S&DS 365 / 665

Intermediate Machine Learning

Nonparametric Bayes: Gaussian Processes

(continued)

October 3

Yale

Reminders

- Assignment 2 posted (CNNs and GPs)
- Quiz 3 on Wednesday, Oct 5
- Midterm on Tuesday, October 17 in class
- Practice exam and review next week

For Today

- Gaussian processes (continued)
- Examples
- Dirichlet process (intro)

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 $\mathcal{D}_n = \{x_1, \dots, x_n\}$, update beliefs and calculate the *posterior distribution* $p(\theta \mid \mathcal{D}_n)$.

Please posted notes for a review of some of the basics of Bayesian inference.

Bayes' Theorem

The posterior distribution can be written as

$$p(\theta \,|\, x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \,|\, \theta)\pi(\theta)}{p(x_1, \ldots, x_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

where $\mathcal{L}_n(\theta)$ is the *likelihood function* and

$$c_n = p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n \,|\, \theta)\pi(\theta)d\theta = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called *evidence*.

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?

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

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.

Regression model

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

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.

As before, if $x_1, \ldots, x_n$ are fixed we 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

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

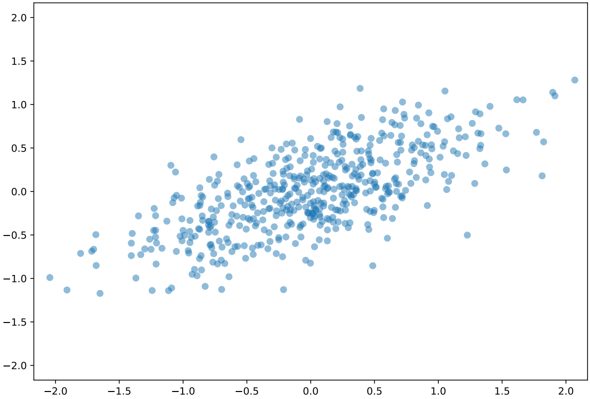
Equivalent to Mercer kernel regression

Let’s look at the notebook demo

(plots from the demo follow)

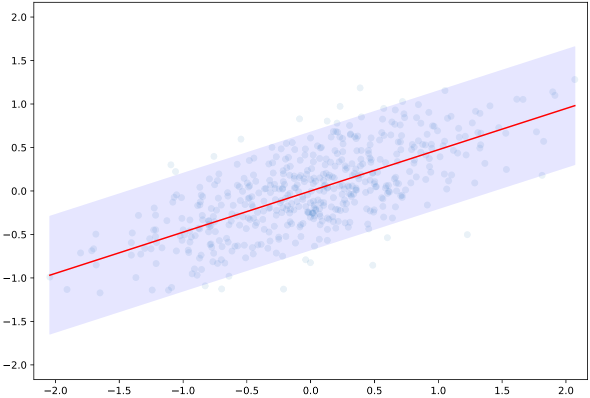
Conditionals of Gaussian

Posterior is calculated using Gaussian conditionals

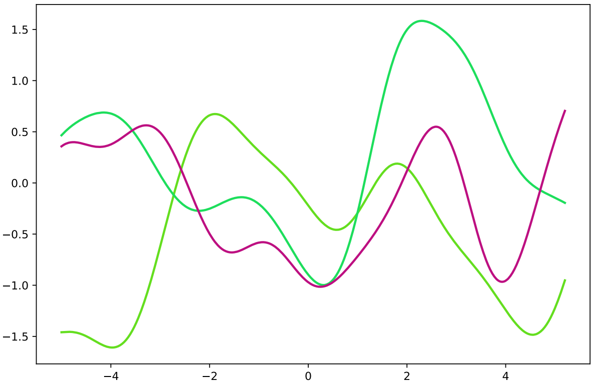


Conditionals of Gaussian

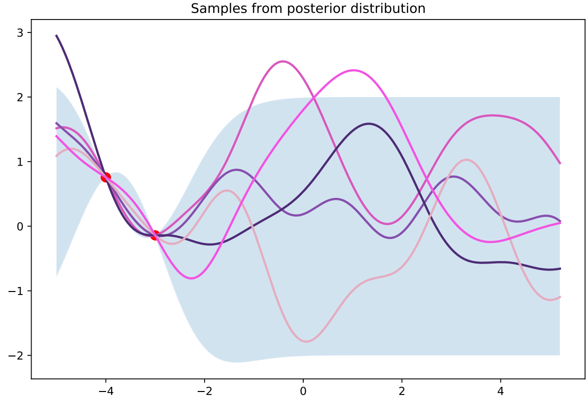
Posterior is calculated using Gaussian conditionals



