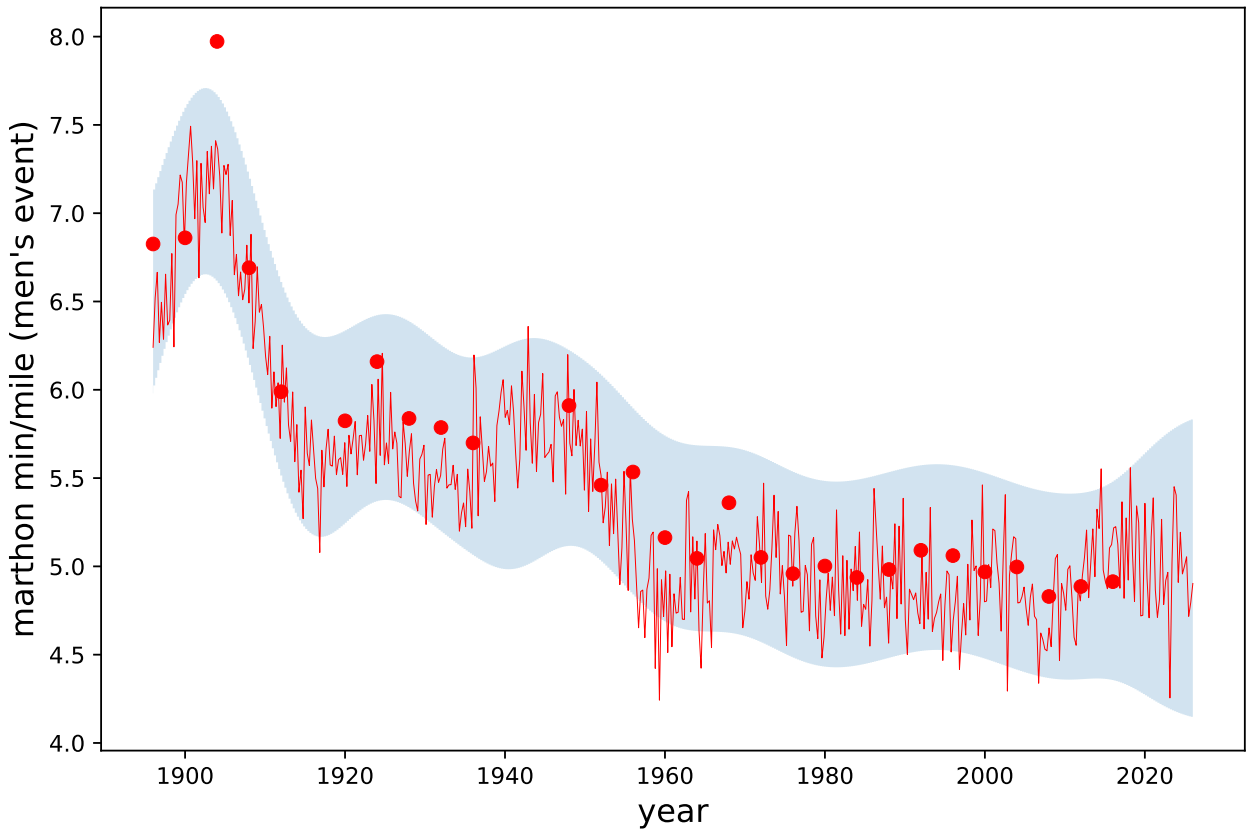
Samples from prior and posterior



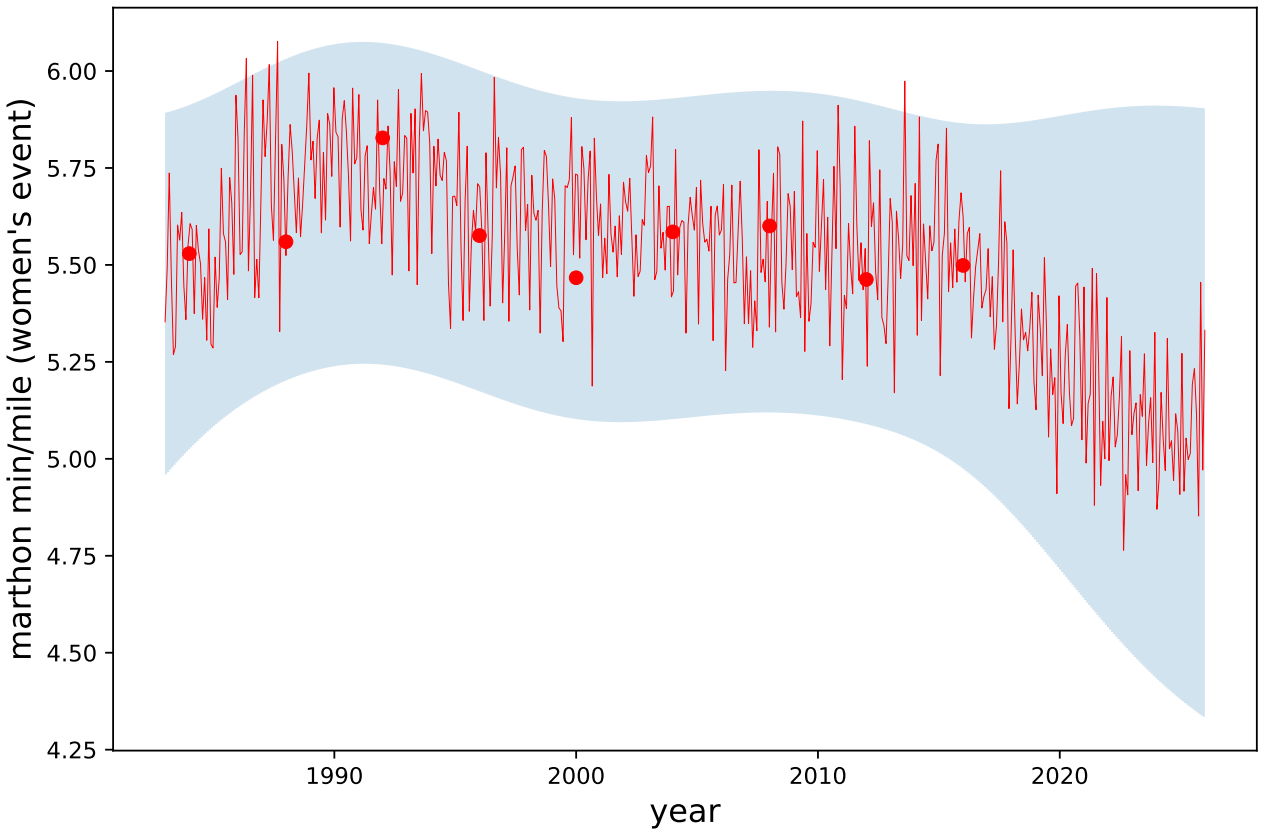
Samples from prior and posterior



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 the input 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

