

S&DS 365 / 665
Intermediate Machine Learning

Nonparametric Bayes: Gaussian Processes

September 30

Reminders

- Assignment 2 is out
- Quiz 3 on Wednesday
- Midterm on Monday, October 14 in class

For today

- Gaussian processes (continued)
- Examples

Bayesian Inference

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

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

In machine learning, Bayesian inference is preferred by some researchers as a way of introducing uncertainty

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*
- 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?

Nonparametric Bayes procedures may not have coverage and consistency properties of frequentist procedures

Essential methods

We'll explore these questions in a couple of 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

NP Bayes

- Nonparametric Bayesian inference can be subtle and technical
- Part of the machine learning toolkit
- Underlying probability theory can be beautiful
- We'll introduce the main techniques to give a flavor
- The notes go into more technical detail

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 a fundamental stochastic process: The Gaussian process

We'll also briefly mention the Dirichlet process — you'll be responsible for the definition

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 (Ω, \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

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

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

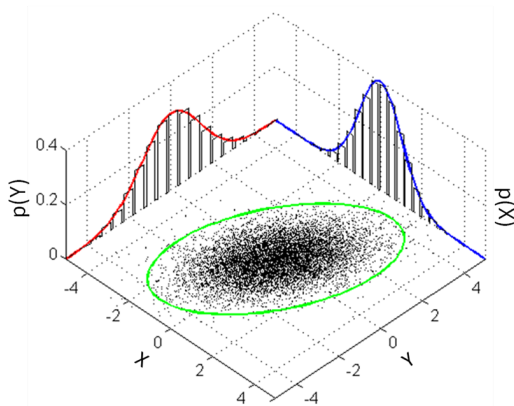
The frequentist kernel estimator for m is

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

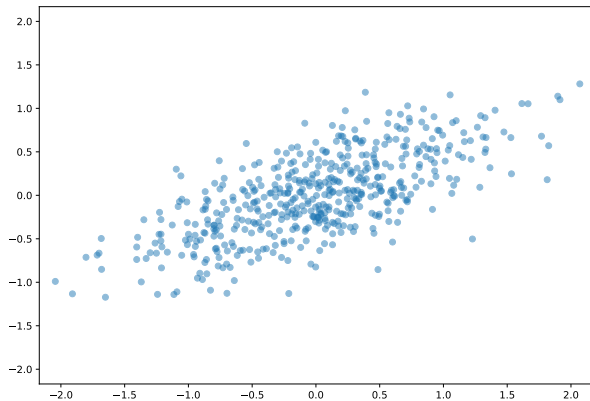
where K is a kernel and h is a bandwidth.

Bayesian version requires prior π on set of regression functions

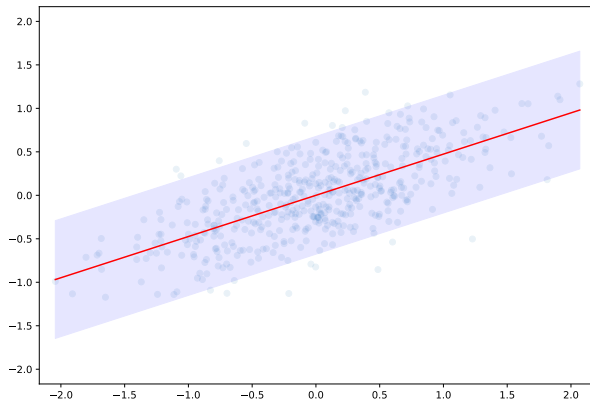
Everything boils down to Gaussian marginals and conditionals



Starting point: Conditionals of Gaussian



Starting point: Conditionals of Gaussian



Gaussian conditionals

If (X_1, X_2) are jointly Gaussian with distribution

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

then the conditional distributions are also Gaussian and given by

$$X_1 | x_2 \sim N \left(\mu_1 + CB^{-1}(x_2 - \mu_2), A - CB^{-1}C^T \right)$$

$$X_2 | x_1 \sim N \left(\mu_2 + C^T A^{-1}(x_1 - \mu_1), B - C^T A^{-1}C \right)$$

Gaussian conditionals

If $(X_1, X_2) \in \mathbb{R}^2$ 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

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

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

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, \dots, x_n the vector $(m(x_1), m(x_2), \dots, m(x_n))^T$ is normally distributed:

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

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

As before, if x_1, \dots, x_n are fixed we denote the $n \times n$ matrix with entries $K(x_i, x_j)$ by \mathbb{K} .

