Fitting models to data

Bayesian Probabilistic Programming

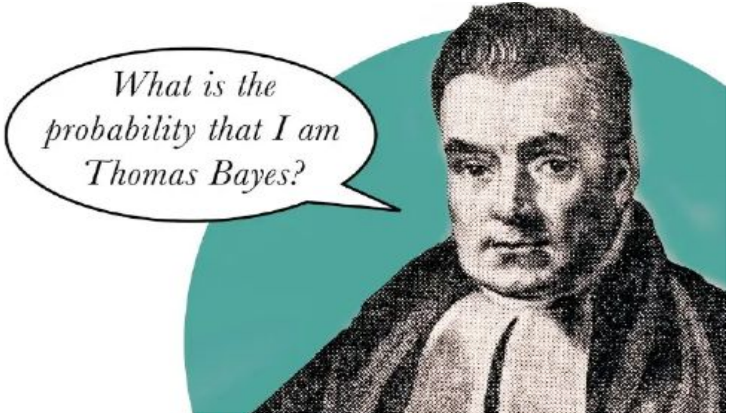
PyMC3

Hitesh Lala, ARI

emcee
The MCMC Hammer

Background

- Basic Statistics: Probability, Conditional Probability, Probability Distributions
- Bayes' Theorem
- Python conversance

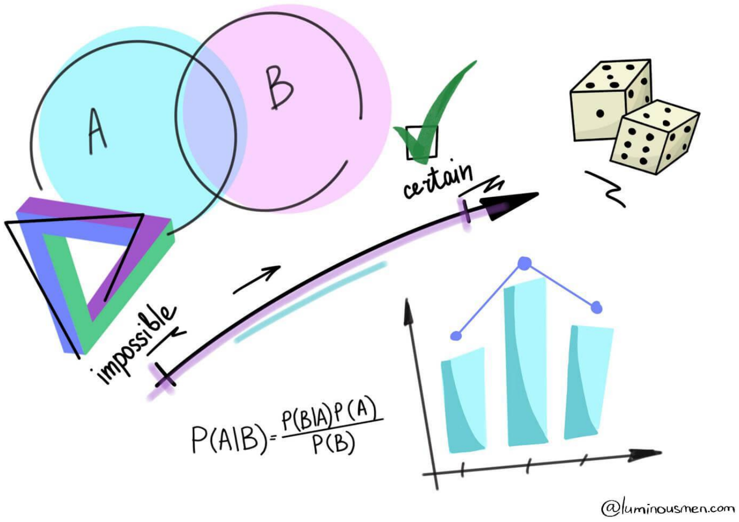


Bayesian Statistics

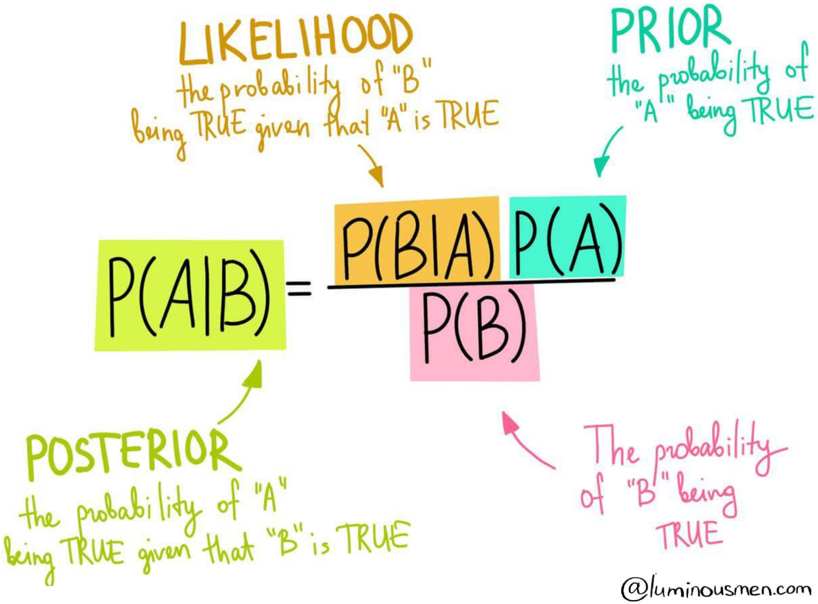
- Interpretation:

