ADVANCED MACHINE LEARNING
GAUSSIAN PROCESSES
LECTURE 2

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Lecture overview

- Lecture 2
  - Estimating the **GP hyperparameters**
  - More on **kernel functions**
  - **Large scale GPs**

Estimating the hyperparameters

- Kernel depends on **hyperparameters** θ. Example SE kernel $[\theta = (\sigma_f, \ell)^T]$

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\ell^2}\right)$$

- Common approach: choose the hyperparameters that maximizes the **marginal likelihood** (**evidence**):

$$p(\mathbf{y}|\mathbf{X},\theta) = \int p(\mathbf{y}|\mathbf{X},\mathbf{f},\theta)p(\mathbf{f}|\mathbf{X},\theta)d\mathbf{f}$$

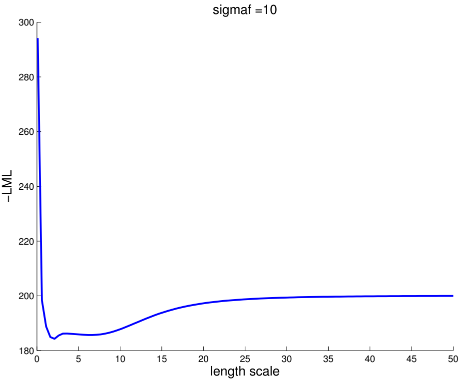
  where $\mathbf{f} = f(\mathbf{X})$ is a vector with function values in the training data.
- For Gaussian process regression:

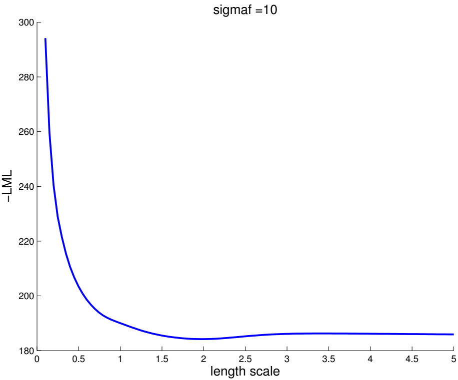
$$\log p(\mathbf{y}|\mathbf{X},\theta) = -\frac{1}{2}\mathbf{y}^T\left(K+\sigma_n^2 I\right)^{-1}\mathbf{y} - \frac{1}{2}\log\left|K+\sigma_n^2 I\right| - \frac{n}{2}\log(2\pi)$$

- Proper **Bayesian inference** **for hyperparameters**

$$p(\theta|\mathbf{y},\mathbf{X}) \propto p(\mathbf{y}|\mathbf{X},\theta)p(\theta).$$

Canadian wages - LML determination of ℓ





More than one input - ARD

- Anisotropic version of isotropic kernels by setting $r^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')$ where $\mathbf{M}$ is positive definite.

- **Automatic Relevance Determination** (**ARD**): $\mathbf{M} = Diag(\ell_1^{-2}, ..., \ell_D^{-2})$ is diagonal with different length scales.

- ARD does 'variable selection' since large ℓ_j means that the jth input essentially drops out of $f(\mathbf{x})$.

More on kernels

- **Periodic kernels**. When $f(x)$ is believed to be periodic with period d. Example:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{2\sin^2\left(\pi\left|x - x'\right|/d\right)}{\ell^2}\right).$$

- **Factor kernels**: $M = \Lambda\Lambda^T + \Psi$, where Λ is $D \times k$ for low rank k.
- Length-scales $\ell(\mathbf{x})$ that vary with $\mathbf{x}$. **Gibbs kernel** in RW Eq. 4.32. **Adaptive smoothness**.

Product of kernels

- Kernels are often combined into **composite kernels**.
- **Product** of kernels is a kernel.
- Example: Product of periodic and square exponential kernels. Locally periodic. Two nearby peaks are more dependent than two distant peaks.

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{2\sin^2\left(\pi\left|x - x'\right|^2/d\right)}{\ell^2}\right) \times \exp\left(-\frac{1}{2}\frac{\left|x - x'\right|^2}{\ell^2}\right)$$

- Example: ARD is a product of D one-dimensional kernels, one for each input variable

$$k_{ARD}(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^{D} k_{SE,\ell_d}(x_d, x_d')$$

Sum of kernels

- **Sum** of kernels is a kernel.
- Let $f_a \sim GP\left[m_a(\mathbf{x}), k_a(\mathbf{x}, \mathbf{x}')\right]$ independently of $f_b \sim GP\left[m_b(\mathbf{x}), k_b(\mathbf{x}, \mathbf{x}')\right]$ then

$$f_a + f_b \sim GP\left[m_a(\mathbf{x}) + m_b(\mathbf{x}), k_a(\mathbf{x}, \mathbf{x}') + k_b(\mathbf{x}, \mathbf{x}')\right]$$

- Adding up kernels is the same as adding up functions.