Gaussian process prior

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

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

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

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

Log-posterior is

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

Calculating the posterior

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

The posterior mean (and mode) is

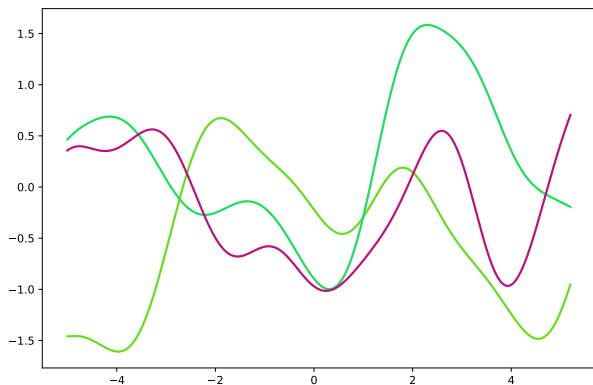
$$\mathbb{E}(\alpha \mid Y) = \left(\mathbb{K} + \sigma^2 I \right)^{-1} Y$$

and thus

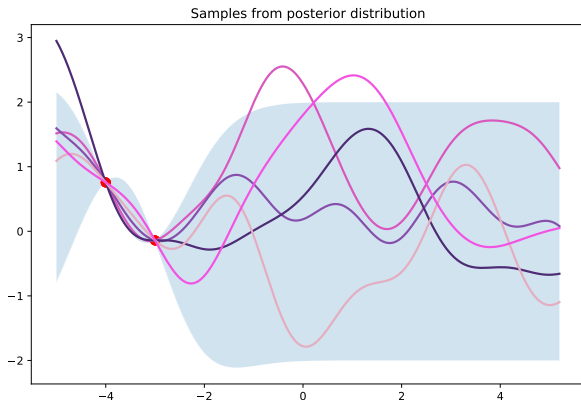
$$\hat{m} = \mathbb{E}(m \mid Y) = \mathbb{K} \left(\mathbb{K} + \sigma^2 I \right)^{-1} Y.$$

Equivalent to Mercer kernel regression

Samples from prior and posterior



Samples from prior and posterior



Predicting at a new point

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

Let k be the vector

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

Then (Y_1, \dots, 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

All from: 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

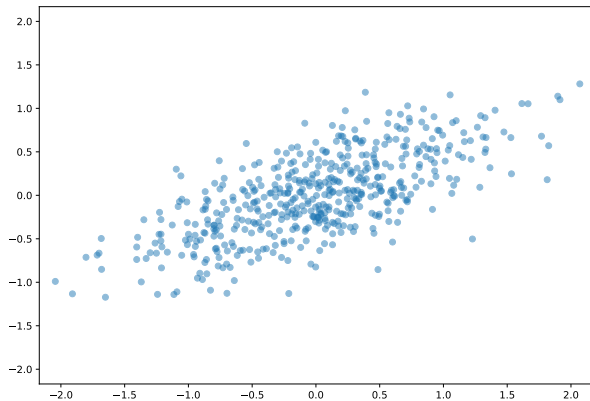
$$X_1 | x_2 \sim N \left(\mu_1 + CB^{-1}(x_2 - \mu_2), A - CB^{-1}C^T \right)$$

$$X_2 | x_1 \sim N \left(\mu_2 + C^T A^{-1}(x_1 - \mu_1), B - C^T A^{-1}C \right)$$

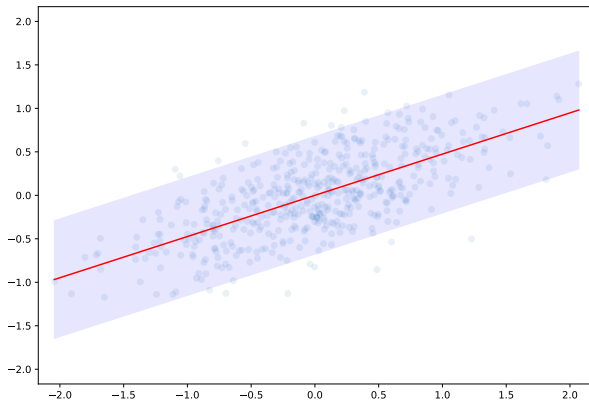
Let's look at the notebook demo

(plots from the demo follow)

Starting point: Conditionals of Gaussian



Starting point: Conditionals of Gaussian



Gaussian conditionals

If (X_1, X_2) are jointly Gaussian with distribution

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

then the conditional distributions are also Gaussian and given by

$$X_1 | x_2 \sim N \left(\mu_1 + CB^{-1}(x_2 - \mu_2), A - CB^{-1}C^T \right)$$

$$X_2 | x_1 \sim N \left(\mu_2 + C^TA^{-1}(x_1 - \mu_1), B - C^TA^{-1}C \right)$$

