Gaussian processes

Nathanaël Carraz Rakotonirina

Mathématiques Informatique et Statistique Appliquées (MISA)
Université d'Antananarivo

Model

- It is a Bayesian and **nonparametric** method. We do not estimate parameters but the function itself from data.
- We observe the function value at a fixed set of M points, $y_i = f(\boldsymbol{x}_i), i = 1...M$
- It defines distributions over functions of the form $f : \mathcal{X} \to \mathbb{R}$. The function values at a set of M points, $f = [f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_M)]$ is jointly Gaussian with
  - mean $\boldsymbol{\mu} = [m(\boldsymbol{x}_1), ..., m(\boldsymbol{x}_M)]$ where m is a mean function
  - covariance $\boldsymbol{\Sigma}_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ where $\mathcal{K}$ is a positive definite (Mercer) kernel
- If $M = N + 1$ with N training points $\boldsymbol{x}_i$ and one test point $\boldsymbol{x}_*$, we can infer $f(\boldsymbol{x}_*)$ from knowledge of $f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_N)$ and the joint Gaussian.

Noise-free observations

We opbserve a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where $y_i = f(\boldsymbol{x}_i)$ is the noise-free observation of the function evaluated at $\boldsymbol{x}_i$.

We want make prediction for new inputs not in $\mathcal{D}$. Given a test set $\boldsymbol{X}_*$ of size $N_* \times D$, we want to predict the function outputs $f_* = [f(\boldsymbol{x}_1^*), \dots, f(\boldsymbol{x}_{N_*}^*)]$.

The joint distribution $p(f_X, f_* | \boldsymbol{X}, \boldsymbol{X}_*)$ has the form

$$\begin{pmatrix} f_X \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \boldsymbol{K}_{X,X} & \boldsymbol{K}_{X,*} \\ \boldsymbol{K}_{X,*}^\top & \boldsymbol{K}_{*,*} \end{pmatrix} \right)$$

where

- $\mu_X = [m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_N)]$
- $\mu_* = [m(\boldsymbol{x}_1^*), \dots, m(\boldsymbol{x}_{N_*}^*)]$
- $\boldsymbol{K}_{X,X} = \mathcal{K}(\boldsymbol{X}, \boldsymbol{X})$ is $N \times N$
- $\boldsymbol{K}_{X,*} = \mathcal{K}(\boldsymbol{X}, \boldsymbol{X}_*)$ is $N \times N_*$
- $\boldsymbol{K}_{*,*} = \mathcal{K}(\boldsymbol{X}_*, \boldsymbol{X}_*)$ is $N_* \times N_*$

The posterior predictive density is also Gaussian

$$p(f_* | f_X, \boldsymbol{X}, \boldsymbol{X}_*) = \mathcal{N}(f_* | \boldsymbol{\mu}_{*|X}, \boldsymbol{\Sigma}_{*|X})$$

$$\boldsymbol{\mu}_{*|X} = \boldsymbol{\mu}_X + \boldsymbol{K}_{X,*}^\top \boldsymbol{K}_{X,X}^{-1} (f_X - \boldsymbol{\mu}_*)$$

$$\boldsymbol{\Sigma}_{*|X} = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{X,*}^\top \boldsymbol{K}_{X,X}^{-1} \boldsymbol{K}_{X,*}$$

Gaussian process regression (noisy observations)

We observe a noisy version of the function $y_i = f(\boldsymbol{x}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_y)$. The joint distribution becomes

$$\begin{pmatrix} f_X \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \boldsymbol{K}_{X,X} + \sigma_y^2 \boldsymbol{I} & \boldsymbol{K}_{X,*} \\ \boldsymbol{K}_{X,*}^\top & \boldsymbol{K}_{*,*} + \sigma_y^2 \boldsymbol{I} \end{pmatrix} \right)$$

The posterior predictive density is also Gaussian

$$\begin{aligned} p(f_* | f_X, \boldsymbol{X}, \boldsymbol{X}_*) = \mathcal{N}(f_* | \boldsymbol{\mu}_{*|X}, \boldsymbol{\Sigma}_{*|X}) \\ \boldsymbol{\mu}_{*|X} &= \boldsymbol{\mu}_X + \boldsymbol{K}_{X,*}^\top (\boldsymbol{K}_{X,X} + \sigma_y^2 \boldsymbol{I})^{-1} (f_X - \boldsymbol{\mu}_*) \\ \boldsymbol{\Sigma}_{*|X} &= \boldsymbol{K}_{*,*} + \sigma_y^2 \boldsymbol{I} - \boldsymbol{K}_{X,*}^\top \boldsymbol{K}_{X,X}^{-1} \boldsymbol{K}_{X,*} \end{aligned}$$

Advantages

- Gaussian process models lead to simple and straightforward linear algebra implementations.
- As Bayesian methods, they allow one to quantify uncertainty in predictions
- Gaussian process regression is non-parametric and hence can model essentially arbitrary functions of the input points.
- They provide a natural way to introduce kernels into a regression modeling framework.
- Methods for model selection and hyperparameter selection in Bayesian methods are immediately applicable to Gaussian processes

There is so much more !

Explore

- Numerical issues (consider Cholesky decomposition instead of direct inversion)
- Comparison to kernel regression and Bayesian linear regression
- Estimating the kernel the Bayesian way
- Gaussian processes for classification
- Scaling to large datasets