S&DS 365 / 665

Intermediate Machine Learning

Nonparametric Bayes: Gaussian Processes

February 21

Yale

Reminders

- Assignment 1 due Wednesday
- Assignment 2 posted Wednesday
- Quiz 1 scores posted
- Quiz 2 on March 2
- Midterm on March 16 in class

For Today

- Review of parametric Bayes
- Introduction to nonparametric Bayes
- Gaussian processes
- 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* $p(x \mid \theta)$ 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* $p(\theta \mid \mathcal{D}_n)$.

Bayes' Theorem

A simple consequence of conditional probability:

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}\end{aligned}$$

Bayes' Theorem

The posterior distribution can be written as

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

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

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

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

Basic Example

Take model $X \sim \text{Bernoulli}(\theta)$.

This is a "coin flip": $X = 1$ means "heads" and $X = 0$ means "tails."

Natural prior is $\text{Beta}(\alpha, \beta)$ distribution

$$\pi_{\alpha,\beta}(\theta) \;\propto\; \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Basic Example

Take model $X \sim \text{Bernoulli}(\theta)$.

Natural prior is $\text{Beta}(\alpha, \beta)$ distribution

$$\pi_{\alpha,\beta}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The scaling constant is scary looking:

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

where $\Gamma(\cdot)$ is the "Gamma function"

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

https://en.wikipedia.org/wiki/Gamma_function

Basic Example

$X \sim \text{Bernoulli}(\theta)$ with data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$. Prior Beta$(\alpha, \beta)$ distribution

$$\pi_{\alpha,\beta}(\theta) \;\propto\; \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Let $s = \sum_{i=1}^n x_i$ be the number of "heads"

Posterior distribution $\theta \,|\, \mathcal{D}_n$ is another beta distribution!

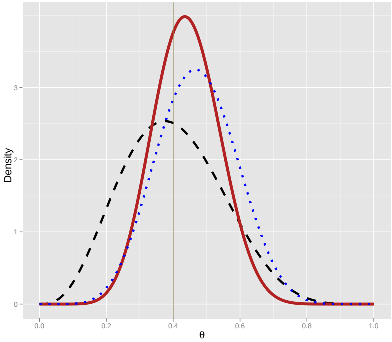
Specifically, with

$$\widetilde{\alpha} = \alpha + \text{number of heads} = \alpha + s$$
$$\widetilde{\beta} = \beta + \text{number of tails} = \beta + n - s$$

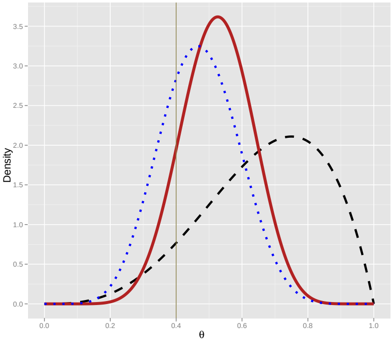
Showing this just uses the simple fact that $\theta^{\alpha-1}\theta^{x} = \theta^{x+\alpha-1}$

Example

$n = 15$ points sampled as $X \sim \text{Bernoulli}(\theta = 0.4)$, with $s = 7$ heads.



”good prior”



”bad prior”

Prior distribution (black-dashed), likelihood function (blue-dotted), posterior distribution (red-solid).

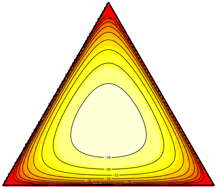
Dirichlet: From coins to dice

Multinomial model with Dirichlet prior is generalization of the Bernoulli/Beta model.

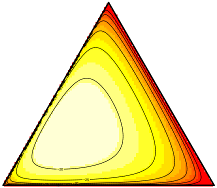
$$\text{Dirichlet}_\alpha(\theta) \;\propto\; \theta_1^{\alpha_1 - 1}\,\theta_2^{\alpha_2 - 1}\;\cdots\theta_K^{\alpha_K - 1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$ is a non-negative vector.