Gaussian conditionals

If (X_1, X_2) are jointly Gaussian with distribution

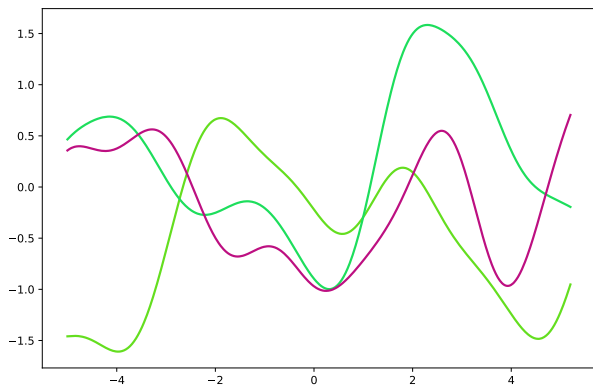
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \right)$$

then the conditional distributions are also Gaussian and given by

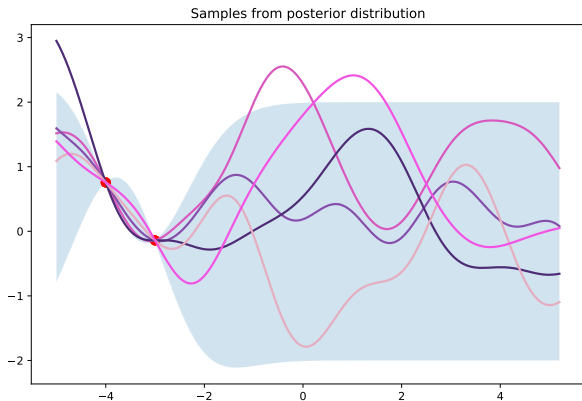
$$X_1 | x_2 \sim N \left(\frac{K_{12}}{K_{22}} x_2, K_{11} - \frac{K_{12}^2}{K_{22}} \right)$$

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

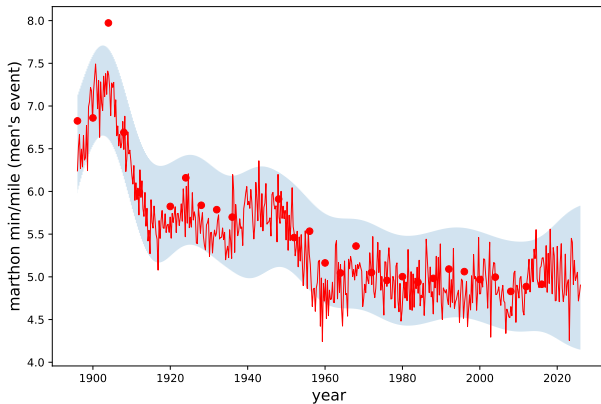
Samples from prior and posterior



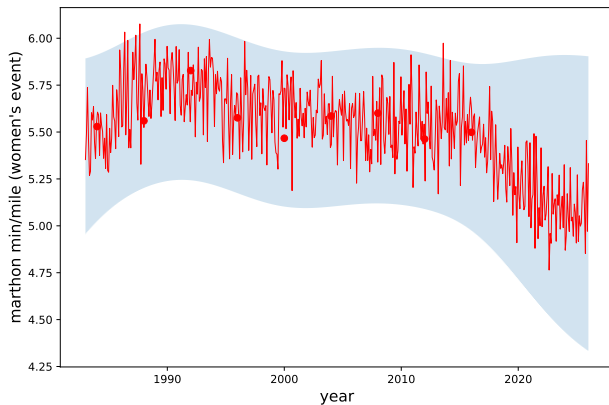
Samples from prior and posterior



Olympic marathon times (men's race)



Olympic marathon times (women's race)



The next few slides give a *very* brief overview of the Dirichlet process.

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: Number of topics in a topic model

Relation to KDEs

- 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!

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, \dots, x_n the random vector $m(x_1), \dots, m(x_n)$ 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, \dots, A_n of the sample space the random vector $F(A_1), \dots, F(A_n)$ has a Dirichlet distribution

$$\text{Dir}(\alpha F_0(A_1), \alpha F_0(A_2), \dots, \alpha F_0(A_n))$$

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

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

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

See notes for an introduction to these ideas (not required for this course)

Summary

- In a Bayesian approach, the parameters are random, and the data are fixed.
- In nonparametric Bayes, the “parameters” are functions
- A Gaussian process is a stochastic process m where each collection of random variables $m(x_1), m(x_2), \dots, m(x_n)$ is jointly Gaussian
- Calculation with GPs uses Gaussian conditioning via Schur complements
- Gaussian processes are Bayesian versions of kernel regression; the posterior mean is equivalent to Mercer kernel regression