CSci 8980: Advanced Topics in Graphical Models

Gaussian Processes

Instructor: Arindam Banerjee

November 15, 2007

Gaussian Processes

- Outline

Gaussian Processes

- Outline
  - Parametric Bayesian Regression

Gaussian Processes

- Outline
  - Parametric Bayesian Regression
  - Parameters to Functions

Gaussian Processes

- Outline
  - Parametric Bayesian Regression
  - Parameters to Functions
  - GP Regression

Gaussian Processes

- Outline
  - Parametric Bayesian Regression
  - Parameters to Functions
  - GP Regression
  - GP Classification

Gaussian Processes

- Outline
  - Parametric Bayesian Regression
  - Parameters to Functions
  - GP Regression
  - GP Classification
- We will use

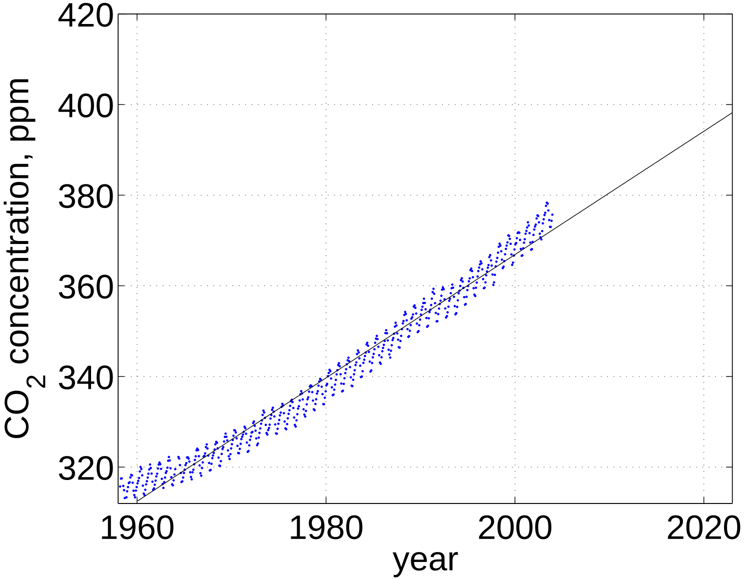
Gaussian Processes

- Outline
  - Parametric Bayesian Regression
  - Parameters to Functions
  - GP Regression
  - GP Classification
- We will use
  - Primary: Carl Rasmussen's GP tutorial slides (NIPS'06)

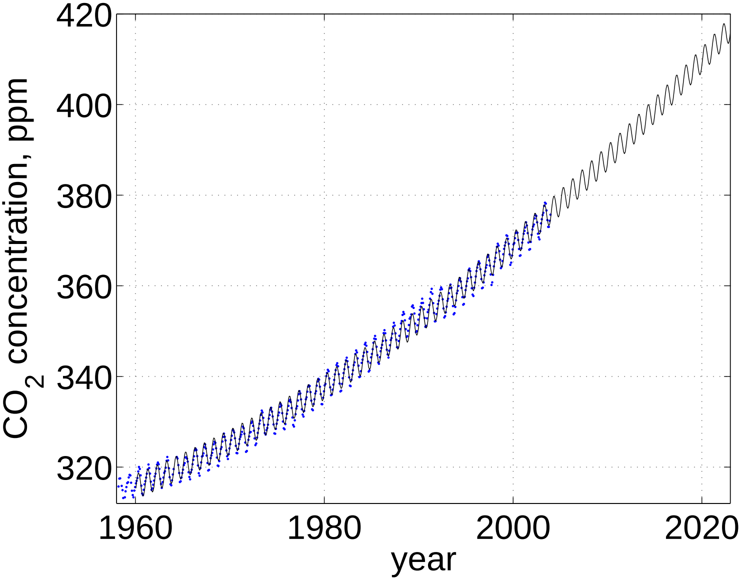
Gaussian Processes

- Outline
  - Parametric Bayesian Regression
  - Parameters to Functions
  - GP Regression
  - GP Classification
- We will use
  - Primary: Carl Rasmussen's GP tutorial slides (NIPS'06)
  - Secondary: Hanna Wallach's slides on regression

