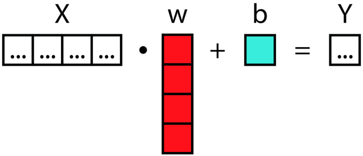
Introduction to Bayesian inference

June 4, 2019

Introduction

$$\boldsymbol{X} \cdot \boldsymbol{w} + \boldsymbol{b} = \boldsymbol{Y}$$



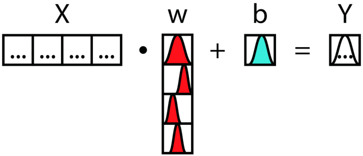
In **red** and **blue** : unknown parameters that we have to learn

- **Classical regression** :
  find each parameter that best explain the data

Image source : `ericmjl.github.io/bayesian-deep-learning-demystified`

Introduction

$$\boldsymbol{X} \cdot \boldsymbol{w} + \boldsymbol{b} = \boldsymbol{Y}$$



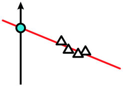
In **red** and **blue** : unknown parameters that we have to learn

- **Classical regression** :
  find each parameter that best explain the data
- **Going Bayesian** :
  Treat each parameter : random variable with other variable to be estimated.
  e.g : mean + standard deviation

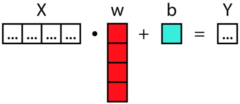
Image source : `ericmjl.github.io/bayesian-deep-learning-demystified`

Introduction

Non-Bayesian

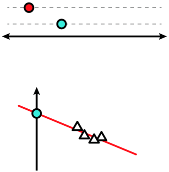


Inferred parameters:
1 slope + 1 intercept

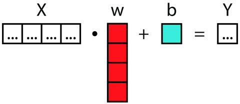


Introduction

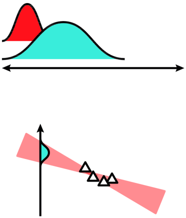
Non-Bayesian



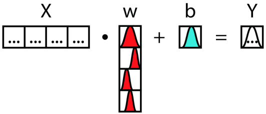
Inferred parameters:
1 slope + 1 intercept



Bayesian



Inferred parameters:
family of slopes + family of intercepts



$\hookrightarrow$ propagates the uncertainties

Image source : `ericmjl.github.io/bayesian-deep-learning-demystified`

The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

Required tools :

Sum rule:

$$P(x) = \sum_y P(x, y)$$

Product rule:

$$P(x, y) = P(x \mid y)P(y)$$

The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}$$

Required tools :

Sum rule:

$$P(x) = \sum_y P(x, y)$$

Product rule:

$$P(x, y) = P(x \mid y)P(y)$$

The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters}) P(\text{parameters})}{P(\text{data})}$$

- $P(\text{data} \mid \text{parameters})$: **likelihood** of a set of parameters in a given model

The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}$$

- $P(\text{data} \mid \text{parameters})$: **likelihood** of a set of parameters in a given model
- $P(\text{parameters})$: **prior** probability of the parameters

The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}$$

- $P(\text{data} \mid \text{parameters})$: **likelihood** of a set of parameters in a given model
- $P(\text{parameters})$: **prior** probability of the parameters
- $P(\text{parameters} \mid \text{data})$: **posterior** probability of the parameter given the data

The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}$$

- $P(\text{data} \mid \text{parameters})$: **likelihood** of a set of parameters in a given model
- $P(\text{parameters})$: **prior** probability of the parameters
- $P(\text{parameters} \mid \text{data})$: **posterior** probability of the parameter given the data
- $P(\text{data})$: **evidence** of the data

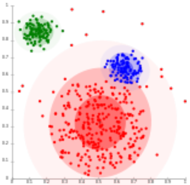
The Bayes rule

How to do inference about hypothesis from data ?

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters}) P(\text{parameters})}{P(\text{data})}$$

Learning : Use the data and the modelling assumption to transform what I knew before the data (prior) $\rightarrow$ gives the posterior

- classification / clustering (yes/no category) (group similar things)



source : wikipedia

- regression (predict values)

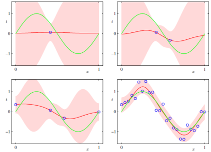


Figure 3.8 Examples of the predictive distribution (3.58) for a model consisting of 9 Gaussian basis functions of the form (3.4) using the synthetic sinusoidal data set of Section 1.1. See the text for a detailed discussion.