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

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

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

$$\begin{aligned}\mathbb{E}(Y_{n+1} \,|\, x_{1:n}, Y_{1:n}) &= k^T(\mathbb{K} + \sigma^2 I)^{-1} Y \\ \mathrm{Var}(Y_{n+1} \,|\, x_{1:n}, Y_{1:n}) &= K(x_{n+1}, x_{n+1}) + \sigma^2 - k^T(\mathbb{K} + \sigma^2 I)^{-1} k\end{aligned}$$

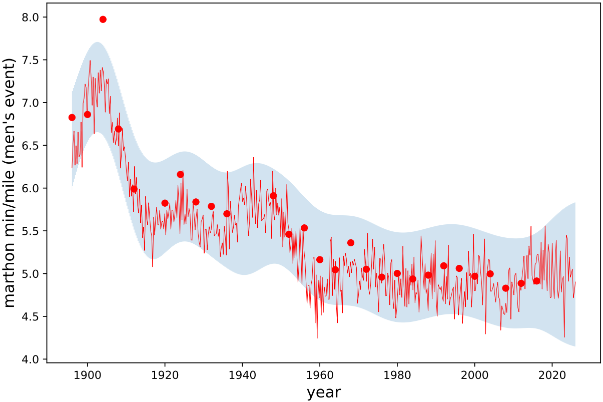
Predictive distribution

- 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

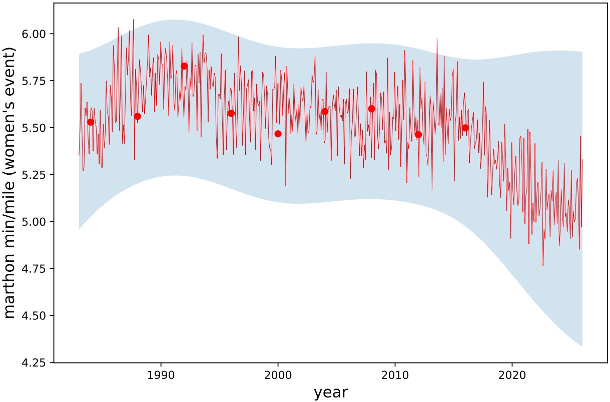
Let’s return to the notebook demo

(plots from the demo follow)

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: Number of topics in a topic model