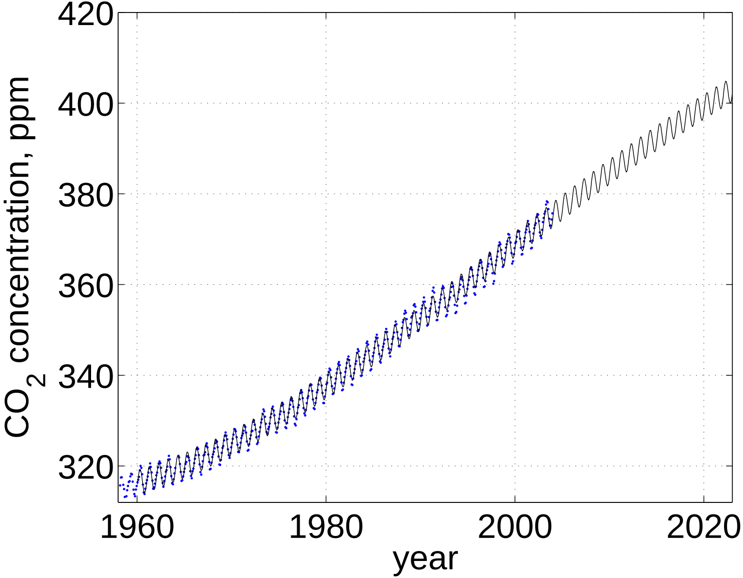
The Prediction Problem



The Prediction Problem



The Prediction Problem



Maximum likelihood, parametric model

Supervised parametric learning:

- data: $\mathbf{x}$, $\mathbf{y}$
- model: $y = f_{\mathbf{w}}(x) + \varepsilon$

Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M_i) \ \propto \ \prod_c \exp(-\tfrac{1}{2}(y_c - f_{\mathbf{w}}(x_c))^2/\sigma^2_{\text{noise}}).$$

Maximize the likelihood:

$$\mathbf{w}_{\text{ML}} \ = \ \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M_i).$$

Make predictions, by plugging in the ML estimate:

$$p(y^*|x^*, \mathbf{w}_{\text{ML}}, M_i)$$

Bayesian Inference, parametric model

Supervised parametric learning:

- data: $\mathbf{x}, \mathbf{y}$
- model: $y = f_{\mathbf{w}}(x) + \varepsilon$

Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M_i) \;\propto\; \prod_c \exp(-\tfrac{1}{2}(y_c - f_{\mathbf{w}}(x_c))^2/\sigma^2_{\text{noise}}).$$

Parameter prior:

$$p(\mathbf{w}|M_i)$$

Posterior parameter distribution by Bayes rule $p(a|b) = p(b|a)p(a)/p(b)$:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, M_i) \;=\; \frac{p(\mathbf{w}|M_i)p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M_i)}{p(\mathbf{y}|\mathbf{x}, M_i)}$$

Bayesian Inference, parametric model, cont.

Making predictions:

$$p(y^*|x^*, \mathbf{x}, \mathbf{y}, M_i) \;=\; \int p(y^*|\mathbf{w}, x^*, M_i)p(\mathbf{w}|\mathbf{x}, \mathbf{y}, M_i)d\mathbf{w}$$

Marginal likelihood:

$$p(\mathbf{y}|\mathbf{x}, M_i) \;=\; \int p(\mathbf{w}|M_i)p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M_i)d\mathbf{w}.$$

Model probability:

$$p(M_i|\mathbf{x}, \mathbf{y}) \;=\; \frac{p(M_i)p(\mathbf{y}|\mathbf{x}, M_i)}{p(\mathbf{y}|\mathbf{x})}$$

Problem: integrals are intractable for most interesting models!

Bayesian Linear Regression (2)

- Likelihood of parameters is:

$$P(\mathbf{y}|X, \mathbf{w}) = \mathcal{N}(X^\top \mathbf{w}, \sigma^2 I).$$

- Assume a Gaussian prior over parameters:

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \Sigma_p).$$

- Apply Bayes' theorem to obtain posterior:

$$P(\mathbf{w}|\mathbf{y}, X) \propto P(\mathbf{y}|X, \mathbf{w})P(\mathbf{w}).$$

Bayesian Linear Regression (3)

- Posterior distribution over **w** is:

$$P(\mathbf{w}|\mathbf{y}, X) = \mathcal{N}(\frac{1}{\sigma^2}A^{-1}X\mathbf{y}, A^{-1}) \text{ where } A = \Sigma_p^{-1} + \frac{1}{\sigma^2}XX^\top.$$

