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

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

Icebreaker

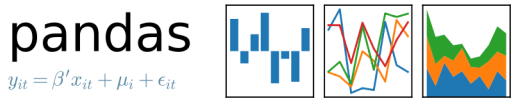
Is your system ready to go?

- 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

Icebreaker

Top 5 libraries for beginners to master

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



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

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

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

- 2D plotting library
- Provides similar interface to MATLAB

We'll see some of these libraries later…

Deep learning frameworks

Probabilistic programming frameworks

Bayesian inference

Recall from Lecture 2

Tools of probabilistic inference

- Bayesian probabilistic inference
  - Start with prior $\mathrm{P}(\theta)$ and likelihood $\mathrm{P}(X|\theta)$
  - Observe data $X = x$
  - Update prior to posterior $\mathrm{P}(\theta|X = x)$

Bayes

- Primary tools to obtain the posterior
  - Bayes Rule: reverses order of conditioning

$$P(\theta|X = x) = \frac{P(X = x|\theta)P(\theta)}{P(X = x)}$$

  - Marginalisation: eliminates unwanted variables

$$\mathrm{P}(X = x) = \sum_t P(X = x, \theta = t)$$

This quantity is called the evidence

These are general tools of probability and not specific to Bayesian stats/ML

24

- 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 $X = x$.

Worksheet 2