The Dirichlet Process

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

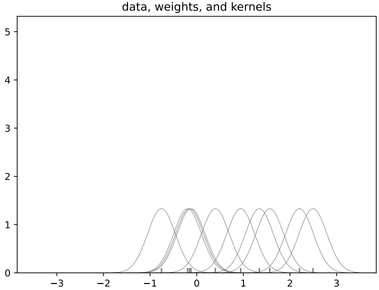
- Stick breaking
- Chinese restaurants

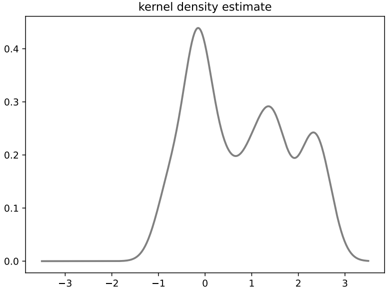
Recall: KDE

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

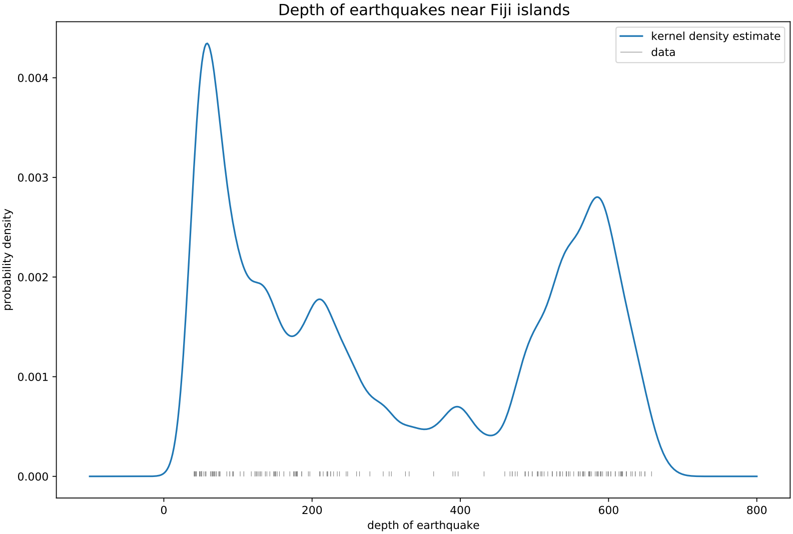
$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{X_i - x}{h} \right)$$

Recall: KDE


data, weights, and kernels


kernel density estimate

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

The “parameters” in the kernel density estimate are the data

We want to construct a *prior* distribution over densities before we see any data

Solution: Use synthetic or “imaginary” data!

Recall the Beta(α, β) prior — α and β can be viewed as imaginary heads and tails.

Dirichlet process

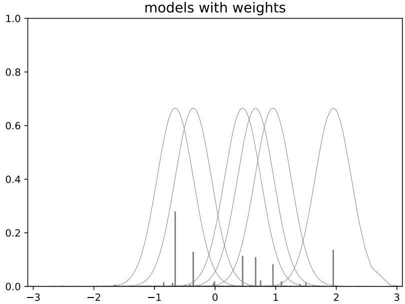
Each sample from a Dirichlet process prior has a *random collection of weights*, assigned to a *random selection of data*

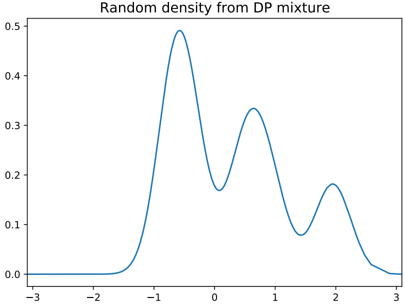
Each sample from Dirichlet process mixture has a random collection of weights assigned to a random selection of *model parameters*

Demo

Two volunteers?

Sample from DP mixture


models with weights


Random density from DP mixture

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 $\theta \sim F_0$

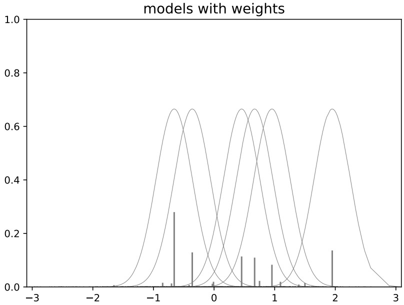
Stick breaking process for DPM

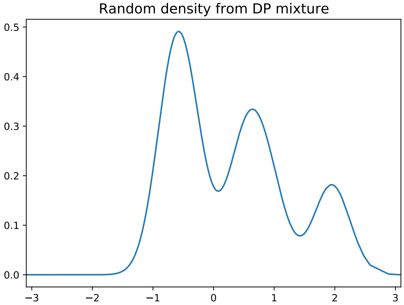
To draw a single random mixture from DPM(α, F_0):

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

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

Sample from DP mixture


models with weights


Random density from DP mixture

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 $x_1, x_2, \ldots, x_n$ 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

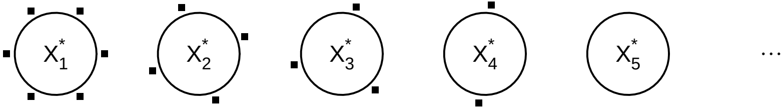
But this is not concrete

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

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

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

- 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 $m(x_1), m(x_2), \ldots, m(x_n)$ is jointly Gaussian
- Gaussian processes are Bayesian versions of kernel regression; the posterior mean is equivalent to Mercer kernel regression
- A Dirichlet process mixture is a Bayesian version of kernel density estimation
- Bayesian nonparametric methods require a lot of conceptual machinery and computation