- Predictive distribution is:

$$\begin{aligned} P(f^\star|\mathbf{x}^\star, X, \mathbf{y}) &= \int f(\mathbf{x}^\star|\mathbf{w})P(\mathbf{w}|X, \mathbf{y})d\mathbf{w} \\ &= \mathcal{N}(\frac{1}{\sigma^2}\mathbf{x}^{\star\top}A^{-1}X\mathbf{y}, \mathbf{x}^{\star\top}A^{-1}\mathbf{x}^\star). \end{aligned}$$

Non-parametric Gaussian process models

In our non-parametric model, the "parameters" is the function itself!

Gaussian likelihood:

$$\mathbf{y}|\mathbf{x}, f(x), M_i \;\sim\; \mathcal{N}(\mathbf{f},\; \sigma^2_{\text{noise}} I)$$

(Zero mean) Gaussian process prior:

$$f(x)|M_i \;\sim\; \mathcal{GP}\big(m(x) \equiv 0,\; k(x, x')\big)$$

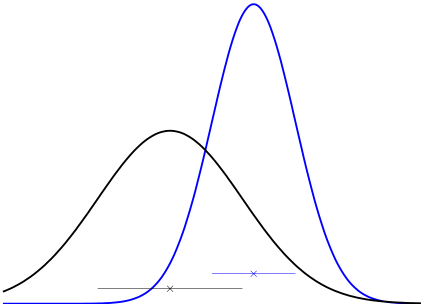
Leads to a Gaussian process posterior

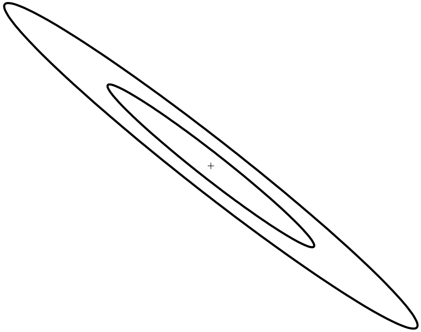
$$\begin{aligned} f(x)|\mathbf{x}, \mathbf{y}, M_i \;\sim\; \mathcal{GP}\big(&m_{\text{post}}(x) = k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y}, \\ &k_{\text{post}}(x, x') = k(x, x') - k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1} k(\mathbf{x}, x')\big). \end{aligned}$$

And a Gaussian predictive distribution:

$$\begin{aligned} y^*|x^*, \mathbf{x}, \mathbf{y}, M_i \;\sim\; \mathcal{N}\big(&\mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y}, \\ &k(x^*, x^*) + \sigma^2_{\text{noise}} - \mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{k}(x^*, \mathbf{x})\big) \end{aligned}$$

The Gaussian Distribution



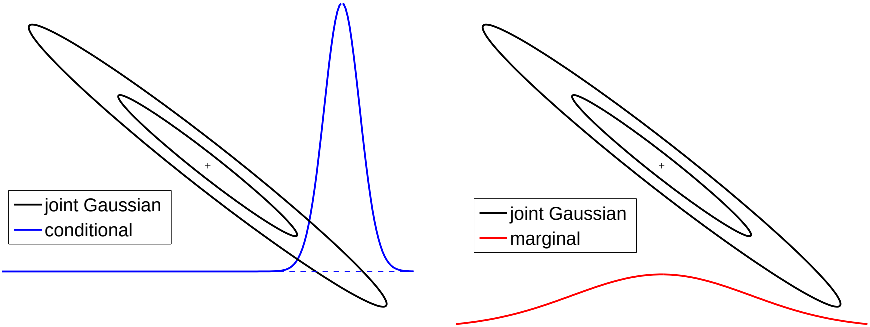


The Gaussian distribution is given by

$$p(\mathbf{x}|\mu, \Sigma) \;=\; \mathcal{N}(\mu, \Sigma) \;=\; (2\pi)^{-D/2}|\Sigma|^{-1/2} \exp\big(-\tfrac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\big)$$

where μ is the mean vector and Σ the covariance matrix.

Conditionals and Marginals of a Gaussian



Both the conditionals and the marginals of a joint Gaussian are again Gaussian.

What is a Gaussian Process?

A *Gaussian process* is a generalization of a multivariate Gaussian distribution to infinitely many variables.