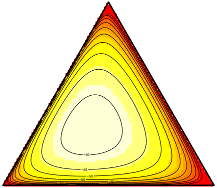
Example



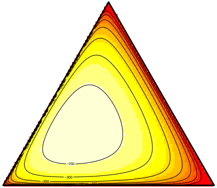
prior with Dirichlet(6,6,6)



likelihood function with $n = 20$



posterior distribution with $n = 20$



posterior distribution with $n = 200$

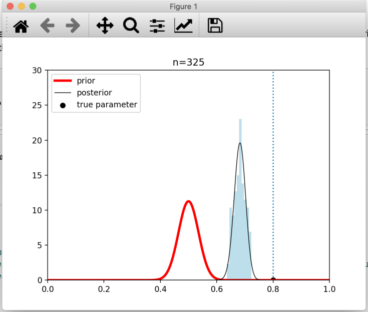
Parametric Bayes demo

Demo code for Bayesian analysis

In this notebook we illustrate some of the basic models and priors for Bayesian inference. These concepts will be important for our discussions about "topic models."

First, we illustrate the situation where the parameter θ that we are modeling is a Bernoulli parameter. This can be thought of as the probability that flipping a certain coin comes up heads. The most commonly used prior distribution for this model is a beta distribution. Under a beta prior, the posterior distribution is again a beta distribution.

This is illustrated in the following simula ior becomes concentrated on the true parameter. But as the variance of t




```
In [5]: import os, gzip
        import numpy as np
        import matplotlib.pyplot as p

In [*]: %matplotlib qt
        from scipy.special import gam
        from scipy import random
        from scipy.stats import beta

        theta = np.linspace(0,1,num=5
        fig = plt.figure(1)
        plt.ion()

        # The following are the param                                          the
        # variance of the prior decre                                         required for
        # the posterior to be centere

        scale = 100
        a0 = scale*1
        b0 = scale*1

        sample_size = 100
```

Nonparametric Bayes

- Bayesian inference for infinite-dimensional spaces
- Alternative to classical/frequentist approaches
- Caution required with interpretation

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?

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

Before diving in...

- Nonparametric Bayesian inference can be subtle and technical
- Part of the machine learning toolkit
- 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 two fundamental stochastic processes:

- Gaussian processes (today)
- Dirichlet processes (next time)

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

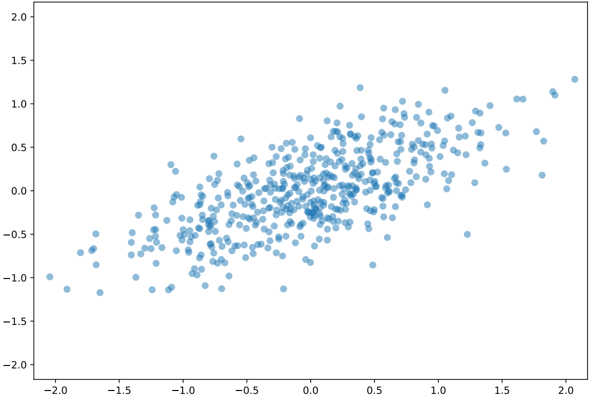
The frequentist kernel estimator for m is

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

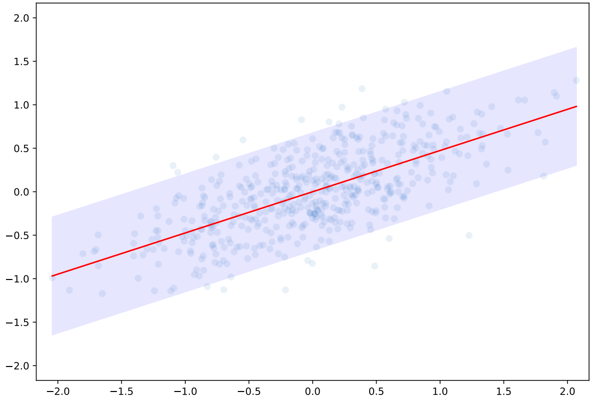
where K is a kernel and h is a bandwidth.

