Workshop 2

COMP90051 Statistical Machine Learning

Semester 1, 2023

About your tutor

Agenda

1. Icebreaker
2. Python ecosystem for ML
3. Refresher: Bayes’ theorem
4. Worksheet on Bayesian inference

Icebreaker

Learning outcomes

At the end of this workshop you should:

- be familiar with the Python ecosystem for machine learning
- develop intuition about the role of prior and posterior in Bayesian inference

Is your system ready to go?

- You should have installed Anaconda on your system before today's workshop. If not, please install it now.
- Anaconda is a Python distribution tailored for scientific computing
- Most of the packages we need are installed by default
- Worksheets will be distributed as Jupyter Notebooks

Top 5 libraries for beginners to master

NumPy

- Library for working with large multidimensional arrays
- High-level functions for arrays

pandas

- Library for analysis and manipulation of tabular data
- Provides similar functionality to DataFrames and dplyr in R

scikit learn

- Machine learning library
- Includes implementations of most models covered in this course (exception: neural nets)

SciPy

- Scientific computing library
- Functionality includes: statistics/random number generation, linear algebra, optimisation, special functions, integration

matplotlib

- 2D plotting library
- Provides similar interface to MATLAB

We’ll see some of these libraries later…

TensorFlow

mxnet

Keras

PYTORCH

Deep learning frameworks

Probabilistic programming frameworks

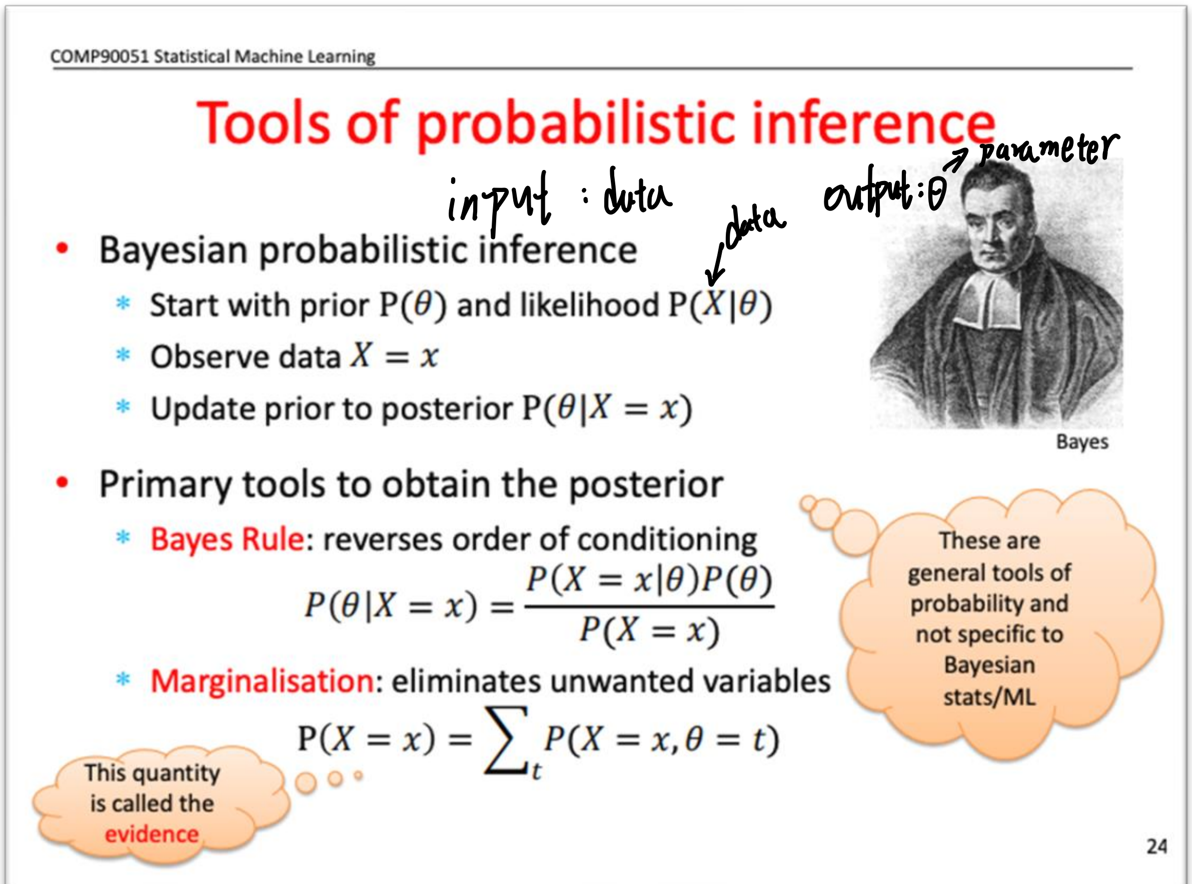
PyStan

PyMC3

PYRO

Bayesian inference

Recall from Lecture 2



- The likelihood $P(X = x|\theta)$ is the conditional probability of the data $X = x$ as a function of θ.
- The prior $P(\theta)$ represents information we have that is not part of the collected data $X = x$.
- The evidence $P(X = x)$ is the average over all possible values of theta.
- $P(\theta|X = x)$ is the posterior distribution, which represents our updated beliefs under our prior $P(\theta)$ now we have observed the data $X = x$.

