

GP and BO

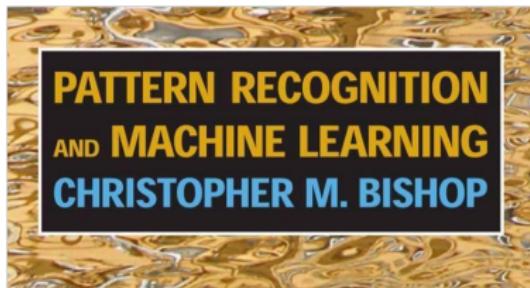
Probabilistic regression and ML assisted optimization

Rodrigo A. Vargas-Hernández

2/Feb/2023

References for today's lecture

Chapter 3 and 6



- Gaussian Processes for ML
- Online material

What about probabilistic regression?

So far we have assumed that for models like,

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

the optimal values of $\boldsymbol{\theta}$ are **deterministic**.

- What if our problem has noise (uncertainty)? $\rightarrow y = \hat{y} + \sigma$

$$y = f(\mathbf{x}) + \epsilon = \boldsymbol{\theta}^T \phi(\mathbf{x}) + \epsilon,$$

where ϵ (noise) can be described probabilistically.

- ϵ is described by a Gaussian distribution with zero mean and variance σ_n^2 ,

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2)$$

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

The probability density of the observations given the parameters,

$$p(\mathbf{y}|\mathbf{X}, \theta) = \prod_i^N p(\mathbf{y}|\mathbf{x}_i, \theta),$$

where,

$$p(\mathbf{y}|\mathbf{x}_i, \theta) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - \theta^T \phi(\mathbf{x}))^2}{2\sigma_n^2}\right)$$

(after some tedious but not complicated algebra we get)

$$p(\mathbf{y}|\mathbf{X}, \theta) = \mathcal{N}(\Phi(\mathbf{X})\theta, \sigma_n^2 \mathbb{I}) = \frac{1}{(2\pi\sigma_n)^{n/2}} \exp\left(-\frac{(\mathbf{y} - \Phi(\mathbf{X})\theta)^2}{2\sigma_n^2}\right)$$

Must probable model

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\Phi(\mathbf{X})\boldsymbol{\theta}, \sigma_n^2 \mathbb{I}) = \frac{1}{(2\pi\sigma_n)^{n/2}} \exp\left(-\frac{(\mathbf{y} - \Phi(\mathbf{X})\boldsymbol{\theta})^2}{2\sigma_n^2}\right)$$

- ▶ What is the most probable $\boldsymbol{\theta}$? (hint: $\ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$)
- ▶ What is the meaning of $\sigma_n^2 \mathbb{I}$?
- ▶ Is $\boldsymbol{\theta}$ deterministic or probabilistic?

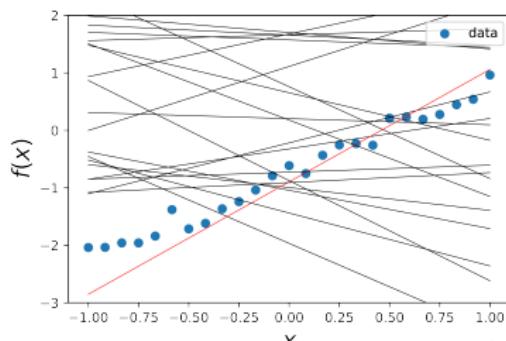
Must probable model

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\Phi(\mathbf{X})\boldsymbol{\theta}, \sigma_n^2 \mathbb{I}) = \frac{1}{(2\pi\sigma_n)^{n/2}} \exp\left(-\frac{(\mathbf{y} - \Phi(\mathbf{X})\boldsymbol{\theta})^2}{2\sigma_n^2}\right)$$

- ▶ What is the most probable $\boldsymbol{\theta}$?

$$\arg \max_{\boldsymbol{\theta}} \ln p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \left(\underbrace{-\frac{1}{2\sigma_n^2} (\mathbf{y} - \Phi(\mathbf{X})\boldsymbol{\theta})^2}_{\text{linear regression}} + \frac{N}{2} \ln \sigma_n - \frac{N}{2} \ln(2\pi) \right)$$

- ▶ Is $\boldsymbol{\theta}$ deterministic or probabilistic?