source : Bishop, Pattern Recognition And Machine Learning

The Bayes rule

Bayes' rule is a way to infer parameter given underlying data
$\hookrightarrow$ Bayesian machine learning is nothing more than learning a probability distribution for each parameter

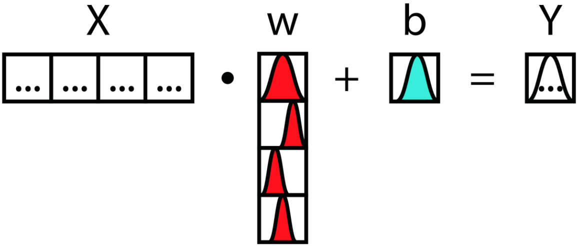


Image source : `ericmjl.github.io/bayesian-deep-learning-demystified`

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$
- Specify prior believes we have about the parameters θ

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$
- Specify prior believes we have about the parameters θ
- Observe the data

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$
- Specify prior believes we have about the parameters θ
- Observe the data
- Compute the posterior $P(\theta \mid \text{data})$

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$
- Specify prior believes we have about the parameters θ
- Observe the data
- Compute the posterior $P(\theta \mid \text{data})$
- Repeat for multiple models/parameter

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$
- Specify prior believes we have about the parameters θ
- Observe the data
- Compute the posterior $P(\theta \mid \text{data})$
- Repeat for multiple models/parameter
- See which model(s)/parameter(s) fit for your data.

Bayesian Machine Learning (in one slide)

- Have a model $\mathcal{M}(\theta)$
- Specify prior believes we have about the parameters θ
- Observe the data
- Compute the posterior $P(\theta \mid \text{data})$
- Repeat for multiple models/parameter
- See which model(s)/parameter(s) fit for your data.

Linear regression

Example with ℓ_2 cost (linear least square problem) :

The objective function is :

$$f = \frac{1}{2}\|y - \hat{y}\|_2^2$$

with ℓ_2 regularization :

$$f = \frac{1}{2}\|y - \hat{y}\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$$

with the model $\hat{y} = Xw$

Linear regression

Frequentist statistics point of view :

$$y = \hat{y} + \epsilon$$

Let's model $\hat{y}$ as a Gaussian random variable : $\hat{y} \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = \hat{y} = Xw$ (the prediction of the model).

$$P(y \mid X, w, \sigma^2) = \mathcal{N}(Xw, \sigma^2) \rightarrow \text{likelihood} \tag{1}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - Xw)^2}{2\sigma^2}\right) \tag{2}$$

Linear regression

Frequentist statistics point of view :

$$y = \hat{y} + \epsilon$$

Let's model $\hat{y}$ as a Gaussian random variable : $\hat{y} \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = \hat{y} = Xw$ (the prediction of the model).

$$\begin{aligned} P(y \mid X, w, \sigma^2) &= \mathcal{N}(Xw, \sigma^2) \rightarrow \text{likelihood} && (1) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - Xw)^2}{2\sigma^2}\right) && (2) \end{aligned}$$

Maximum Likelihood Estimate : $\boxed{w_{\text{MLE}} = \arg\max \mathcal{N}(Xw, \sigma^2)}$

Linear regression

Frequentist statistics point of view :

$$y = \hat{y} + \epsilon$$

Let's model $\hat{y}$ as a Gaussian random variable : $\hat{y} \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = \hat{y} = Xw$ (the prediction of the model).

$$P(y \mid X, w, \sigma^2) = \mathcal{N}(Xw, \sigma^2) \rightarrow \text{likelihood} \quad (1)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - Xw)^2}{2\sigma^2}\right) \quad (2)$$

Maximum Likelihood Estimate : $\boxed{w_{\text{MLE}} = \arg\max \mathcal{N}(Xw, \sigma^2)}$

or, minimizing the neg-log likelihood :

$$\boxed{w_{\text{MLE}} = \arg\min (y - Xw)^2}$$

$\hookrightarrow$ MLE on Gaussian Likelihood is equivalent to the least squares

Linear regression

Bayesian point of view

We introduce the prior and then maximize the posterior:

$$\underbrace{P(w \mid y, X)}_{\text{posterior}} \propto \underbrace{P(y, X, w)}_{\text{likelihood}} \underbrace{P(w \mid \mu_w, \sigma_w^2)}_{\text{prior}}$$

Linear regression

Bayesian point of view

We introduce the prior and then maximize the posterior:

$$\underbrace{P(w \mid y, X)}_{\text{posterior}} \propto \underbrace{P(y, X, w)}_{\text{likelihood}} \underbrace{P(w \mid \mu_w, \sigma_w^2)}_{\text{prior}}$$