Be patient with yourself!

...and with me :-)

Starting point: CDF

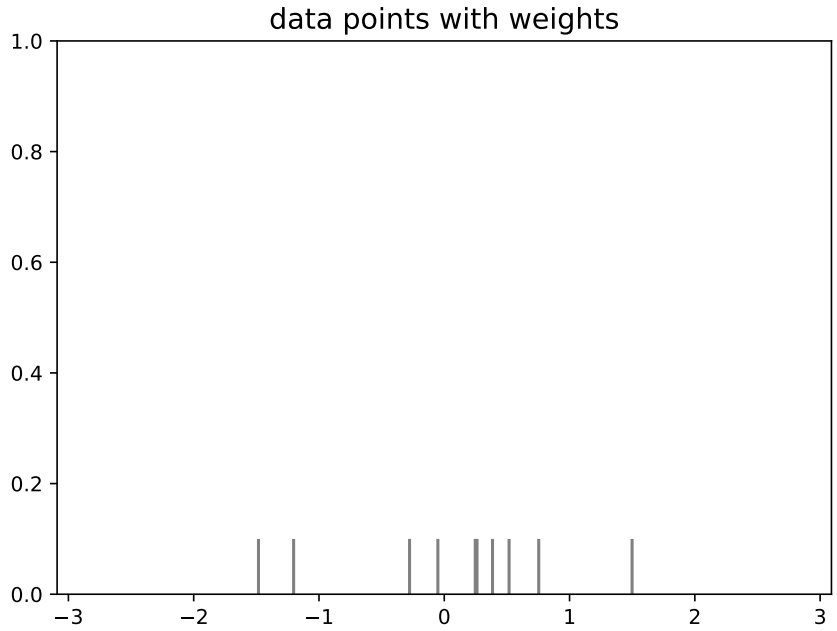
The *empirical distribution* of a set of data is the probability distribution that places probability mass $\frac{1}{n}$ on each data point $x_1, x_2, \ldots, x_n$.

