Bayesian Kernel Methods

THE AUSTRALIAN NATIONAL UNIVERSITY

Unit 1: Bayes Rule, Approximate Inference, Hyperparameters

Unit 2: Gaussian Processes, Covariance Function, Kernel

Unit 3: GP: Regression

Unit 4: GP: Classification

Unit 5: Implementation: Laplace Approximation, Low Rank Methods

Unit 6: Implementation: Low Rank Methods, Bayes Committee Machine

Unit 7: Relevance Vector Machine: Priors on Coefficients

Unit 8: Relevance Vector Machine: Efficient Optimization and Extensions

http://mlg.anu.edu.au/~smola/summer2002/

Overview of Unit 7: Relvance Vector Machine

THE AUSTRALIAN NATIONAL UNIVERSITY

01: Data-Dependent Priors
02: Applications
03: Recall: Coefficient Priors
04: Example: Neurons
05: Example: Independent Sources
06: Example: Kernel Expansions
07: Convergence to Gaussian Processes
08: Proof
09: Posterior
10: The RVM Idea
11: Example: Gamma-Hyperprior
12: Example: Normal-Hyperprior
13: General Hyperpriors
14: More Examples
15: Practical Problem: Inference

Data-Dependent Priors

Problem

We are wasting information if we ignore the training patterns in specifying our prior.

Solution: Revisiting Bayes' Rule

$$P(f|X,Y) = \frac{P(Y|f,X)P(f|X)}{P(Y|X)}$$

This means that we already have a **data dependent prior**. The problem with data-independence only arose from the standard approximation $p(f|X) = p(f)$.

Note: the same connection applies to densities.

Consequence

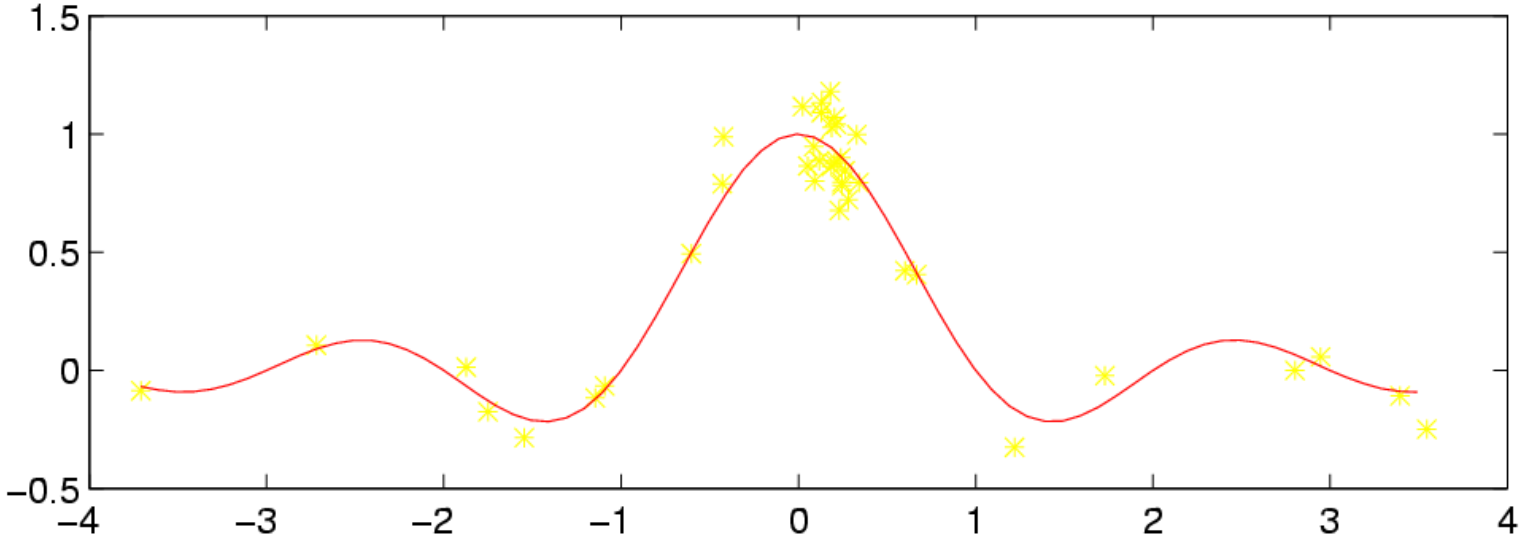
We need to find suitable data-dependent priors which correspond to useful priors over function spaces. If we know $p(X)$, we obviously have

$$p(f) = \int p(f|X)p(X)dX.$$

Examples

Density Dependent Capacity

We can allow for a higher complexity function where we have a large amount of data.