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

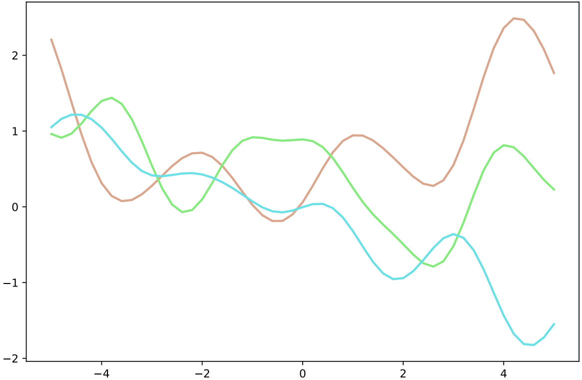
$$\mathbb{E}(\alpha \mid Y) = \left(\mathbb{K} + \sigma^2 I\right)^{-1} Y$$

and thus

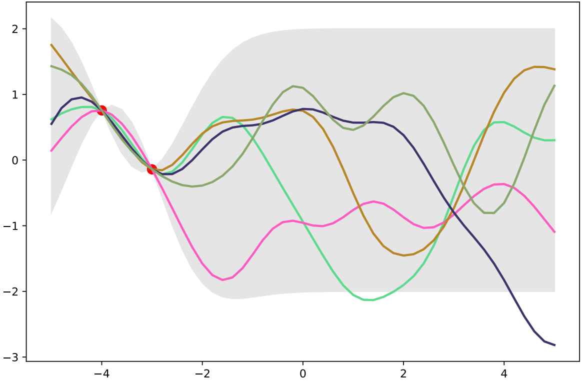
$$\widehat{m} = \mathbb{E}(m \mid Y) = \mathbb{K} \left(\mathbb{K} + \sigma^2 I\right)^{-1} Y.$$

Equivalent to Mercer kernel regression

Samples from prior and posterior



Samples from prior and posterior



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

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

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

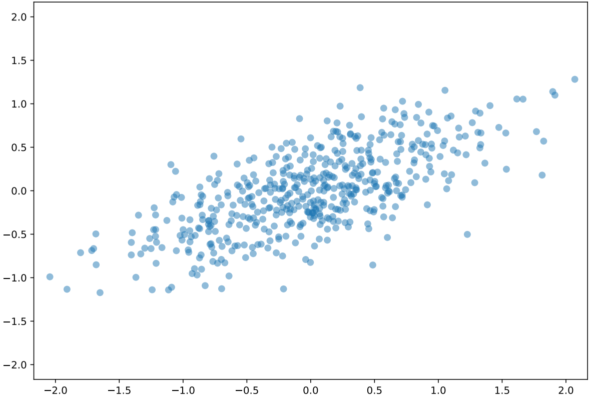
$$\begin{aligned}
\mathbb{E}(Y_{n+1} \mid x_{1:n}, Y_{1:n}) &= k^T(\mathbb{K} + \sigma^2 I)^{-1} Y \\
\mathrm{Var}(Y_{n+1} \mid x_{1:n}, Y_{1:n}) &= K(x_{n+1}, x_{n+1}) + \sigma^2 - k^T(\mathbb{K} + \sigma^2 I)^{-1} k
\end{aligned}$$

Predictive distribution

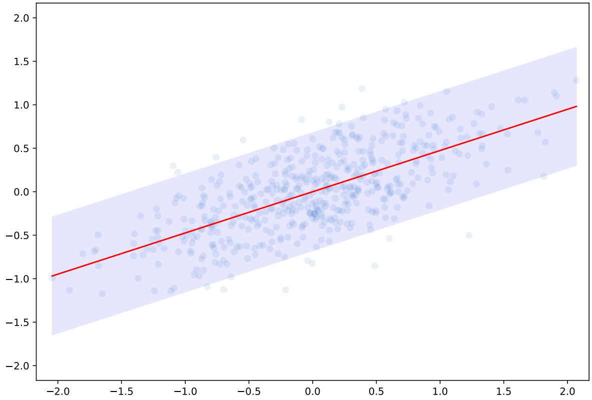
- Note that the mean is identical to what we saw for Mercer kernel regression
- But now we get a measure of uncertainty (the variance), which comes from the Gaussian process assumption