Informally: infinitely long vector $\simeq$ function

> **Definition:** *a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.* □

A Gaussian distribution is fully specified by a mean vector, μ, and covariance matrix Σ:

$$\mathbf{f} = (f_1, \ldots, f_n)^\top \sim \mathcal{N}(\mu, \Sigma), \quad \text{indexes } i = 1, \ldots, n$$

A Gaussian process is fully specified by a mean function $m(x)$ and covariance function $k(x, x')$:

$$f(x) \sim \mathcal{GP}\big(m(x), k(x, x')\big), \quad \text{indexes: } x$$

The marginalization property

Thinking of a GP as a Gaussian distribution with an infinitely long mean vector and an infinite by infinite covariance matrix may seem impractical…

…luckily we are saved by the *marginalization property*:

Recall:

$$p(\mathbf{x}) \;=\; \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

For Gaussians:

$$p(\mathbf{x}, \mathbf{y}) \;=\; \mathcal{N}\Big(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}\Big) \;\Longrightarrow\; p(\mathbf{x}) \;=\; \mathcal{N}(\mathbf{a},\; A)$$

Random functions from a Gaussian Process

Example one dimensional Gaussian process:

$$p(f(x)) \sim \mathcal{GP}\big(m(x) = 0,\ k(x, x') = \exp(-\tfrac{1}{2}(x - x')^2)\big).$$

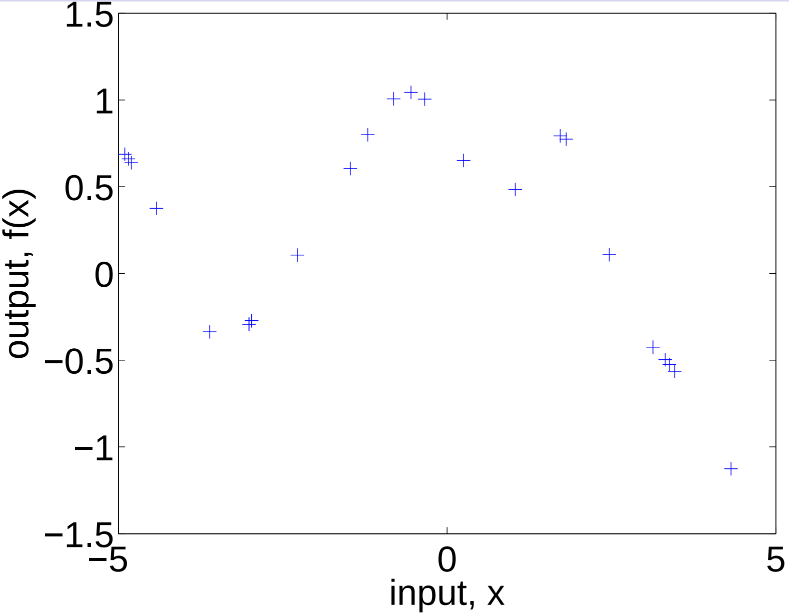
To get an indication of what this distribution over functions looks like, focus on a finite subset of function values $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))^\top$, for which

$$\mathbf{f} \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma_{ij} = k(x_i, x_j)$.

Then plot the coordinates of f as a function of the corresponding x values.

Some values of the random function



Non-parametric Gaussian process models

In our non-parametric model, the "parameters" is the function itself!

Gaussian likelihood:

$$\mathbf{y}|\mathbf{x}, f(x), M_i \;\sim\; \mathcal{N}(\mathbf{f},\; \sigma^2_{\text{noise}} I)$$

(Zero mean) Gaussian process prior:

$$f(x)|M_i \;\sim\; \mathcal{GP}\big(m(x) \equiv 0,\; k(x, x')\big)$$

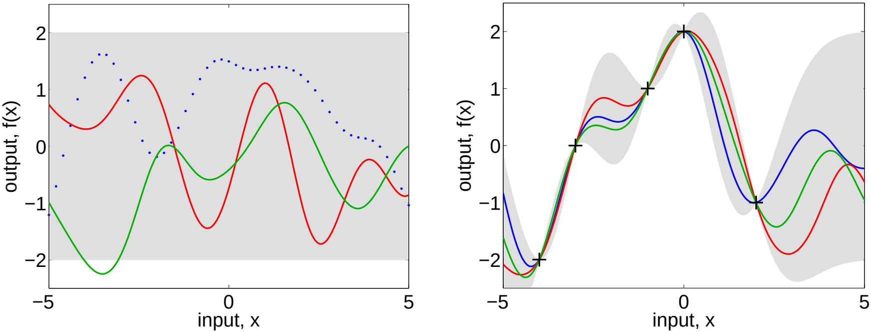
Leads to a Gaussian process posterior

$$\begin{aligned} f(x)|\mathbf{x}, \mathbf{y}, M_i \;\sim\; \mathcal{GP}\big(&m_{\text{post}}(x) = k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y}, \\ &k_{\text{post}}(x, x') = k(x, x') - k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1} k(\mathbf{x}, x')\big). \end{aligned}$$

And a Gaussian predictive distribution:

$$\begin{aligned} y^*|x^*, \mathbf{x}, \mathbf{y}, M_i \;\sim\; \mathcal{N}\big(&\mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y}, \\ &k(x^*, x^*) + \sigma^2_{\text{noise}} - \mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{k}(x^*, \mathbf{x})\big) \end{aligned}$$

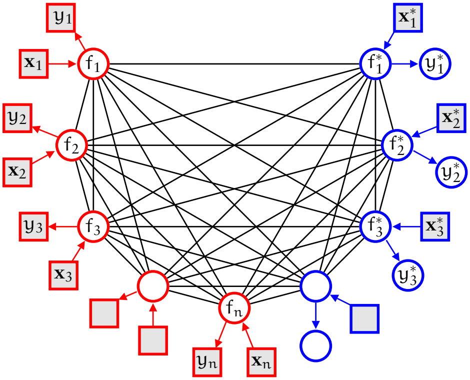
Prior and Posterior



Predictive distribution:

$$p(y^*|x^*, \mathbf{x}, \mathbf{y}) \;\sim\; \mathcal{N}\big(\mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{y},$$
$$k(x^*, x^*) + \sigma^2_{\text{noise}} - \mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{k}(x^*, \mathbf{x})\big)$$

Graphical model for Gaussian Process



Square nodes are observed (clamped), round nodes stochastic (free).

All pairs of latent variables are connected.

Predictions y^* depend only on the corresponding single latent f^*.

Notice, that adding a triplet x_m^*, f_m^*, y_m^* does not influence the distribution. This is guaranteed by the marginalization property of the GP.

This explains why we can make inference using a finite amount of computation!

Some interpretation

Recall our main result:

$$\mathbf{f}_*|X_*, X, \mathbf{y} \;\sim\; \mathcal{N}\big(K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y},\\ K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*)\big).$$

The mean is linear in two ways:

$$\mu(\mathbf{x}_*) \;=\; k(\mathbf{x}_*, X)[K(X, X) + \sigma_n^2]^{-1}\mathbf{y} \;=\; \sum_{c=1}^{n} \beta_c y^{(c)} \;=\; \sum_{c=1}^{n} \alpha_c k(\mathbf{x}_*, \mathbf{x}^{(c)}).$$

The last form is most commonly encountered in the kernel literature.

The variance is the difference between two terms:

$$V(\mathbf{x}_*) \;=\; k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{k}(X, \mathbf{x}_*),$$

the first term is the *prior variance*, from which we subtract a (positive) term, telling how much the data X has explained. Note, that the variance is independent of the observed outputs $\mathbf{y}$.

The marginal likelihood

Log marginal likelihood:

$$\log p(\mathbf{y}|\mathbf{x}, M_i) \; = \; -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi)$$

is the combination of a data fit term and complexity penalty. Occam's Razor is automatic.

Learning in Gaussian process models involves finding

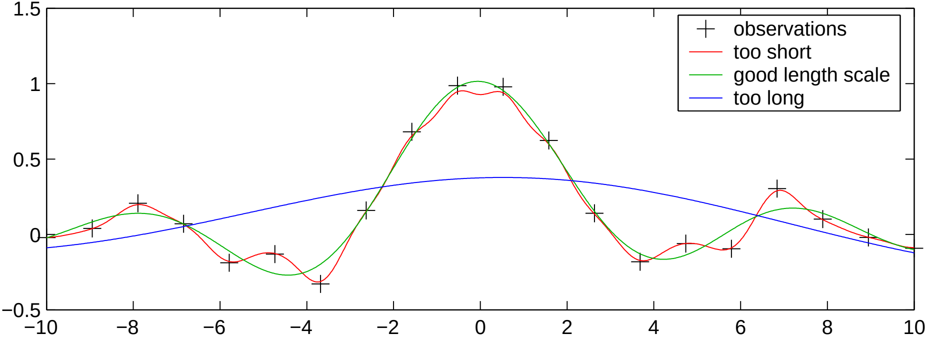
- the form of the covariance function, and
- any unknown (hyper-) parameters θ.