Different Regimes

Data might come from N different sources, which can be distinguished solely based on $x_1, \ldots, x_m$. So, depending on which source, we will switch between priors $p_1(f), \ldots, p_N(f)$.

Recall: Coefficient Priors

Function Expansion

Assume that f can be expanded into a linear model of type

$$f(x) = \sum_{i=1}^{M} \alpha_i f_i(x)$$

where $\{f_i(x)\}$ is a suitable set of functions. This could, e.g., be a kernel, i.e., $m = M$ and $f_i = k(x_i, x)$. Note: k **is arbitrary**, e.g., we do not require positivity.

Factorizing Priors

Analogously to a factorizing assumption on the observations we may also assume

$$p(f) = \prod_{i=1}^{m} p(\alpha_i) \text{ where } f = \sum_{i=1}^{m} \alpha_i f_i$$

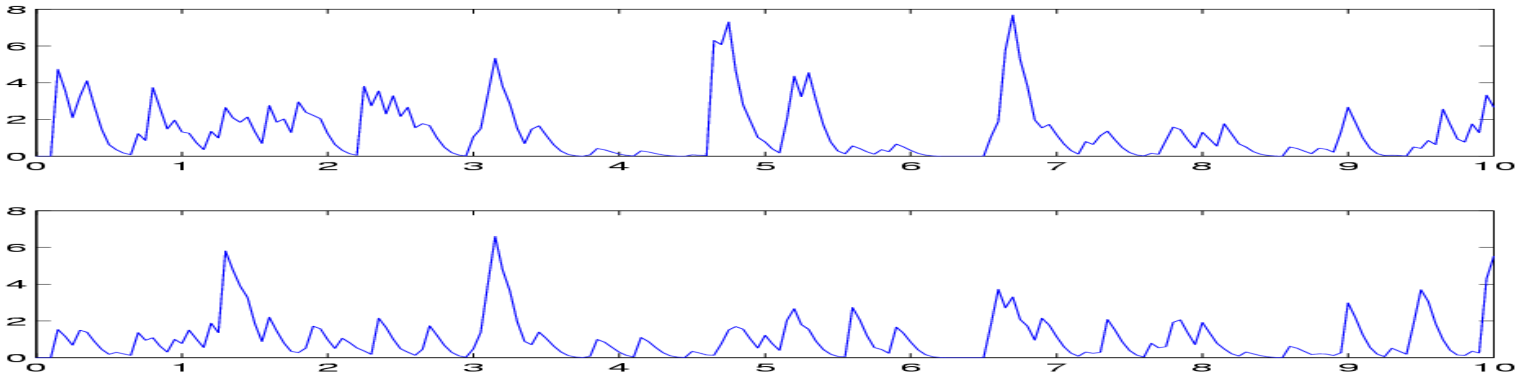
Motivation

The basis functions f_i correspond to independent "factors" causing the observations, e.g., neurons firing independently but rarely, image elements occurring, etc.

Examples

Brain Signals (gross oversimplification)

Neurons fire independently, and very rarely, however, we only observe the signal from several neurons at the same time, possibly several observations with different linear combinations thereof.



Cocktail Party Problem

Assume many speakers, talking (not necessarily to each other) independently. We have many microphones, what is the signal we receive on each microphone? What were the underlying signals?

Example: Kernel Expansions

THE AUSTRALIAN NATIONAL UNIVERSITY

Expansion

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) \text{ and } p(f) = \prod_{i=1}^{m} p(\alpha_i)$$

Rationale

- Convenient way of specifying data-dependent prior
- Increases capacity automatically where much data occurs
- Easy to optimize
- Easy to explain (linear model)
- Nice theoretical properties

Examples

$p(\alpha) \propto \exp(-|\alpha|^p)$, $p(\alpha) = \text{BesselK}(0, |\alpha|)$, $p(\alpha) = \frac{1}{s_i}\alpha^{-s}$, ...