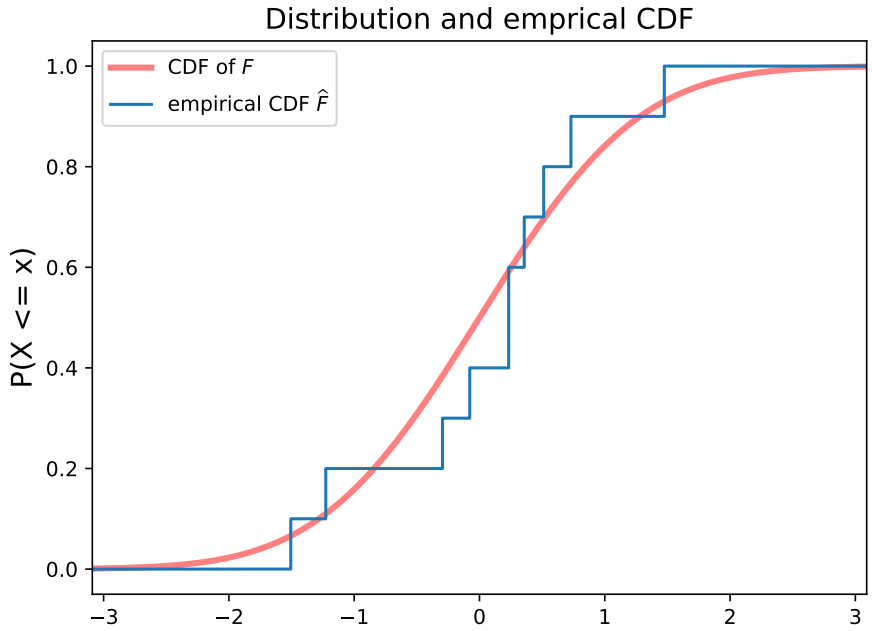
The *empirical CDF* is the function

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

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

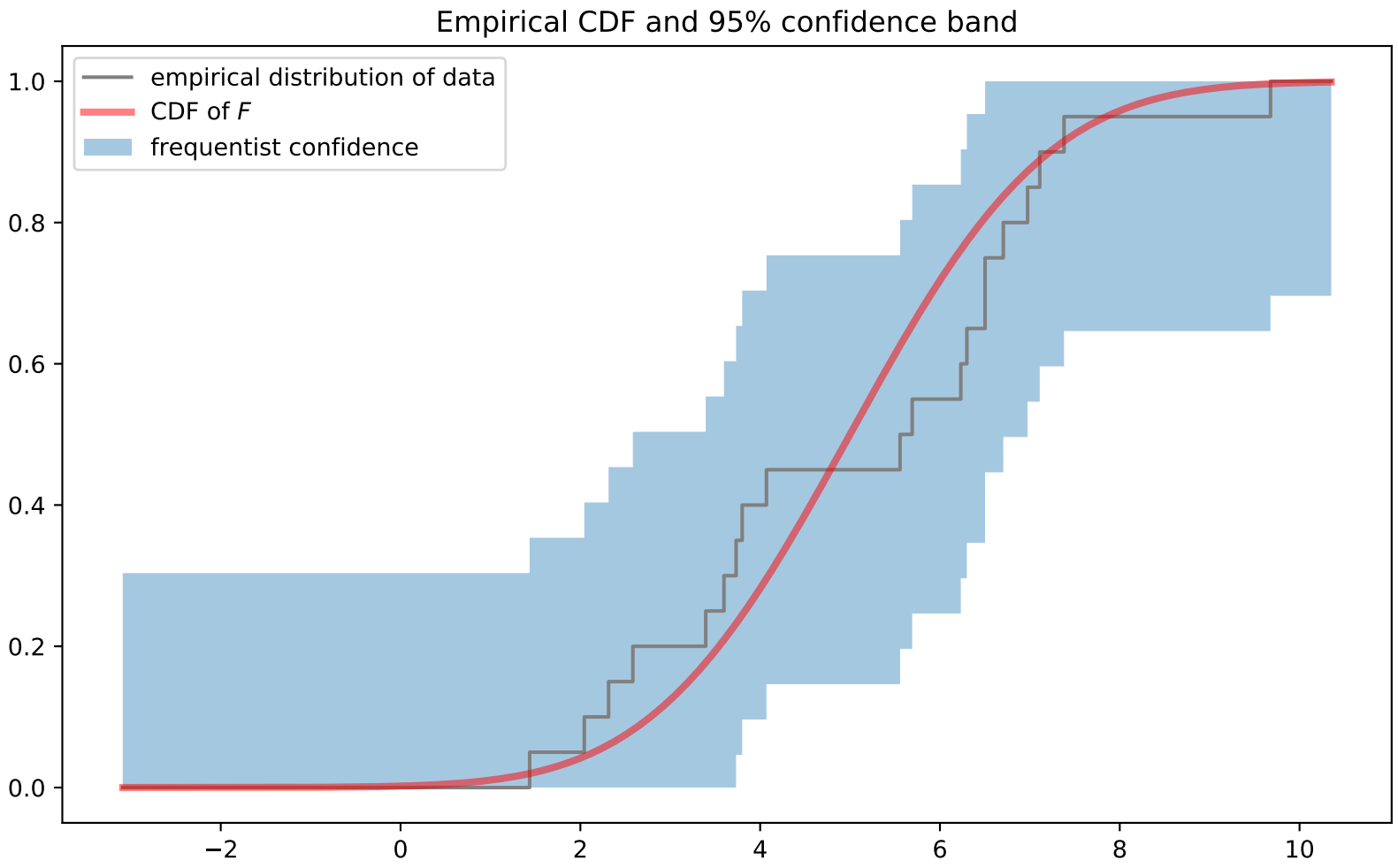
Empirical CDF





Empirical CDF