This can be done by optimizing the marginal likelihood:

$$\frac{\partial \log p(\mathbf{y}|\mathbf{x}, \theta, M_i)}{\partial \theta_j} \; = \; \frac{1}{2}\mathbf{y}^\top K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\mathbf{y} - \frac{1}{2}\,\mathrm{trace}(K^{-1}\frac{\partial K}{\partial \theta_j})$$

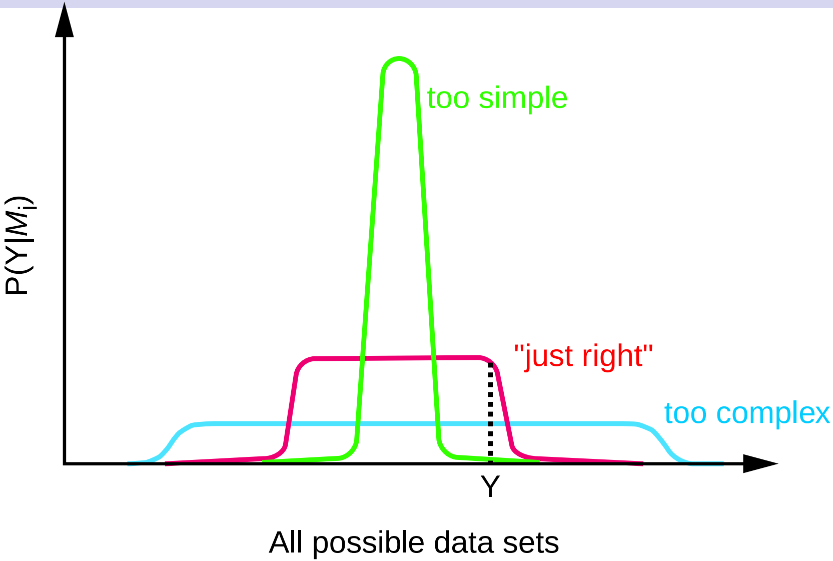
Example: Fitting the length scale parameter

Parameterized covariance function: $k(x, x') = v^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right) + \sigma_n^2 \delta_{xx'}$.



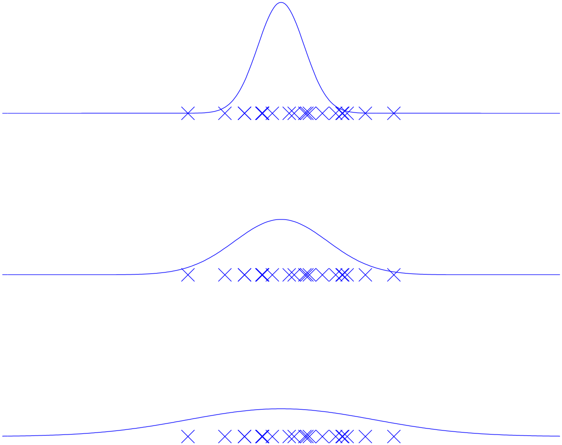
The mean posterior predictive function is plotted for 3 different length scales (the green curve corresponds to optimizing the marginal likelihood). Notice, that an almost exact fit to the data can be achieved by reducing the length scale – but the marginal likelihood does not favour this!

Why, in principle, does Bayesian Inference work?
Occam's Razor



An illustrative analogous example

Imagine the simple task of fitting the variance, σ^2, of a zero-mean Gaussian to a set of n scalar observations.



