Learning Curves in Gaussian Process Regression

April 17, 2024

Introduction to GP

- Bayesian Linear Regression: Model $y|X, w \sim \mathcal{N}(Xw, \sigma_n^2 I_n)$ and prior $w \sim \mathcal{N}(0, \Sigma_p)$. Here $X \in \mathbb{R}^{p \times n}$ is the data matrix.

Introduction to GP

- Bayesian Linear Regression: Model $y|X, w \sim \mathcal{N}(Xw, \sigma_n^2 I_n)$ and prior $w \sim \mathcal{N}(0, \Sigma_p)$. Here $X \in \mathbb{R}^{p \times n}$ is the data matrix.
- Posterior becomes

$$p(w|y, X) \propto p(y|X, w)p(w) = \mathcal{N}\left(\sigma_n^{-2} A^{-1} X y, A^{-1}\right)$$

where $A = \sigma_n^{-2} X X^T + \Sigma_p^{-1} \in \mathbb{R}^{p \times p}$.

Introduction to GP

- Bayesian Linear Regression: Model $y|X, w \sim \mathcal{N}(Xw, \sigma_n^2 I_n)$ and prior $w \sim \mathcal{N}(0, \Sigma_p)$. Here $X \in \mathbb{R}^{p \times n}$ is the data matrix.
- Posterior becomes

$$p(w|y, X) \propto p(y|X, w)p(w) = \mathcal{N}\left(\sigma_n^{-2} A^{-1} X y, A^{-1}\right)$$

  where $A = \sigma_n^{-2} X X^T + \Sigma_p^{-1} \in \mathbb{R}^{p \times p}$.
- Predictive distribution at new x...

$$p(f(x)|x, X, y) = \mathcal{N}\left(\sigma_n^{-2} x^T A^{-1} X y, x^T A^{-1} x\right)$$

Introduction to GP

- Bayesian Linear Regression: Model $y|X, w \sim \mathcal{N}(Xw, \sigma_n^2 I_n)$ and prior $w \sim \mathcal{N}(0, \Sigma_p)$. Here $X \in \mathbb{R}^{p \times n}$ is the data matrix.
- Posterior becomes

$$p(w|y, X) \propto p(y|X, w)p(w) = \mathcal{N}\left(\sigma_n^{-2} A^{-1} X y, A^{-1}\right)$$

  where $A = \sigma_n^{-2} X X^T + \Sigma_p^{-1} \in \mathbb{R}^{p \times p}$.
- Predictive distribution at new x...

$$p(f(x)|x, X, y) = \mathcal{N}\left(\sigma_n^{-2} x^T A^{-1} X y, x^T A^{-1} x\right)$$

- Can directly be kernelized with Φ replacing X.

Introduction to GP

- A bit more mathematically sound: We define $f \sim GP(\mu, k)$ as a Gaussian process on $\mathbb{R}^d$ if for all datasets $X = \{x_1, \dots, x_n\}$, we get

  $$f(X) \sim \mathcal{N}(\mu(x_i), k(x_i, x_j))$$

  This specifies the distribution completely.

Introduction to GP

- A bit more mathematically sound: We define $f \sim GP(\mu, k)$ as a Gaussian process on $\mathbb{R}^d$ if for all datasets $X = \{x_1, \dots, x_n\}$, we get

  $$f(X) \sim \mathcal{N}(\mu(x_i), k(x_i, x_j))$$

  This specifies the distribution completely.
- The GP regression model is $y = f(x) + \epsilon$ with noise ϵ at level σ_n^2 and for $f \sim GP(\mu, k)$.

