Gaussian Process

Jiali Lin

Virginia Tech

January 15, 2017

Outline

Introduction

GPs for regression

GPs meet GLMs

Introduction

- We observe some inputs $\boldsymbol{x}_i$ and some outputs y_i (i.i.d).
- We assume linear relationships

$$y = f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}, \quad y = f(\boldsymbol{x}) + \epsilon, \quad \epsilon \sim N(0, \sigma_n^2)$$

- Prior: $\boldsymbol{w} \sim N(0, \boldsymbol{\Sigma}_p)$.
- To make predictions on new input $\boldsymbol{x}_*$, we use posterior predictive distribution

$$p(y_* | \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) = \int p(y_* | f, \boldsymbol{x}_*) p(f | \boldsymbol{X}, \boldsymbol{y}) df$$

- Alternatively one could first map $\boldsymbol{x}$ to some basis function, then let

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

Main Ideas

- Want more flexible form for $f(\boldsymbol{x})$, treat the whole $f(.)$ as a parameter.
- Assume $f(.)$ takes values from a function space.
- Assume $f(.)$ is random. In particular, Gaussian Process.
- **Gaussian processes** or **GPs**: defines a prior over functions, which can be converted into a posterior over functions once we have seen some data.

Outline

Introduction

GPs for regression

GPs meet GLMs

Gaussian process

- **Definition**: Gaussian process is a collection of random variables, any *finite* number of which have a joint Gaussian distribution.
- Let the prior on the regression function be a GP, denoted by

$$f(\boldsymbol{x}) \sim \mathsf{GP}(m(\boldsymbol{x}), \kappa(\boldsymbol{x}, \boldsymbol{x'}))$$

- $m(\boldsymbol{x})$ is the mean function and $\kappa(\boldsymbol{x}, \boldsymbol{x'}))$ is the covariance function

$$\begin{aligned} m(\boldsymbol{x}) &= E[f(\boldsymbol{x})] \\ \kappa(\boldsymbol{x}, \boldsymbol{x'}) &= E[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))^T] \end{aligned}$$

- Require $\kappa()$ be a positive definite kernel. For any finite set of points, this process defines a joint Gaussian

$$p(\boldsymbol{f}|\boldsymbol{X}) = N(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{K})$$

- $K_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{\mu} = (m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_N))$. Note that it is common to use a mean function of $m(\boldsymbol{x}) = 0$, since the GP is flexible enough to model the mean arbitrarily well.

Predictions using noise-free observations

- Given noise-free training data

$$\mathcal{D} = \{\boldsymbol{x}^{(i)}, f^{(i)} | i = 1, \ldots, n\}$$

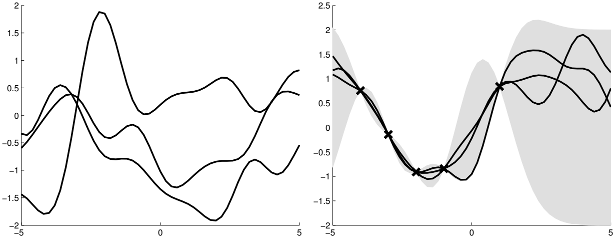
- We want to make predictions $\boldsymbol{f}_*$ at test points $\boldsymbol{X}_*$.
- By definition of the GP, the joint distribution has the following form

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \quad \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{X}, \boldsymbol{X}_*) \\ K(\boldsymbol{X}_*, \boldsymbol{X}) & K(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix} \right)$$

- Posterior: $p(\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f}) \sim N(\boldsymbol{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$

$$\begin{aligned} \boldsymbol{\mu}_* &= \mu(\boldsymbol{X}_*) + K(\boldsymbol{X}_*, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} (\boldsymbol{f} - \mu(\boldsymbol{X})) \\ \boldsymbol{\Sigma}_* &= K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}, \boldsymbol{X}_*) K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{X}_*, \boldsymbol{X}) \end{aligned}$$

Predictions using noise-free observations (cont'd)



(a)

(b)

Figure: Left: some functions sampled from a GP prior with SE kernel. Right: some samples from a GP posterior, after conditioning on 5 noise-free observations. The shaded area represents $E[f(\boldsymbol{x})] \pm 2\text{std}(f(\boldsymbol{x}))$. Based on Figure 2.2 of (Rasmussen and Williams 2006). Figure generated by `GprDemo`.