Convergence to Gaussian Processes

Theorem

- Denote by α_i independent random variables (we do not require identical distributions on α_i) with **unit variance and zero mean**.
- Assume that there exists a **distribution** $p(x)$ **on** $\mathcal{X}$ according to which a sample $\{x_1, \ldots, x_m\}$ is drawn.
- Assume that $k(x, x')$ is **bounded** on $\mathcal{X} \times \mathcal{X}$.

Then the random variable $y(x)$ given by

$$y(x) = \frac{1}{m} \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

converges for $m \to \infty$ to a Gaussian process with zero mean and covariance function

$$\tilde{k}(x, x') = \int_{\mathcal{X}} k(x, \bar{x}) k(x', \bar{x}) p(\bar{x}) d\bar{x}.$$

Proof

Normal Distribution of Linear Combinations

We need only check is that $y(x)$ and any linear combination $\sum_j y(x_j)$ (for arbitrary $x'_j \in \mathcal{X}$) converge to a normal distribution. By application of a theorem of Cramér, this is sufficient to prove that $y(x)$ is distributed according to a Gaussian Process.

Computing $y(x)$

The random variable $y(x)$ is a sum of m independent random variables with bounded variance (since $k(x, x')$ is bounded on $\mathcal{X} \times \mathcal{X}$). Therefore in the limit $m \to \infty$, by virtue of the Central Limit Theorem, we have

$$y(x) \sim \mathcal{N}(0, \sigma^2(x)) \text{ for some } \sigma^2(x) \in \mathbb{R}$$

Linear Combinations For arbitrary $x'_j \in \mathcal{X}$, linear combinations of $y(x'_j)$ also have Gaussian distributions since

$$\sum_{j=1}^{n} \beta_i y(x'_j) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{n} \beta_i k(x_i, x'_j).$$

Proof, Part II

Central Limit Theorem on Linear Combination

We may apply the Central Limit Theorem to the sum since the inner sum $\sum_{j=1}^{n} \beta_i k(x_i, x'_j)$ is bounded for any x_i. This also implies that $\sum_{j=1}^{n} \beta_j y(x'_j) \sim \mathcal{N}(0, \sigma^2)$ for $m \to \infty$ and some $\sigma^2 \in \mathbb{R}$, which proves that $y(x)$ is distributed according to a GP.

Computing an equivalent Gaussian Process

Note that $y(x)$ has zero mean. Thus the covariance function for finite m can be found as expectation with respect to the random variables α_i,

$$E[y(x)y(x')] = E\left[\frac{1}{m}\sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x) k(x_j, x')\right] = \frac{1}{m}\sum_{i=1}^{m} k(x_i, x) k(x_j, x'),$$

since the α_i are independent and have zero mean. This converges to the Riemann integral over $\mathcal{X}$ with the density $p(x)$ as $m \to \infty$. Thus

$$E[y(x)y(x')] \underset{m\to\infty}{\longrightarrow} \int_{\mathcal{X}} k(x, \bar{x}) k(x', \bar{x}) p(\bar{x}) d\bar{x}.$$

Effective Kernels

Example: Linear Kernel

$k(x, x') = \langle x, x' \rangle$ and coefficient-based prior. Here we have

$$\tilde{k}(x, x') = \int_X k(x, \bar{x}) k(x', \bar{x}) p(\bar{x}) d\bar{x} = x^\top \left(\int \bar{x}\bar{x}^\top p(\bar{x}) d\bar{x} \right) x' = x^\top \left(\mathrm{Cov}[x] \right) x'.$$

Example: Gaussian Kernel

For a kernel $k(x, x') = \exp(-\frac{1}{2}\|x - x'\|^2)$ and $p(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$ we obtain for $\tilde{k}$

$$\tilde{k}(x, x') = \frac{1}{\sqrt{5}} \exp\left(-\frac{3}{5}(x - x')^2 \right) \exp\left(-\frac{2}{5}\langle x, x' \rangle \right)$$

Note

The specific form of $p(\alpha_i)$ is irrelevant for $\tilde{k}$, as long as the variance is bounded (of course, this holds only in the limit).

Consequence

We can look for priors which allow for many zero coefficients α_i.

The RVM Idea

Posterior

For a kernel expansion, the posterior can be found as

$$p(\alpha|X,Y) \propto \prod_{i=1}^{m} p(y_i|f(x_i))p(\alpha_i) \text{ where } f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x).$$

Problem

For rather arbitrary priors, this is a difficult optimization problem. We would rather like to have a Gaussian prior ...