Introduction to GP

- A bit more mathematically sound: We define $f \sim GP(\mu, k)$ as a Gaussian process on $\mathbb{R}^d$ if for all datasets $X = \{x_1, \dots, x_n\}$, we get
$$f(X) \sim \mathcal{N}(\mu(x_i), k(x_i, x_j))$$
This specifies the distribution completely.
- The GP regression model is $y = f(x) + \epsilon$ with noise ϵ at level σ_n^2 and for $f \sim GP(\mu, k)$.
- Given data X with observations y, and a new x, we have
$$f(x)|X, y, x \sim \mathcal{N}\left(\bar{f}(x), \sigma^2(f(x))\right)$$
where
$$\bar{f}(x) = k(x, X)\left(K + \sigma_n^2 I_n\right)^{-1} y$$
$$\sigma^2(x) = k(x, x) - k(x, X)\left(K + \sigma_n^2 I_n\right)^{-1} k(X, x)$$

Introduction to GP

- A bit more mathematically sound: We define $f \sim GP(\mu, k)$ as a Gaussian process on $\mathbb{R}^d$ if for all datasets $X = \{x_1, \ldots, x_n\}$, we get

$$f(X) \sim \mathcal{N}(\mu(x_i), k(x_i, x_j))$$

  This specifies the distribution completely.
- The GP regression model is $y = f(x) + \epsilon$ with noise ϵ at level σ_n^2 and for $f \sim GP(\mu, k)$.
- Given data X with observations y, and a new x, we have

$$f(x)|X, y, x \sim \mathcal{N}\left(\bar{f}(x), \sigma^2(f(x))\right)$$

  where

$$\bar{f}(x) = k(x, X)\left(K + \sigma_n^2 I_n\right)^{-1} y$$
$$\sigma^2(x) = k(x, x) - k(x, X)\left(K + \sigma_n^2 I_n\right)^{-1} k(X, x)$$

- It's kernel regression with uncertainty estimation.

Generalization Error for GP

- Suppose $f \sim GP(0, k_0)$ and y has noise level σ_0. We estimate f using GP regression with kernel k_1 and noise σ_1.

Generalization Error for GP

- Suppose $f \sim GP(0, k_0)$ and y has noise level σ_0. We estimate f using GP regression with kernel k_1 and noise σ_1.
- Generalization error (averaging over new data x and the prior f) is

$$E^{gen}(X) = \int k_0(x,x)dp(x) - 2\,Trace\left(K_{1,\sigma_1^2}^{-1}\int k_0(X,x)k_1(x,X)dp(x)\right) + Trace\left(K_{1,\sigma_1^2}^{-1}K_{0,\sigma_0^2}K_{1,\sigma_1^2}^{-1}\int k_1(X,x)k_0(x,X)dp(x)\right)$$

Generalization Error for GP

- Suppose $f \sim GP(0, k_0)$ and y has noise level σ_0. We estimate f using GP regression with kernel k_1 and noise σ_1.
- Generalization error (averaging over new data x and the prior f) is

$$E^{gen}(X) = \int k_0(x,x)dp(x) - 2\,Trace\left(K_{1,\sigma_1^2}^{-1}\int k_0(X,x)k_1(x,X)dp(x)\right) + Trace\left(K_{1,\sigma_1^2}^{-1}K_{0,\sigma_0^2}K_{1,\sigma_1^2}^{-1}\int k_1(X,x)k_0(x,X)dp(x)\right)$$

- In the well-specified case

$$E^{gen}(X) = \int k_0(x,x)dp(x) - Trace\left(K_{0,\sigma_0^2}^{-1}\int k_0(X,x)k_0(x,X)dp(x)\right)$$

Generalization Error for GP

- Mercer's Theorem: $k(x, x') = \sum_i \lambda_i \phi_i(x)\phi_i(x')$. This gives

$$E^{gen}(X) = Trace\left(\left(\Lambda + \sigma^{-2}\Phi\Phi^T\right)^{-1}\right)$$

here ϕ_i are L^2 orthonormal, i.e. $\int \phi_i(x)\phi_j(x)dp(x) = \delta_{ij}$ for all i, j.

