Bayesian Linear Regression

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

November 16, 2021

Reference

Mathematics for Machine Learning:

https://mml-book.com

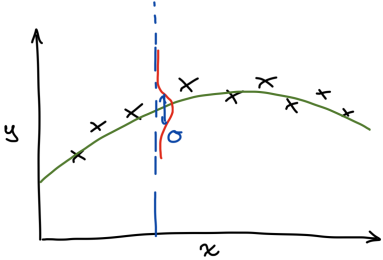
Chapter 9

Probabilistic Models

Probabilistic model: Model of the data is a probability distribution.

$$y_n = f(\mathbf{x}_n; \boldsymbol{\theta}) + \epsilon_n \qquad\qquad \epsilon_n \sim \mathcal{N}\left(0, \sigma^2\right)$$

We can now also estimate the **unpredictability** of our problem:

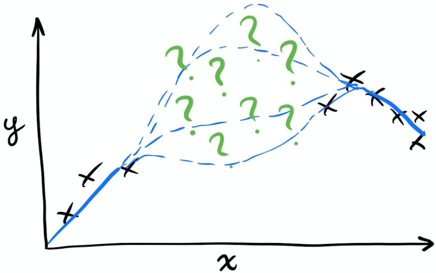


$$(\boldsymbol{\theta}^*, {\sigma^2}^*) = \operatorname{argmax}_{\boldsymbol{\theta}, \sigma^2} \log p(\mathbf{y}|\boldsymbol{\theta}, \sigma^2, X) \qquad (1)$$

Unpredictability remains even if we **know** underlying function.
Goes by many names... e.g. **aleatoric uncertainty**.

Uncertainty in Parameters/Function

Aren't we also uncertain when we have a lack of data?



This is uncertainty in the parameters that define the function!
Also goes by many names... e.g. **epistemic uncertainty**.

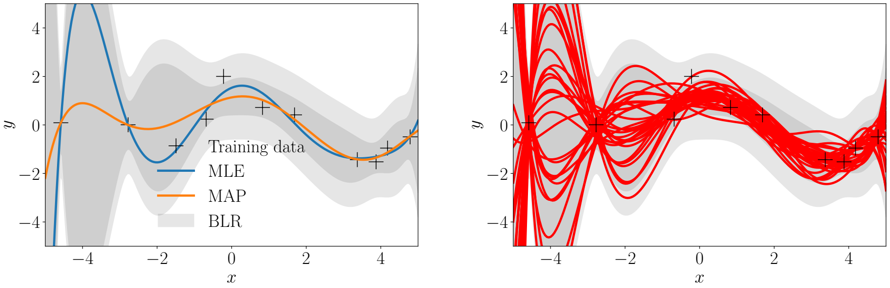
Quantifying Uncertainty with Bayesian Inference

If we knew that for a series of problems, our parameters $\boldsymbol{\theta}$ were sampled from $p(\boldsymbol{\theta})$, then Bayes' rule would give us the probability distribution after observing our data $\mathbf{y}$:

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|\boldsymbol{\theta})}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y})}_{\text{evidence}}} \tag{2}$$

- Allows us to quantify uncertatinty in parameters $\boldsymbol{\theta}$.
- Bayesian inference makes a leap of faith: Choose a prior and assume this is the correct one.
- Choosing priors is important $\implies$ Probabilistic Inference (Spring).

Example



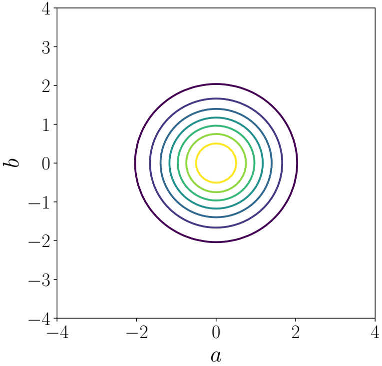
- Light-gray: uncertainty due to noise (aleatoric uncertainty / unpredictability)
- Dark-gray: uncertainty due to parameter uncertainty (epistemic uncertainty)
- Right: Plausible functions under the parameter distribution (every single parameter setting describes one function)