Idea

Rewrite $p(\alpha)$ **as** $\int p(\alpha|s)p(s)ds$**, i.e., by means of a Hyperparameterwhere** $p(\alpha|s)$ **is Gaussian** (and optimize via MAP2).

Result

$$p(\alpha|X,Y) = \int_{\mathbb{R}^m} \prod_{i=1}^{m} p(y_i|f(x_i))p(\alpha_i|s_i)ds_1 \dots ds_m$$

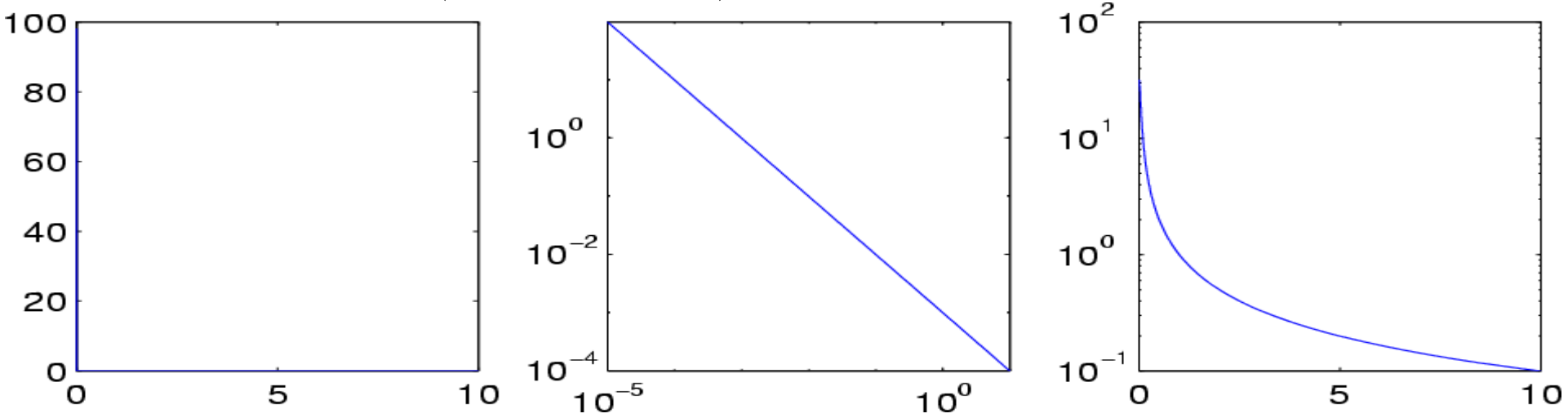
Example: Gamma-Hyperprior

THE AUSTRALIAN NATIONAL UNIVERSITY

Gamma Distribution

$$p(s) = \Gamma(s|a,b) := \frac{s^{a-1}b^a \exp(-sb)}{\Gamma(a)} \text{ for } s_i > 0.$$

For non-informative (flat in logspace) priors, one typically chooses $a = b = 10^{-4}$.



Effective Prior

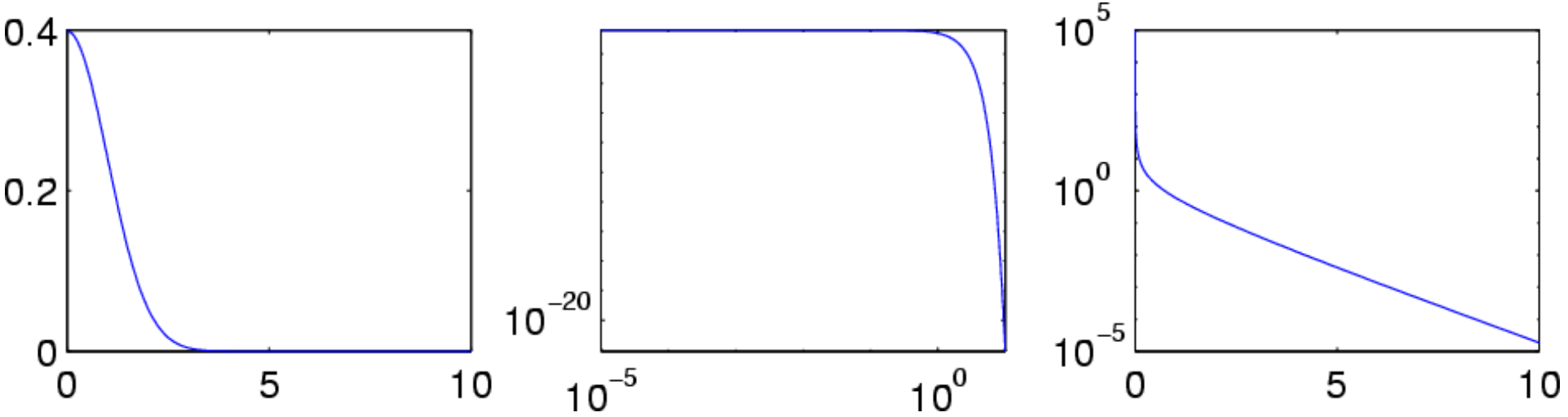
For the normal prior $p(\alpha|s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2s^2}\alpha^2\right)$ we have