Generalization Error for GP

- Mercer's Theorem: $k(x, x') = \sum_i \lambda_i \phi_i(x)\phi_i(x')$. This gives

$$E^{gen}(X) = Trace\left(\left(\Lambda + \sigma^{-2}\Phi\Phi^T\right)^{-1}\right)$$

  here ϕ_i are L^2 orthonormal, i.e. $\int \phi_i(x)\phi_j(x)dp(x) = \delta_{ij}$ for all i, j.
- Using $\mathbb{E}[\Phi\Phi^T] = nI$ we get the simple approximation

$$E^{gen} \approx Trace\left(\left(\Lambda + \sigma^{-2}nI\right)^{-1}\right) = \sum_i \frac{\lambda_i \sigma^2}{\sigma^2 + n\lambda_i}$$

Generalization Error for GP

- Mercer's Theorem: $k(x, x') = \sum_i \lambda_i \phi_i(x) \phi_i(x')$. This gives

$$E^{gen}(X) = Trace\left(\left(\Lambda + \sigma^{-2}\Phi\Phi^T\right)^{-1}\right)$$

  here ϕ_i are L^2 orthonormal, i.e. $\int \phi_i(x)\phi_j(x)dp(x) = \delta_{ij}$ for all i, j.

- Using $\mathbb{E}[\Phi\Phi^T] = nI$ we get the simple approximation

$$E^{gen} \approx Trace\left(\left(\Lambda + \sigma^{-2}nI\right)^{-1}\right) = \sum_i \frac{\lambda_i \sigma^2}{\sigma^2 + n\lambda_i}$$

- This is a lower bound on E^{gen}.

Sollich's 1st Approximation

- Let's see how new data affects $G(n) := \left(\Lambda + \sigma^{-2}\Phi\Phi^T\right)^{-1}$. We have

$$G(n+1) := \left(\Lambda + \sigma^{-2}\Phi\Phi^T + \sigma^{-2}\phi\phi^T\right)^{-1} = \left(G^{-1}(n) + \sigma^{-2}\phi\phi^T\right)^{-1}$$

  Woodbury's formula for rank 1 updates of matrix inverses gives...

$$G(n+1) - G(n) = -\frac{G(n)\phi\phi^T G(n)}{\sigma^2 + \phi^T G(n)\phi}$$

Sollich's 1st Approximation

- Let's see how new data affects $G(n) := \left(\Lambda + \sigma^{-2}\Phi\Phi^T\right)^{-1}$. We have

$$G(n+1) := \left(\Lambda + \sigma^{-2}\Phi\Phi^T + \sigma^{-2}\phi\phi^T\right)^{-1} = \left(G^{-1}(n) + \sigma^{-2}\phi\phi^T\right)^{-1}$$

  Woodbury's formula for rank 1 updates of matrix inverses gives...

$$G(n+1) - G(n) = -\frac{G(n)\phi\phi^T G(n)}{\sigma^2 + \phi^T G(n)\phi}$$

- We now average over the new point ϕ (note $\mathbb{E}[\phi\phi^T] = I$) and treat n as continuous. We then average this **informally** over X by (i) taking expectations over numerator and denominator separately and (ii) assuming $\mathbb{E}[G(n)^2] = \mathbb{E}[G(n)]^2 =: \bar{G}(n)^2$

Sollich's 1st Approximation

- We get a differential equation

$$\bar{G}' = -\frac{\bar{G}^2}{\sigma^2 + Trace(\bar{G})}$$

that can be solved (see paper).

Sollich's 1st Approximation

- We get a differential equation

$$\bar{G}' = -\frac{\bar{G}^2}{\sigma^2 + Trace(\bar{G})}$$

  that can be solved (see paper).
- Finally, with n' satisfying $n' + \log Trace(I + n'\sigma^{-2}\Lambda) = n$, we have

$$E^{gen} \approx \sum_i \frac{\lambda_i \sigma^2}{\sigma^2 + n'\lambda_i}$$

Sollich's 1st Approximation

- We get a differential equation

$$\bar{G}' = -\frac{\bar{G}^2}{\sigma^2 + Trace(\bar{G})}$$

  that can be solved (see paper).
- Finally, with n' satisfying $n' + \log Trace(I + n'\sigma^{-2}\Lambda) = n$, we have

$$E^{gen} \approx \sum_i \frac{\lambda_i \sigma^2}{\sigma^2 + n'\lambda_i}$$

- Note that $n' < n$, so this is indeed a larger bound than the naive approximation from before.