Discrete covariates

- Suppose: x_1 is continuous (mg/week) and x_2 is binary (sex).
- Linear regression: just use x_2 coded as $x_2 = 0$ if male, $x_2 = 1$ if female.
- Implicit model:

$$y = \begin{cases} \beta_0 + \beta_1 x_1 & \text{if } x_2 = 0 \\ \beta_0 + \tilde{\beta}_0 + (\beta_1 + \tilde{\beta}_1)x_1 & \text{if } x_2 = 1 \end{cases}$$

- GP: add the 0-1 coded covariate and use ARD kernel:

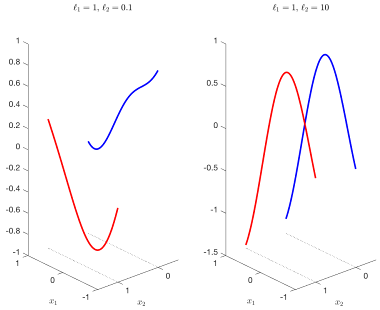
$$\exp\left(-\frac{1}{2}\left(\frac{x_1 - x_1'}{\ell_1}\right)^2\right)\exp\left(-\frac{1}{2}\left(\frac{x_2 - x_2'}{\ell_2}\right)^2\right)$$

So the covariance between $(x_1, 0)$ and $(x_1, 1)$ is

$$\exp\left(-\frac{1}{2}\left(\frac{1}{\ell_2}\right)^2\right)$$

Discrete covariates

- Large ℓ_2: men and female are believed to have similar profiles with respect to x_1.
- Small ℓ_2: men and female are believed to have potentially very different profiles with respect to x_1.



- Categorical covariates with K levels: create K *one-hot* variables.

Large scale GPs

- GPs are **computationally challenging**. Need to invert $n \times n$ matrices such as $\left[K(\mathbf{x},\mathbf{x}) + \sigma^2 I\right]^{-1}$. **Scales as** $O(n^3)$.
- **Banded covariance functions**.
  - Special covariance functions that makes $K(\mathbf{x},\mathbf{x})$ sparse.
  - Observations more than a certain distance apart are uncorrelated.
  - Sparse matrix algebra.
  - Still $O(n^3)$, but with much smaller proportionality constant (i.e. much faster for a given n).

Large scale GPs

- Introduce m latent **inducing variables** $\mathbf{u} = \{u_1, ..., u_m\}$ with corresponding inducing inputs $\mathbf{X}_u = \{\mathbf{x}_{u_1}, \mathbf{x}_{u_2}, ..., \mathbf{x}_{u_m}\}$. Pseudo inputs.
- The **Fully Independent Conditional** (**FIC**) method *assumes* that the elements in $\mathbf{f}$ are independent given the $\mathbf{u}$

$$p(\mathbf{f}|\mathbf{X}, \mathbf{X}_u, \mathbf{u}, \theta) = \prod_{i=1}^{n} p_i(f_i|\mathbf{X}, \mathbf{X}_u, \mathbf{u}, \theta)$$

- Computations are now $O(m^2 n)$. If $m \ll n$, much faster computations.
- **Partially Independent Conditional** (**PIC**). Extension where blocks of $\mathbf{f} = (\mathbf{f}_1, ..., \mathbf{f}_k)$ and each $\mathbf{f}_i$ is a block of b elements from $\mathbf{f}$.
  PIC assumes that the blocks are independent given the inducing variables $\mathbf{u}$, but that the elements within each block are dependent.
  $b = 1$ gives FIC. $b = n$ gives the original GP.
- The locations of the inducing variables $\mathbf{X}_u$ are learned by optimization.

Example Matlab's own toolbox

- Statistics and Machine Learning Toolbox.
- Many kernels, fitting methods etc.
- Limited to **regression** (continuous response).
- Can include explicit basis functions.

- ```
  gprMdl = fitrgp(Xtrain,ytrain,'FitMethod','fic',
  'KernelFunction','ardsquaredexponential',
  'KernelParameters',[sigmaM0;sigmaF0],
  'Sigma',sigma0);
  ```

- See `MatlabGPexample.m`
  ```

Example R - kernlab

- The kernlab package includes many Kernel methods (e.g. SVM), including also GPs.
- Non-traditional parametrization of kernel functions.
- Can do both **regression** (continuous response) or **classification** (categorical response).

- `GPfit <- gausspr(logWage ~ age, kernel = 'rbfdot', par = list(sigma = 1))`

- See `KernLabDemo.R`