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

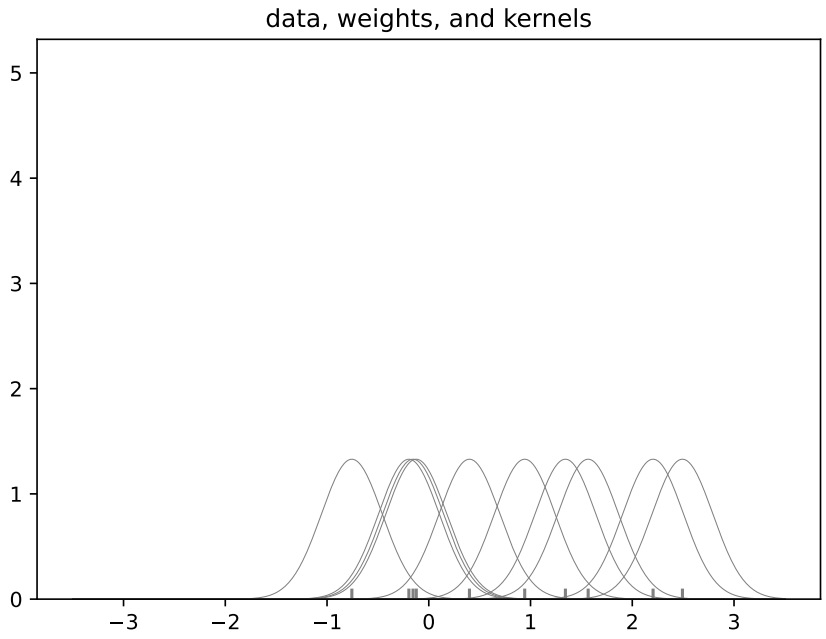


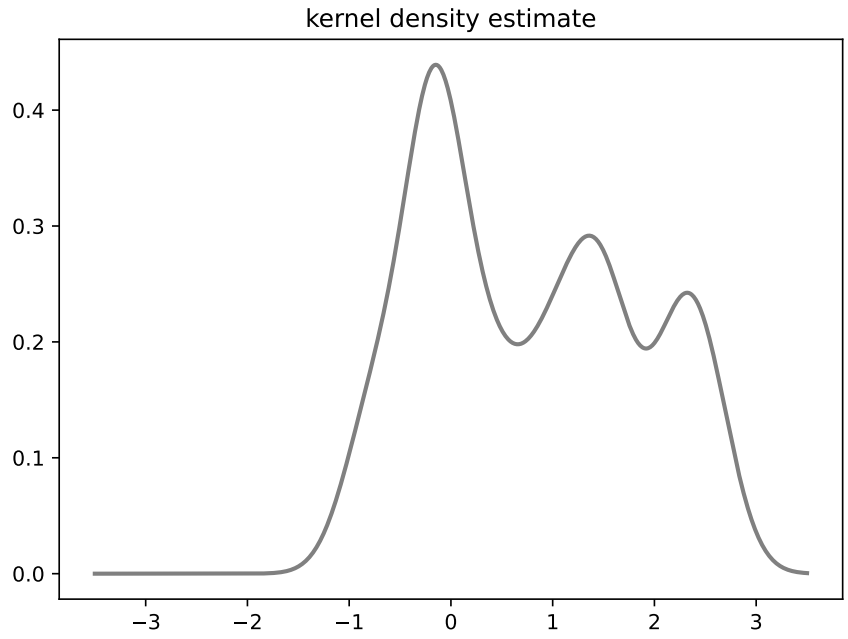
Recall: KDE

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

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

Kernel density estimate