Predictions using noisy observations

- $y = f(\boldsymbol{x}) + \epsilon$, where $\epsilon \sim N(0, \Sigma_y^2)$.
- The covariance of the observed noisy responses is

$$\mathsf{cov}[y_p, y_q] = \kappa(\boldsymbol{x}_p, \boldsymbol{x}_q) + \Sigma_y^2 \delta_{pq}, \quad \text{where} \quad \delta_{pq} = \mathbb{I}(p = q)$$

- We assume the noise terms were independent.
- Set mean function $m(\boldsymbol{x}) = 0$. Thus, $f(\boldsymbol{x}) \sim \mathsf{GP}(0, \kappa(\boldsymbol{x}, \boldsymbol{x}'))$
- Now the joint distribution is given by

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \quad \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I & K(\boldsymbol{X}, \boldsymbol{X}_*) \\ K(\boldsymbol{X}_*, \boldsymbol{X}) & K(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix} \right)$$

- Posterior: $p(\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f}) \sim N(\boldsymbol{f}_* | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$

$$\boldsymbol{\mu}_* = K(\boldsymbol{X}_*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}]^{-1} y$$
$$\boldsymbol{\Sigma}_* = K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}, \boldsymbol{X}_*)[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 n \boldsymbol{I}]^{-1} K(\boldsymbol{X}_*, \boldsymbol{X})$$

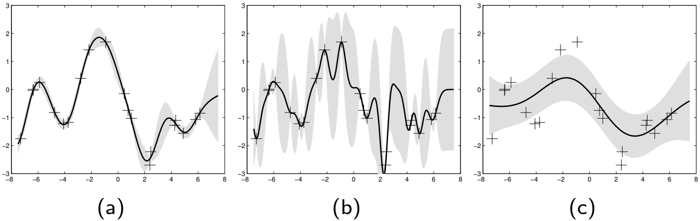
Effect of the kernel parameters

- The predictive performance of GPs depends exclusively on the chosen kernel.
- Suppose we choose **squared-exponential** (SE) kernel for the noisy observations

$$\kappa_y(x_p, x_q) = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(x_p - x_q)^2) + \sigma_y^2 \delta_{pq}$$

- ℓ is the horizontal scale over which the function changes, σ_f^2 controls the vertical scale of the function, and σ_y^2 is the noise variance.

Effect of the kernel parameters (cont'd)



(a) (b) (c)

Figure: Some 1d GPs with SE kernels but diferent hyper-parameters fit to 20 noisy observations. The kernel has the form in Equation (14). The hyper-parameters $(\ell, \sigma_f, \sigma_y)$ are as follows: (a) $(1, 1, 0.1)$ (b) $(0.3, 0.1.08, 0.00005)$, (c) $(3.0, 1.16, 0.89)$. Figure generated by `gprDemoChangeHparams`, written by Carl Rasmussen.

- In Figure (a), the result is a good fit.
- In Figure (b), the function looks more "wiggly". Also, the uncertainty goes up faster.
- In Figure (c), the function looks smoother.

Estimating the kernel parameters

- Frequentist: exhaustive search over a discrete grid of values (slow!).
- Consider empirical Bayes approach.
- Maximization log likelihood can be done using efficient gradient-based optimization algorithms.
- In absence of prior $p(\boldsymbol{\theta})$, the posterior for hyperparameter $\boldsymbol{\theta}$ is proportional to the marginal likelihood

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

- Choose $\boldsymbol{\theta}$ to optimize the marginal log-likelihood

$$\log p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) \propto -\frac{1}{2} \log |K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}| - \frac{1}{2} \boldsymbol{y}^T (K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}$$

Outline

Introduction

GPs for regression

GPs meet GLMs

Gaussian Process Classification

- In the binary case, we have

$$
\begin{aligned}
y_i &\in \{0, 1\} \\
p(y_i|\boldsymbol{x}_i, f_i) &= \exp\{y_i f_i - A(f_i)\} \\
p(\boldsymbol{y}) &= N(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}), \quad \text{where} \quad K_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
p(y_i|\boldsymbol{x}_i, f_i) &= \mathsf{Ber}(y_i|\mathsf{Sigm}(f_i))
\end{aligned}
$$

- Equivalent to logistic regression, but uses kernels rather than features.
- Gaussian prior on weights replaced by Gaussian prior on training log-odds.
- Basic inference finds MAP estimate of function f at all training points

$$
\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmax}} \log p(\boldsymbol{f}) + \sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i, f_i)
$$

Gaussian Process Classification (cont'd)

- Interpretation of function values f_i
  - Postive: $p(y_i = 1|f_i) > 0.5$.
  - Zero: $p(y_i = 1|f_i) = 0.5$.
  - Negative: $p(y_i = 1|f_i) < 0.5$.
- Interpretation of kernel values K_{ij}
  - Postive: likely have same label.
  - Zero: inputs are totally independent.
  - Negative: likely have different labels.

Gaussian Process Classification (cont'd)

First, compute the distribution of the latent variable for a test case

$$p(f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) = N(\mathsf{E}[f_*], \mathsf{var}[f_*])$$

Second, produce a predictive distribution for binary responses

$$\pi_* = p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) \approx \int \mathsf{Sigm}(f_*) p(f_*|\boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}) df_*$$

- Note that $p(\boldsymbol{y}|\boldsymbol{f})$ has link function involved, conjugacy of $\boldsymbol{f}$ are lost.
- Also integrations are difficult.
- **Solutions:** use analytic approximations of integrals to approximate $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y})$, or Monte Carlo sampling.