Let’s 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

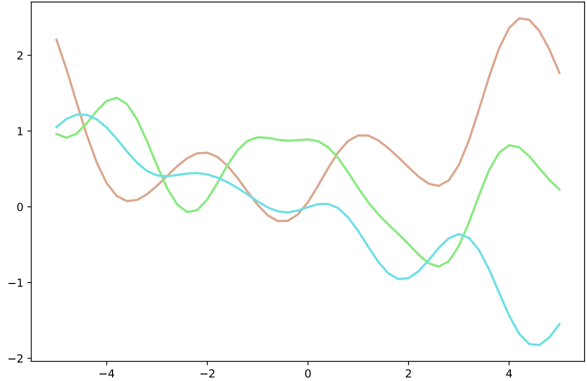
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}\right)$$

then the conditional distributions are also Gaussian and given by

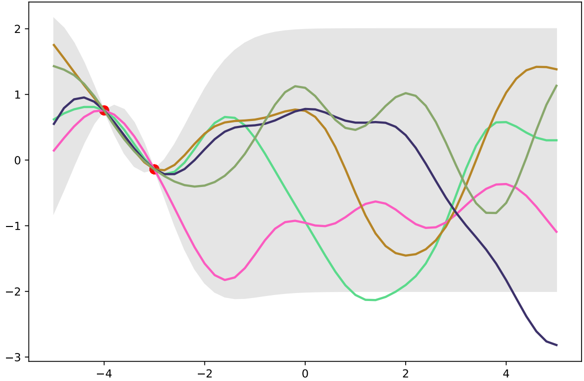
$$X_1 \mid x_2 \sim N\left(\frac{K_{12}}{K_{22}} x_2,\ K_{11} - \frac{K_{12}^2}{K_{22}}\right)$$

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

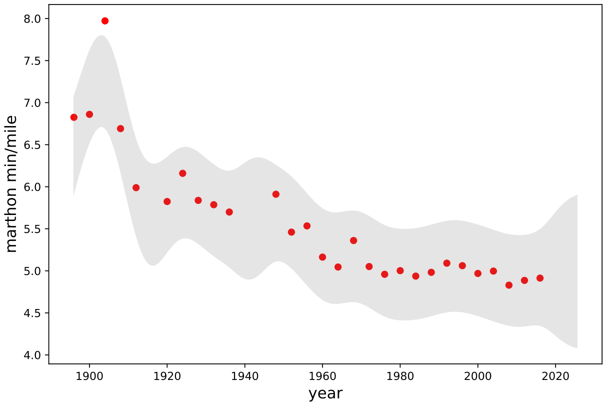
Samples from prior and posterior



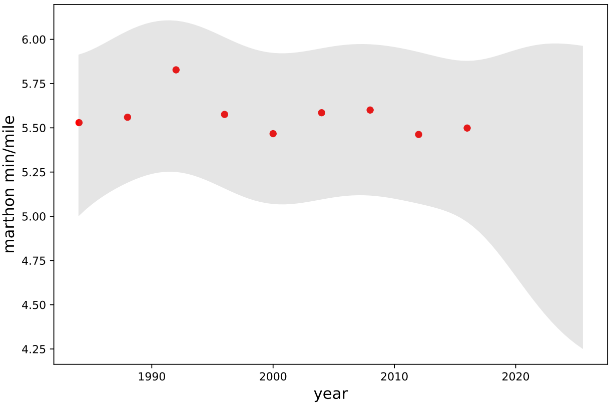
Samples from prior and posterior



Olympic marathon times (men's race)



Olympic marathon times (women's race)



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