Gaussian prior for w : $P(w \mid \mu_w, \sigma_w^2) = \mathcal{N}(0, \sigma_0)$,

$$P(w \mid y, X) \propto \exp\left(-\frac{(y - Xw)^2}{\sigma^2}\right) \exp\left(-\frac{(w - \mu_w)^2}{\sigma_w^2}\right)$$

Linear regression

Bayesian point of view

We introduce the prior and then maximize the posterior:

$$\underbrace{P(w \mid y, X)}_{\text{posterior}} \propto \underbrace{P(y, X, w)}_{\text{likelihood}} \underbrace{P(w \mid \mu_w, \sigma_w^2)}_{\text{prior}}$$

Gaussian prior for w : $P(w \mid \mu_w, \sigma_w^2) = \mathcal{N}(0, \sigma_0)$,

$$P(w \mid y, X) \propto \exp\left(-\frac{(y - Xw)^2}{\sigma^2}\right)\exp\left(-\frac{(w - \mu_w)^2}{\sigma_w^2}\right)$$

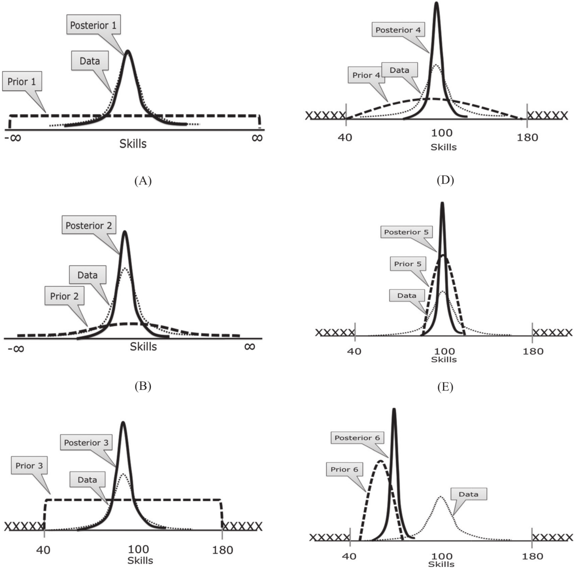
Then, minimize the neg-log posterior writes:

$$w_{\text{MAP}} = \arg\min \|\hat{y} - Xw\|_2^2 + \lambda\|w\|_2^2$$

$\hookrightarrow$ Gaussian prior leads to ℓ_2 regularization.

- *Note 1: if the prior on w is uniform (non informative prior), Maximum a Posteriori = Maximum Likelihood estimate*
- *Note 2: good informative prior $\rightarrow$ efficient regularization*

How the prior drives the posterior



from van de Schoot *et al.*, 2014

Norms vs Priors

- The median minimize the L1 norm :

$$\text{median}(x) = \arg\min_s \sum_i \|x_i - s\|_1$$

$\hookrightarrow$ least absolute deviation estimate = maximum likelihood estimate with errors having a Laplace distribution
(fat-tailed distribution $\rightarrow$ less prone to outliers)

Norms vs Priors

- The median minimize the L1 norm :

$$\text{median}(x) = \arg\min_s \sum_i \|x_i - s\|_1$$

$\hookrightarrow$ least absolute deviation estimate = maximum likelihood estimate with errors having a Laplace distribution
(fat-tailed distribution $\rightarrow$ less prone to outliers)

- The mean minimize the L2 norm :

$$\text{mean}(x) = \arg\min_s \sum_i \|x_i - s\|_2$$

$\hookrightarrow$ least squares estimate = ML estimate with Gaussian errors (more sensitive to outliers)

Norms vs Priors

- The median minimize the L1 norm :

$$\text{median}(x) = \arg\min_s \sum_i \|x_i - s\|_1$$

$\hookrightarrow$ least absolute deviation estimate = maximum likelihood estimate with errors having a Laplace distribution
(fat-tailed distribution $\rightarrow$ less prone to outliers)

- The mean minimize the L2 norm :

$$\text{mean}(x) = \arg\min_s \sum_i \|x_i - s\|_2$$

$\hookrightarrow$ least squares estimate = ML estimate with Gaussian errors (more sensitive to outliers)

Gauss proved the central limit theorem $\rightarrow$ justify the use of least squares