Bayesian linear model

- "In the Bayesian formalism we need to specify a prior over the parameters, expressing our beliefs about the parameters before we look at the observations" ¹

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\theta})$$

Goal: Infer the posterior distribution over the weights.

Bayes rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})}_{posterior} \approx \underbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}_{likelihood} \underbrace{p(\boldsymbol{\theta})}_{prior}$$

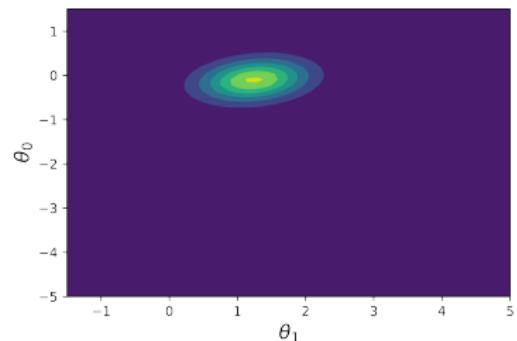
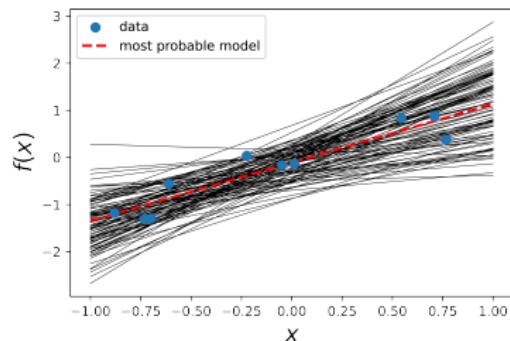
- What is the solution of $\arg \max_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$?

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) &\approx \mathcal{N}(\boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\theta}, \sigma_n^2 \mathbb{I}) \quad \mathcal{N}(\mathbf{0}, \lambda^{-1} \mathbb{I}) \\ \mathcal{N}(\mathbf{0}, \lambda^{-1} \mathbb{I}) &\propto \exp\left(-\frac{1}{\lambda} \boldsymbol{\theta}^T \boldsymbol{\theta}\right) \end{aligned}$$

¹GP book

Bayesian linear model

$$[\theta_1, \theta_0] \sim p(\theta | \mathbf{X}, \mathbf{y})$$



Gaussian Process

$$p(\theta|\Phi(\mathbf{X}), \mathbf{y}) \approx p(\mathbf{y}|\Phi(\mathbf{X}), \theta) p(\theta)$$

- ▶ Same as kernel ridge regression!!
- ▶ Let's use the matrix identity trick and write the equations in terms of kernel function
- ▶ What about prediction?

$$p(y(\mathbf{x})|\mathbf{x}, \Phi(\mathbf{X}), \mathbf{y}) = \int \boldsymbol{\theta}^T \phi(\mathbf{x}) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) d\boldsymbol{\theta}$$

- ▶ Ensemble average!
- ▶ $p(y(\mathbf{x})|\mathbf{x}, \Phi(\mathbf{X}), \mathbf{y})$ is also a Gaussian distribution!!

$$\mu(\mathbf{x}) = \mathbf{k}(\mathbf{x}, \mathbf{X})^\top \left(K(\mathbf{X}, \mathbf{X} + \sigma_n^2 \mathbb{I}) \right)^{-1} \mathbf{y} = \sum_i^N \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

$$\sigma(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x}, \mathbf{X})^\top \left(K(\mathbf{X}, \mathbf{X} + \sigma_n^2 \mathbb{I}) \right)^{-1} \mathbf{k}(\mathbf{x}, \mathbf{X})$$