Advanced Bayesian Learning
Gaussian Processes
Spring 2014

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Topic overview

- Gaussian process regression
  - Recall: Bayesian inference for multivariate normal model
  - Gaussian processes for flexible regression
  - Covariance kernels
  - Properties of GPs
  - Selecting the kernel and hyperparameters
- Gaussian process classification
  - Flexible classification
  - Laplace approximation of the posterior
- Main literature: Rasmussen and Williams (2006). *Gaussian Processes for Machine Learning*.

Flexible nonlinear regression

- **Linear regression**

$$y = f(\mathbf{x}) + \text{"}$$

$$f(\mathbf{x}) = \mathbf{x}^T \cdot \mathbf{w}$$

  and " $\sim N(0, \sigma_n^2)$ and iid over observations.
- The weights $\mathbf{w}$ are called regression coefficients (β) in statistics.
- **Polynomial regression**: $\mathbf{x} = (1, x, x^2, x^3, ..., x^k)^T$.
- **Spline regression**:

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \cdot \mathbf{w}$$

  where $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), ..., \phi_N(\mathbf{x}))^T$ for N basis functions
- Example: **thin plate splines** with N knots $\kappa_1, ..., \kappa_N$ in $\mathbf{x}$-space

$$\phi_k(\mathbf{x}) = \ln\left(\|\mathbf{x} - \kappa_k\|\right)\|\mathbf{x} - \kappa_k\|$$

- Note: these models are still linear in the weights.

Bayesian linear regression - inference

- $\mathbf{w}$ is unknown. σ_n is assumed known.
- **Prior** [note: RW does *not* use $\Sigma_p = \sigma_n^2 \Omega$.]

$$\mathbf{w} \sim N\left(0, \Sigma_p\right)$$

- **Posterior**

$$\begin{aligned} \mathbf{w}|\mathbf{X},\mathbf{y} &\sim N\left(\bar{\mathbf{w}}, \mathbf{A}^{-1}\right) \\ \mathbf{A} &= \sigma_n^{-2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1} \\ \bar{\mathbf{w}} &= \sigma_n^{-2}\mathbf{A}^{-1}\mathbf{X}\mathbf{y} \end{aligned}$$

- Recall: **Posterior precision = Data Precision + Prior Precision**
- Posterior is student t when σ_n^2 is unknown with Inv-χ^2 prior.

Bayesian linear regression - prediction

- **Predictive density for mean** $f(\mathbf{x}_*)$ at new location $\mathbf{x}_*$

$$f(\mathbf{x}_*)|\mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim N\left(\mathbf{x}_*^T \bar{\mathbf{w}}, \mathbf{x}_*^T \mathbf{A}^{-1}\mathbf{x}_*\right)$$

- Proof: $f(\mathbf{x}_*) = \mathbf{x}_*^T \mathbf{w}$ and $\mathbf{w}$ has a normal posterior. Use that linear combs of normals is normal.

- **Predictive density for new response** y_*

$$\mathbf{y}_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim N\left(\mathbf{x}_*^T \bar{\mathbf{w}}, \mathbf{x}_*^T \mathbf{A}^{-1}\mathbf{x}_* + \sigma_n^2\right)$$

Non-parametric regression

- **Non-parametric regression**: avoiding a parametric form for $f(\cdot)$.
- **Function space view**
  - Treat f as an **unknown function**.
  - Put a **prior over a set of functions**.
- **Weight space view**
  - Restrict attention to a grid of (ordered) x-values: $x_1, x_2, .., x_k$.
  - Put a joint prior on the k function values: $f(x_1), f(x_2), ..., f(x_k)$.

Gaussian process regression

- Natural choice. Multivariate normal (Gaussian):

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

- But how do we specify the $k \times k$ **covariance matrix K**?

$$Cov\left(f(x_p), f(x_q)\right)$$

Gaussian process regression

- Natural choice. Multivariate normal (Gaussian):

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

- But how do we specify the $k \times k$ **covariance matrix K**?

$$Cov(f(x_p), f(x_q))$$

- **Squared exponential covariance function**

$$Cov(f(x_p), f(x_q)) = K(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_p - x_q)^2\right)$$

- The covariance between $f(x_p)$ and $f(x_q)$ is a function of x_p and x_q.
- Nearby x's have highly correlated function ordinates $f(x)$.
- We can compute $Cov(f(x_p), f(x_q))$ for *any* x_p and x_q (no need for a pre-determined grid)

Gaussian process regression, cont.