The log likelihood is $\log p(\mathbf{y}|\mu, \sigma^2) = -\frac{1}{2}\sum(y_i - \mu)^2/\sigma^2 - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi)$

From random functions to covariance functions

Consider the class of linear functions:

$$f(x) \;=\; ax + b, \quad \text{where} \;\; a \sim \mathcal{N}(0, \alpha), \quad \text{and} \;\; b \sim \mathcal{N}(0, \beta).$$

We can compute the mean function:

$$\mu(x) \;=\; E[f(x)] \;=\; \iint f(x)p(a)p(b)dadb \;=\; \int axp(a)da + \int bp(b)db = 0,$$

and covariance function:

$$\begin{aligned} k(x, x') \; &= E[(f(x) - 0)(f(x') - 0)] \;=\; \iint (ax + b)(ax' + b)p(a)p(b)dadb \\ &= \int a^2xx'p(a)da + \int b^2p(b)db + (x + x')\int abp(a)p(b)dadb = \alpha xx' + \beta. \end{aligned}$$

From random functions to covariance functions II

Consider the class of functions (sums of squared exponentials):

$$
\begin{aligned}
f(x) &= \lim_{n\to\infty} \frac{1}{n} \sum_i \gamma_i \exp(-(x-i/n)^2), \quad \text{where } \gamma_i \sim \mathcal{N}(0,1),\ \forall i \\
&= \int_{-\infty}^{\infty} \gamma(u) \exp(-(x-u)^2) du, \quad \text{where } \gamma(u) \sim \mathcal{N}(0,1),\ \forall u.
\end{aligned}
$$

The mean function is:

$$
\mu(x) = E[f(x)] = \int_{-\infty}^{\infty} \exp(-(x-u)^2) \int_{-\infty}^{\infty} \gamma p(\gamma) d\gamma du = 0,
$$

and the covariance function:

$$
\begin{aligned}
E[f(x)f(x')] &= \int \exp\big(-(x-u)^2 - (x'-u)^2\big) du \\
&= \int \exp\big(-2(u - \frac{x+x'}{2})^2 + \frac{(x+x')^2}{2} - x^2 - x'^2\big)\big) du \ \propto\ \exp\big(-\frac{(x-x')^2}{2}\big).
\end{aligned}
$$

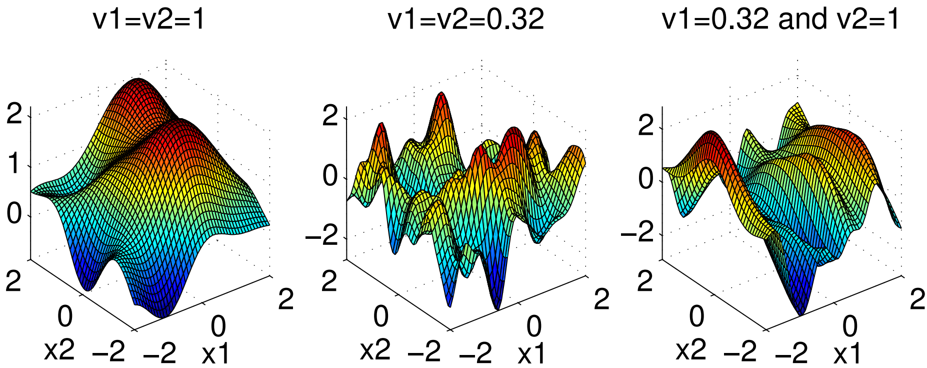
Thus, the squared exponential covariance function is equivalent to regression using infinitely many Gaussian shaped basis functions placed everywhere, not just at your training points!

Model Selection in Practise; Hyperparameters

There are two types of task: *form* and *parameters* of the covariance function.

Typically, our prior is too weak to quantify aspects of the covariance function. We use a hierarchical model using hyperparameters. Eg, in ARD:

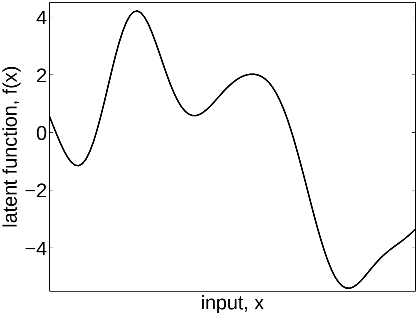
$$k(\mathbf{x}, \mathbf{x}') = v_0^2 \exp\Big(-\sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{2v_d^2}\Big), \qquad \text{hyperparameters } \theta = (v_0, v_1, \ldots, v_d, \sigma_n^2).$$