Hierarchical Bayesian models

Example:

$$y \mid \theta \sim \mathcal{N}(\theta, 1)$$

Hierarchical Bayesian models

Example:

$$y \mid \theta \sim \mathcal{N}(\theta, 1)$$

θ is a parameter of the model. It can have its own distribution (the prior):

$$\theta \mid \mu \sim \mathcal{N}(\mu, 1)$$

Hierarchical Bayesian models

Example:

$$y \mid \theta \sim \mathcal{N}(\theta, 1)$$

θ is a parameter of the model. It can have its own distribution (the prior):

$$\theta \mid \mu \sim \mathcal{N}(\mu, 1)$$

μ is called an hyperparameter. It can also have its own distribution (the hyperprior)

$$\mu \sim \mathcal{N}(0, 1)$$

Hierarchical Bayesian models

Example:

$$y \mid \theta \sim \mathcal{N}(\theta, 1)$$

θ is a parameter of the model. It can have its own distribution (the prior):

$$\theta \mid \mu \sim \mathcal{N}(\mu, 1)$$

μ is called an hyperparameter. It can also have its own distribution (the hyperprior)

$$\mu \sim \mathcal{N}(0, 1)$$

The full posterior is then:

$$\begin{aligned} P(\theta, \mu \mid y) &\propto P(y \mid \theta, \mu) P(\theta, \mu) \\ &\propto P(y \mid \theta) P(\theta \mid \mu) P(\mu) \\ &\propto \mathcal{N}(\theta, 1)\mathcal{N}(\mu, 1)\mathcal{N}(0, 1) \end{aligned}$$

Hierarchical Bayesian models

Example:

$$y \mid \theta \sim \mathcal{N}(\theta, 1)$$

θ is a parameter of the model. It can have its own distribution (the prior):

$$\theta \mid \mu \sim \mathcal{N}(\mu, 1)$$

μ is called an hyperparameter. It can also have its own distribution (the hyperprior)

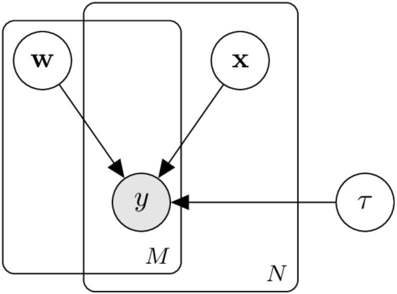
$$\mu \sim \mathcal{N}(0, 1)$$

The full posterior is then:

$$\begin{aligned} P(\theta, \mu \mid y) &\propto P(y \mid \theta, \mu) P(\theta, \mu) \\ &\propto P(y \mid \theta) P(\theta \mid \mu) P(\mu) \\ &\propto \mathcal{N}(\theta, 1)\mathcal{N}(\mu, 1)\mathcal{N}(0, 1) \end{aligned}$$

If the full posterior does not have a closed-form, it can be approximated by numerical methods such as Monte Carlo Markov Chains.

Hierarchical Bayesian models

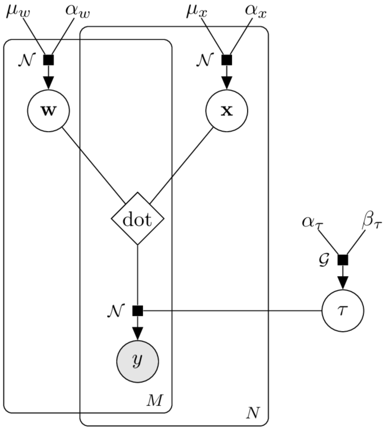


Read :

$$y_{m,n} = f_1(w_m)\ [+ \text{or} \times]\ f_2(x_n)\ [+ \text{or} \times]\ f_3(\tau)$$

from `github.com/jluttine/tikz-bayesnet`

Hierarchical Bayesian models



Read :

$$y_{m,n} = f_1(w_m) \times f_2(x_n) + f_3(\tau)$$

$\hookrightarrow$ Probabilistic Principal Component Analysis model
from `github.com/jluttine/tikz-bayesnet`

Related topics

- Machine learning
- Pattern recognition
- Neural networks and deep learning
- Data mining / Data science
- Statistic modeling
- Artificial intelligence

For different fields:

- Engineering (signal processing, system identification, ...)
- Computer Science
- Statistics (data science, estimation,...)
- Cognitive science and psychology (perception, linguistics,...)
- Economics (decision theory, game theory, e-commerce...)

References

Book : *Pattern Recognition and Machine Leaning*, Bishop

New article : *Machine learning in acoustics: a review*, Bianco & coaut.