by data x_1, x_2, x_3

prior: $p(\theta) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$

likelihood: $p(x_1|\theta) = \theta^{x_1}(1-\theta)^{1-x_1} =$

$p(\theta | x_1) = \frac{p(x_1|\theta) \cdot p(\theta)}{p(x_1)} \propto p(x_1|\theta) \cdot p(\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$

Worksheet 2

$\propto \theta^{\underbrace{x_1+a-1}_{a'}} (1-\theta)^{\underbrace{1-x_1+b-1}_{b'}}$

$p(\theta | x_1, \dots, x_n) \propto p(x_1, \dots, x_n | \theta) \, p(\theta)$

Every example is independent

$= \prod_{i=1}^{n} p(x_i|\theta) \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$

$= \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}$

$$= \theta^{n_H} (1-\theta)^{n-n_H} \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$= \theta^{(n_H + a - 1)} (1-\theta)^{(n - n_H + b - 1)}$$

(MAP) maximum a posteriori probability n_H # of heads.

~~pos~~ posterior is Beta

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid x_1, \cdots, x_n)$$

$\frac{\partial L}{\partial \theta}$ 对后验求导

$$= \frac{n_H + a - 1}{n + a + b - 2}$$

MLE only based on likelihood not prior.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(x_1, \cdots, x_n \mid \theta) = \frac{n_H}{n}$$

对likelihood求导

实验次数↑ → coverge to together

解 $\frac{n_H}{n}$ 逐渐趋同

$a=1$ $b=1$ 时 两个一样 MAP与MLE identical

uniform dist → we don't have prior info about data.

数据越多 ⇒ Beta 越 more concentrated → more data ⇒ more infor used to update posterior ⇒ more confident to estimation

$$P(\theta) = \frac{1}{Beta(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$P(x_1|\theta) = \theta^{x_1} (1-\theta)^{1-x_1}$$

$$P(x_2|\theta) = \theta^{x_2} (1-\theta)^{1-x_2}$$

$$P(\theta|x_1, x_2) \propto P(x_1|\theta) P(x_2|\theta) P(\theta)$$

$$= \theta^{x_1+x_2} (1-\theta)^{1-x_1+1-x_2} \times \theta^{a-1} (1-\theta)^{b-1}$$

$$= \theta^{(x_1+x_2+a)-1} (1-\theta)^{(2-x_1-x_2)+b-1}$$

$$= \theta^{(x_1+x_2+a)-1} (1-\theta)^{(2-x_1-x_2+b)-1}$$

$$Beta(x_1+x_2+a,\ 2-x_1-x_2+b)$$

$$\prod_{i=1}^{n} P(x_i|\theta) P(\theta)$$

$$\propto \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \theta^{a-1} (1-\theta)^{b-1}$$

$$\propto \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \theta^{a-1} (1-\theta)^{b-1}$$

$$\propto = \underbrace{\theta^{(\underset{\downarrow\, n_H}{\sum x_i}+a)-1} (1-\theta)^{(\underset{\downarrow\, n_T}{n-\sum x_i}+b)-1}}$$

正面的次数

derivation for MAP: $\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta | x_1, \cdots, x_n)$

$$= \frac{n_H + a - 1}{n + a + b - 2}$$

MLE: $\hat{\theta}_{MLE} = \arg\max_{\theta} P(x_1, \cdots, x_n | \theta)$

$$= \frac{n_H}{n}$$

MAP: $P(\theta / x_1, \cdots, x_n) = \theta^{\Sigma x_i + a - 1} (1-\theta)^{n - \Sigma x_i + b - 1}$

logarithm: $\log P(\theta | x_1, \cdots, x_n) = (\Sigma x_i + a - 1) \log(\theta) + (n - \Sigma x_i + b - 1) \log(1-\theta)$

derivative: and equal to zero $\frac{\partial L}{\partial \theta} = \frac{\Sigma x_i + a - 1}{\theta} - \frac{n - \Sigma x_i + b - 1}{1-\theta} = 0$

$$\theta = \frac{n_H + a - 1}{n + a + b - 2}$$

MLE: $P(x_1 \cdots x_n | \theta) = \prod_{i=1}^{n} \theta^{x_1} (1-\theta)^{1-x_1}$

log: $\Sigma x_i \log \theta + \Sigma (1 - x_i) \log(1-\theta)$

$$\frac{\partial \log P(x_1, \cdots, x_n | \theta)}{\partial \theta} = \frac{\Sigma x_i}{\theta} - \frac{\Sigma(1-x_i)}{1-\theta} = 0$$

$$\Sigma_{i=1}^{n} x_i - \Sigma x_i \theta = (n - \Sigma x_i)\theta.$$

$$n\theta = \Sigma x_i$$

$$\theta = \frac{\Sigma x_i}{n}$$

数据越多，更新越多，后验分布越集中 (more concentrated