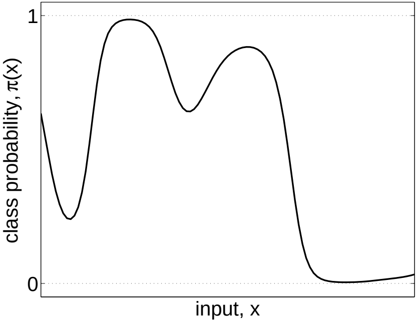

v1=v2=1

v1=v2=0.32

v1=0.32 and v2=1

Binary Gaussian Process Classification





The class probability is related to the *latent* function, f, through:

$$p(y = 1|f(\mathbf{x})) \;=\; \pi(\mathbf{x}) \;=\; \Phi\big(f(\mathbf{x})\big),$$

where Φ is a sigmoid function, such as the logistic or cumulative Gaussian. Observations are independent given f, so the likelihood is

$$p(\mathbf{y}|\mathbf{f}) \;=\; \prod_{i=1}^{n} p(y_i|f_i) \;=\; \prod_{i=1}^{n} \Phi(y_i f_i).$$

Prior and Posterior for Classification

We use a Gaussian process prior for the latent function:

$$\mathbf{f}|X, \theta \sim \mathcal{N}(\mathbf{0}, K)$$

The posterior becomes:

$$p(\mathbf{f}|\mathcal{D}, \theta) = \frac{p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f}|X, \theta)}{p(\mathcal{D}|\theta)} = \frac{\mathcal{N}(\mathbf{f}|\mathbf{0}, K)}{p(\mathcal{D}|\theta)} \prod_{i=1}^{m} \Phi(y_i f_i),$$

which is non-Gaussian.

The latent value at the test point, $f(\mathbf{x}^*)$ is

$$p(f_*|\mathcal{D}, \theta, \mathbf{x}_*) = \int p(f_*|\mathbf{f}, X, \theta, \mathbf{x}_*) p(\mathbf{f}|\mathcal{D}, \theta) d\mathbf{f},$$

and the predictive class probability becomes

$$p(y_*|\mathcal{D}, \theta, \mathbf{x}_*) = \int p(y_*|f_*) p(f_*|\mathcal{D}, \theta, \mathbf{x}_*) df_*,$$

both of which are intractable to compute.

Gaussian Approximation to the Posterior

We approximate the non-Gaussian posterior by a Gaussian:

$$p(\mathbf{f}|\mathcal{D}, \theta) \;\simeq\; q(\mathbf{f}|\mathcal{D}, \theta) \;=\; \mathcal{N}(\mathbf{m},\; A)$$

then $q(f_*|\mathcal{D}, \theta, \mathbf{x}_*) = \mathcal{N}(f_*|\mu_*, \sigma_*^2)$, where

$$\begin{aligned} \mu_* \;&=\; \mathbf{k}_*^\top K^{-1}\mathbf{m} \\ \sigma_*^2 \;&=\; k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K^{-1} - K^{-1}AK^{-1})\mathbf{k}_*. \end{aligned}$$

Using this approximation with the cumulative Gaussian likelihood

$$q(y_* = 1|\mathcal{D}, \theta, \mathbf{x}_*) \;=\; \int \Phi(f_*)\, \mathcal{N}(f_*|\mu_*, \sigma_*^2)df_* \;=\; \Phi\Big(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\Big)$$

Laplace's method and Expectation Propagation

How do we find a good Gaussian approximation $\mathcal{N}(\mathbf{m}, A)$ to the posterior?

Laplace's method: Find the Maximum A Posteriori (MAP) lantent values $\mathbf{f}_{\text{MAP}}$, and use a local expansion (Gaussian) around this point as suggested by Williams and Barber [10].

Variational bounds: bound the likelihood by some tractable expression
A **local variational bound for each likelihood term** was given by Gibbs and MacKay [1]. A **lower bound based on Jensen's inequality** by Opper and Seeger [7].

Expectation Propagation: use an approximation of the likelihood, such that the moments of the marginals of the approximate posterior match the (approximate) moment of the posterior, Minka [6].

Laplace's method and EP were compared by Kuss and Rasmussen [3].

Conclusions

Complex non-linear inference problems can be solved by manipulating plain old Gaussian distributions

- Bayesian inference is tractable for GP regression and
- Approximations exist for classification
- predictions are probabilistic
- compare different models (via the marginal likelihood)

GPs are a simple and intuitive means of specifying prior information, and explaining data, and equivalent to other models: RVM's, splines, closely related to SVMs.

Outlook:

- new interesting covariance functions
- application to structured data
- better understanding of sparse methods