$$p(\alpha) = \int \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2s^2}\alpha^2\right) \frac{s_i^{a-1}b^a \exp(-s_i b)}{\Gamma(a)} ds = \exp\left(-(a+1/2)\ln\left(b+\frac{\alpha^2}{2}\right)\right.$$

Example: Normal-Hyperprior

Normal Distribution

$$p(s) = \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2\omega^2}s^2\right)$$



Effective Prior

For the normal prior $p(\alpha|s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2s^2}\alpha^2\right)$ we have

$$p(\alpha) = \int \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2s^2}\alpha^2\right) \frac{1}{\sqrt{2\pi\omega^2}} \exp\left(-\frac{1}{2\omega^2}s^2\right) ds \propto \text{BesselK}(0, |\omega|).$$

General Hyperpriors

Problem

How can we find a suitable hyperprior $p(s)$ for a given $p(\alpha)$ such that

$$p(\alpha) = \int p(\alpha|s)p(s)ds = \int \frac{1}{\sqrt{2\pi s^2}} \exp\left(\frac{1}{2s^2}\alpha^2\right) p(s)ds$$

Solution (after Girosi, 1991)

Parameter transformation $\beta = \frac{1}{2\omega^2}$ leads to

$$p(\alpha) = \int \exp\left(-\beta\alpha\right) \left[\frac{1}{\sqrt{8\pi\beta}} p\left(\frac{1}{\sqrt{2\beta}}\right)\right] d\beta$$

That is, $p(\alpha)$ is the **Laplace Transform** of $\left[\frac{1}{\sqrt{8\pi\beta}} p\left(\frac{1}{\sqrt{2\beta}}\right)\right]$.

Strategy

Given $p(\alpha)$ we only need to find its **inverse** Laplace Transform $\mathcal{L}^{-1}p$ to obtain $p(s)$.

More Examples

Polynomial Priors

For $p(\alpha) = \exp(-|\alpha|^{-a})$ for $a > 1$ we have

$$\left[\mathcal{L}^{-1}p\right](s) = \frac{s^{a-1}}{\Gamma(a)} \text{ hence } p(s) = \sqrt{2\pi}\,\frac{2^{1-a}}{\Gamma(a)}\omega^{-2a}$$

Consequence

- We can deal quite conveniently with priors which do not lead to a lower-bounded optimization problem.
- Large a leads to priors highly peaked at 0 (hence a very sparse code).
- For $a > 1.5$ the variance of s is bounded, hence we get a limiting Gaussian Process.
- For more examples see Bronstein & Semendjajev, Abramovitz & Stegun, etc.

Practical Problem: Inference

THE AUSTRALIAN NATIONAL UNIVERSITY

MAP2 Approximation

Instead of computing the integral over m hyperparameters, we approximate by maximizing

$$p(Y|X,Y,s) \propto \left(\int_{\alpha} \prod_{i=1}^{m} \underbrace{p(y_i|f(x_i))}_{\text{Likelihood}} \underbrace{p(\alpha_i|s_i)}_{\text{Prior}} d\alpha \right) \underbrace{p(s_i)}_{\text{Hyperprior}}$$

Fixed Point Iteration

For given s, find new s' that approximately minimizes the first derivative of $p(s|X,Y)$. Repeat until a (local) minimum has been obtained.

Confidence

For fixed s we use the normal distribution in Y as a measure for the confidence.