Definition

A **Gaussian process** (**GP**) is a collection of random variables, any finite number of which have a multivariate Gaussian distribution.

- A Gaussian process is really a **probability distribution over functions** (curves). This is exactly what we want! No need for a grid!
- A GP is completely specified by a mean and a covariance function

$$m(x) = \mathrm{E}\left[f(x)\right]$$

$$K(x, x') = E\left[\left(f(x) - m(x)\right)\left(f(x') - m(x')\right)\right]$$

  for any two inputs x and x' (note: this is *not* the transpose here).
- A Gaussian process (prior) is denoted by

$$f(x) \sim GP\left(m(x), K(x, x')\right)$$

Gaussian process regression, cont.

- Example:

$$m(x) = \sin(x)$$
$$K(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\frac{x_p - x_q}{\ell}\right)^2\right)$$

  where $\ell > 0$ is the length scale.

- Larger l gives more smoothness in $f(x)$.
- Simulate draw from $f(x) \sim GP\left(m(x), K(x, x')\right)$ over a grid $x_* = (x_1, ..., x_n)$ by using that

$$f(x_*) \sim N\left(m(x_*), K(x_*, x_*)\right)$$

Simulating a GP

- The joint way: Choose a grid $x_1, ..., x_k$. Simulate the k-vector

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

- GRAPH HERE
- More intuition from the conditional decomposition $(y_i = f(x_i))$

$$p(y_1, y_2, ..., y_k) = p(y_1)p(y_2|y_1) \cdots p(y_k|y_1, ..., y_{k-1})$$

[Simulate $p(y_1)$ from bands, $y_2|y_1$ from conditional bands etc]

Gaussian process regression, cont.

- **Model**

$$y_i = f(\mathrm{x}_i) + \varepsilon_i, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- **Prior**

$$f(x) \sim GP\left(0, K(x, x')\right)$$

- You have observed the data: $\mathbf{x} = (x_1, ..., x_n)$' and $\mathbf{y} = (y_1, ..., y_n)'$.
- Goal: the posterior of $f(\cdot)$ over a grid of x-values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.

Gaussian process regression, cont.

- **Model**

$$y_i = f(\mathrm{x}_i) + \varepsilon_i, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- **Prior**

$$f(x) \sim GP\left(0, K(x, x')\right)$$

- You have observed the data: $\mathbf{x} = (x_1, ..., x_n)'$ and $\mathbf{y} = (y_1, ..., y_n)'$.
- Goal: the posterior of $f(\cdot)$ over a grid of x-values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.
- Intermediate step: joint distribution of $\mathbf{y}$ and $\mathbf{f}_*$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} K(\mathbf{x},\mathbf{x}) + \sigma^2 I & K(\mathbf{x},\mathbf{x}_*) \\ K(\mathbf{x}_*,\mathbf{x}) & K(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix} \right\}$$

Gaussian process regression, cont.

- **Model**

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- **Prior**

$$f(x) \sim GP\left(0, K(x, x')\right)$$

- You have observed the data: $\mathbf{x} = (x_1, ..., x_n)'$ and $\mathbf{y} = (y_1, ..., y_n)'$.
- Goal: the posterior of $f(\cdot)$ over a grid of x-values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.

- Intermediate step: joint distribution of $\mathbf{y}$ and $\mathbf{f}_*$

$$\left(\begin{array}{c} \mathbf{y} \\ \mathbf{f}_* \end{array}\right) \sim N\left\{\left(\begin{array}{c} \mathbf{0} \\ \mathbf{0} \end{array}\right), \left[\begin{array}{cc} K(\mathbf{x},\mathbf{x}) + \sigma^2 I & K(\mathbf{x},\mathbf{x}_*) \\ K(\mathbf{x}_*,\mathbf{x}) & K(\mathbf{x}_*,\mathbf{x}_*) \end{array}\right]\right\}$$

- The **posterior**

$$\mathbf{f}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_* \sim N\left(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)\right)$$

$$\bar{\mathbf{f}}_* = K(\mathbf{x}_*, \mathbf{x}) \left[K(\mathbf{x},\mathbf{x}) + \sigma^2 I\right]^{-1} \mathbf{y}$$

$$\mathrm{cov}(\mathbf{f}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{x}) \left[K(\mathbf{x},\mathbf{x}) + \sigma^2 I\right]^{-1} K(\mathbf{x}, \mathbf{x}_*)$$