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.

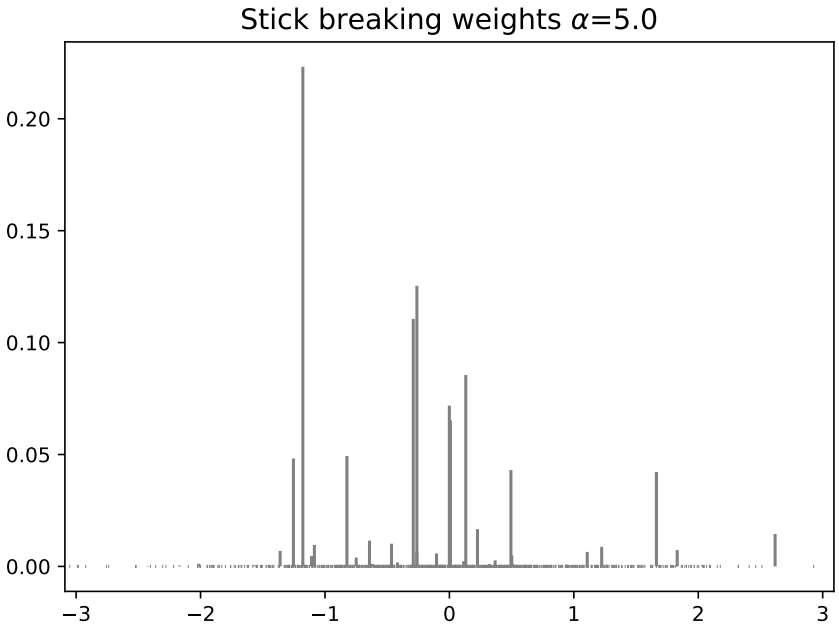
Dirichlet process

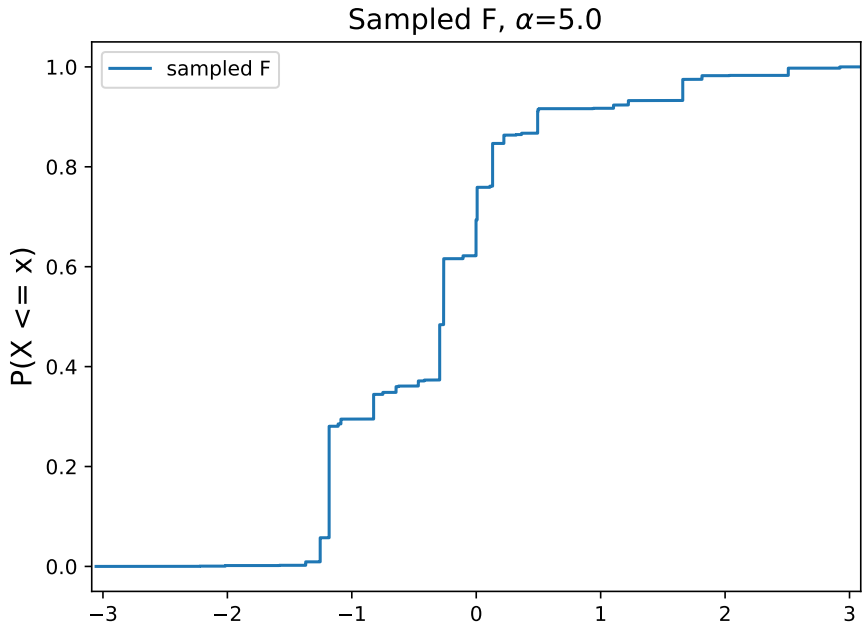
The Dirichlet process has a *random collection of weights*, assigned to a *random selection of data*