Distribution over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$

$$p(a, b) = \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right)$$



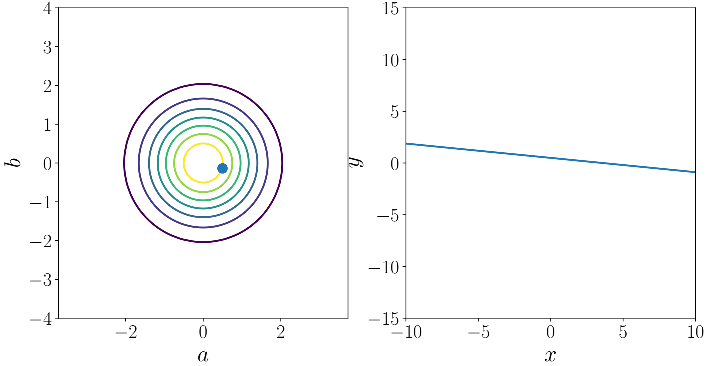
Sampling from the Prior over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$

$$p(a, b) = \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right)$$

$$f_i(x) = a_i + b_i x, \quad [a_i, b_i] \sim p(a, b)$$



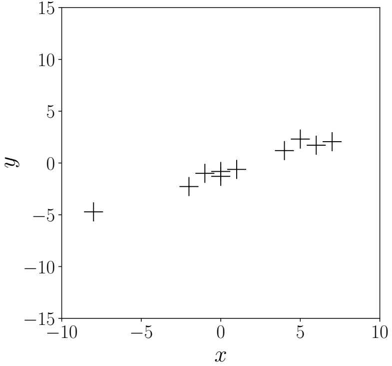
Sampling from the Posterior over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$

$$p(a, b) = \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right)$$

$$\boldsymbol{X} = [x_1, \ldots, x_N],\ \boldsymbol{y} = [y_1, \ldots, y_N] \quad \text{Training inputs/targets}$$



Sampling from the Posterior over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$

$$p(a, b) = \mathcal{N}\left(\mathbf{0},\, \boldsymbol{I}\right)$$

$$p(a, b|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{m}_N,\, \boldsymbol{S}_N\right) \qquad \text{Posterior}$$

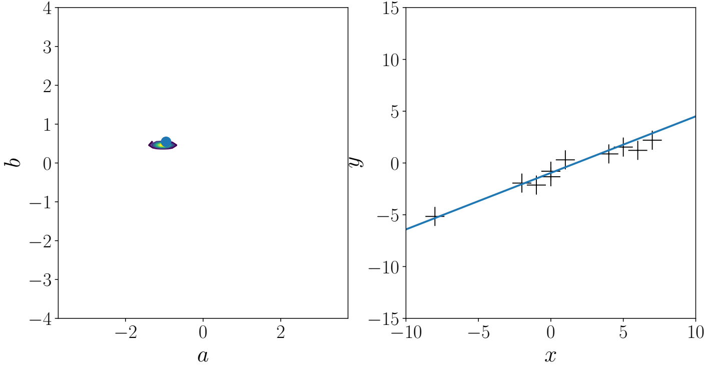
Sampling from the Posterior over Functions

Consider a linear regression setting

$$y = f(x) + \epsilon = a + bx + \epsilon\,, \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_n^2\right)$$

$$[a_i, b_i] \sim p(a, b|\boldsymbol{X}, \boldsymbol{y})$$

$$f_i = a_i + b_i x$$



Model: Bayesian Linear Regression

We never put a distribution on any $\mathbf{x}_n$, so we drop from conditioning.

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; 0, \mathrm{I}_M) \tag{3}$$

$$p(y_n|\boldsymbol{\theta}) = \mathcal{N}\left(y_n; \boldsymbol{\phi}(\mathbf{x}_n)^\top \boldsymbol{\theta}, \sigma^2\right) \tag{4}$$

$$p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y}; \Phi(\boldsymbol{X})\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_N\right) \tag{5}$$

Two goals:

- Find posterior over parameters $p(\boldsymbol{\theta}|\mathbf{y})$
- Find predictive posterior $p(\mathbf{y}^*|\mathbf{y})$

Posterior over Parameters

Board:

- Equating coefficients (tests your matrix algebra skills!)
- Joint Gaussian
- Woodbury identity

Method 1: Crunching densities

$$\log p(\boldsymbol{\theta}|\mathbf{y}) = \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{y}) \tag{6}$$

$$= c - \frac{1}{2\sigma^2}(\mathbf{y} - \Phi(\boldsymbol{X})\boldsymbol{\theta})^\top(\mathbf{y} - \Phi(\boldsymbol{X})\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} \tag{7}$$

This is a vector quadratic in $\boldsymbol{\theta}$! board $\implies$ Gaussian.

- Equate coefficients. Can rearrange... or find $\mathbb{E}$+$\mathbb{V}$ by other means
- Find maximum to find mean board
- Find Hessian to find covariance board

$$p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\theta}; \left[\frac{1}{\sigma^2}\Phi(X)^\top\Phi(X) + \boldsymbol{I}_M\right]^{-1}\frac{1}{\sigma^2}\Phi(X)^\top\mathbf{y}, \right. \tag{8}$$

$$\left. \left[\frac{1}{\sigma^2}\Phi(X)^\top\Phi(X) + \boldsymbol{I}_M\right]^{-1}\right) \tag{9}$$

Method 2: Joint Gaussian

Find

$$p(\boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\theta} \\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] \\ \mathbb{E}_{\mathbf{y}}[\mathbf{y}]\end{bmatrix}, \begin{bmatrix}\mathbb{V}[\boldsymbol{\theta}] & \mathbb{C}[\theta, \mathbf{y}] \\ \mathbb{C}[\mathbf{y}, \boldsymbol{\theta}] & \mathbb{V}[\mathbf{y}]\end{bmatrix}\right) \tag{10}$$

board

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) = \mathcal{N}\Big(&\boldsymbol{\theta}; \Phi(X)^\intercal\left[\Phi(X)\Phi(X)^\intercal + \sigma^2 \boldsymbol{I}_N\right]^{-1}\mathbf{y}, \\ &\boldsymbol{I}_M - \Phi(X)^\intercal\left[\Phi(X)\Phi(X)^\intercal + \sigma^2 \boldsymbol{I}_N\right]^{-1}\Phi(X)\Big) \end{aligned} \tag{11}$$

Computational Considerations

Typical algorithms (i.e. not optimal ones) take:

- $O(NM^2)$ to multiply matrices of shape $M \times N$ with $N \times M$

  (you must be able to derive this)

- $O(N^3)$ to find a matrix inverse

- The two results are certainly different in computational complexity!
- From joint is worse when $N \gg M$
- Are they different in value?

Woodbury identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1} \tag{12}$$

- Can go back and forth between the two forms.
- Allows you to implement the most efficient, based on setting of M and N.
- See exercise to practice.

Predictive posterior

- Crunching densities (see pdf)
- Equating coefficients (tests your matrix algebra skills!) (see pdf)
- Joint Gaussian (see exercise)
- May also need to apply the Woodbury identity

Method 1: Crunching densities

First, how to express our target in terms of densities we know.

$$p(\mathbf{y}^*|\mathbf{y}) \overset{\text{AT}}{=} \int \frac{p(\mathbf{y}^*, \boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} \mathrm{d}\boldsymbol{\theta} \tag{13}$$

$$\overset{\text{MA}}{=} \int p(\mathbf{y}^*|\boldsymbol{\theta}) \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})} \mathrm{d}\boldsymbol{\theta} \tag{14}$$

Next do the integrals / equating coefficients.

$\implies$ I recommend other method.

Method 2: Expectation identities

Conclusion

- Bayesian Linear Regression quantifies uncertainty due to lack of data (epistemic uncertainty)
- Gaussians are easy to deal with when conditioning