The Dirichlet process mixture has a random collection of weights assigned to a random selection of *models*

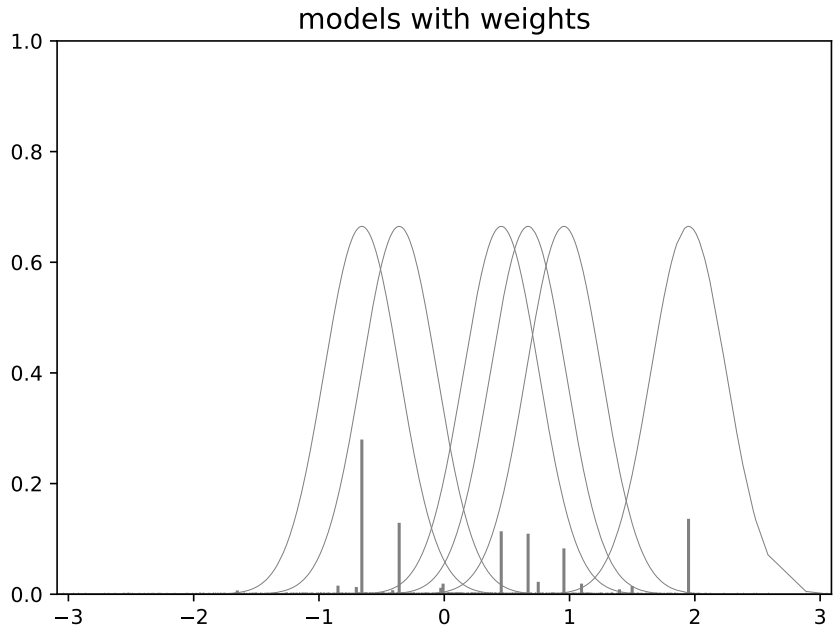
Let’s go to the demo

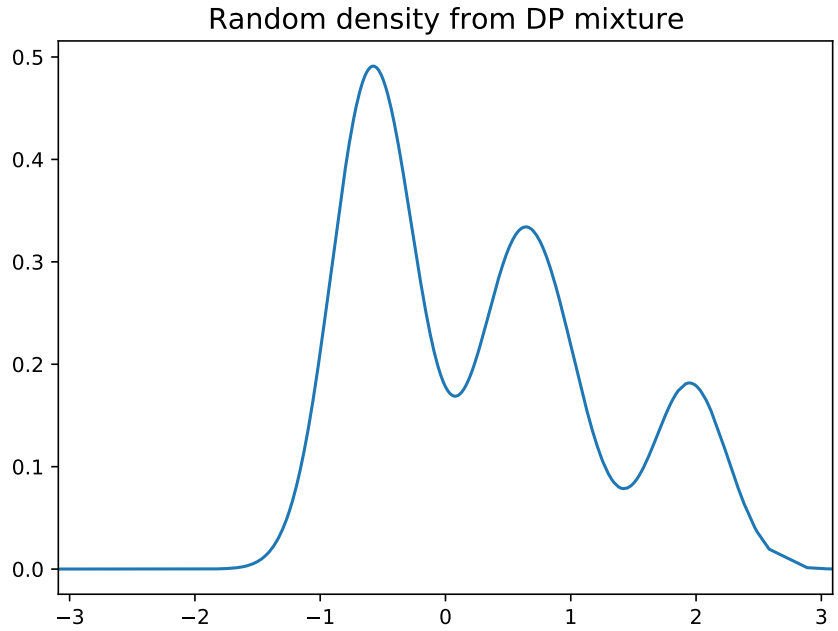
Sample from DP prior





Sample from DP mixture





More details next class!

Summary

- In nonparametric Bayes, the "parameters" are functions
- A Gaussian process is a stochastic process m where each collection of random variables $m(x_1), m(x_2), \ldots, m(x_n)$ is jointly Gaussian
- Gaussian processes are Bayesian versions of kernel regression; the posterior mean is equivalent to Mercer kernel regression
- A Dirichlet process is a prior over distribution functions
- A Dirichlet process mixture is a Bayesian version of kernel regression