  Quantifying Personal Belief (Subjectivity)

- Methodology:
  - Set up Random Variables (incorporating uncertainty)
  - Determine Prior Probability
  - Compute Posterior Probability using Bayes' Theorem



Bayes' Theorem



Published in [1763]

Bayes' Theorem

$$\overbrace{p(\theta|D)}^{\text{posterior distribution}} = \frac{\overbrace{p(D|\theta)}^{\text{likelihood function}}\ \overbrace{p(\theta)}^{\text{prior distribution}}}{\int p(D|\theta)\, p(\theta)\, d\theta}$$

D : Data

Θ : Model Parameters

Likelihood

- Generative Probabilistic Model
- A functional form

Evidence

- No one cares about the evidence
- Really hard to calculate
- Might not be possible to calculate in practical settings(?)

Fit a line to data

a.k.a. Linear Regression

$$Y = \alpha + (\beta \times X) + \sigma$$

MCMC

- Markov Chain Monte Carlo
- Metropolis Hastings
- Hamiltonian Monte Carlo (HMC)
- No U-turn Sampler (NUTS)

Visualize

pymc3 vs emcee

pymc3	emcee
Provides a plethora of samplers - NUTS, etc.	Ensemble Sampler with Affine Invariance
Much more than just MCMC - Neural Nets, etc.	Expedited sampling of non-differentiable distributions
Inefficient sampler can be slow	Embarrassingly fast in low-dimensions
HMC-based efficient handling of large spaces	Slower in higher dimensions
Out-of-the-box handling	No built-in distribution

OLS vs Robust Regression

- Using a Student’s-T distribution as the likelihood
- Fatter tails as compared to a Normal distribution
- Immune to Outliers

Regression is a powerful tool for forecasting.
Economists using it successfully predicted ten
out of the last two recessions.

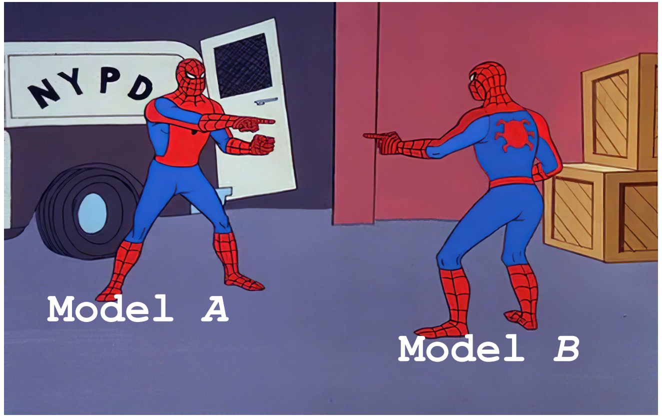
Model Selection

- LOO (Leave-one-out Cross-Validation)
- WAIC (Widely Applicable Information Criterion)
- Posterior Predictive Checks

> *Inference is normal science.*
>
> *Model–checking is revolutionary science.*
>
> Andrew Gelman

Model Comparison

- Kullback-Leibler Divergence - Mutual Information - Relative Entropy
- Null Hypothesis Significance Testing - T-test, F-test, χ^2 - test
- BEST (Bayesian Estimation Supersedes the T-Test)

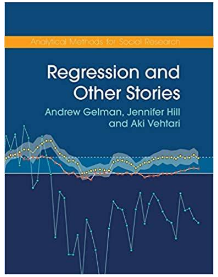


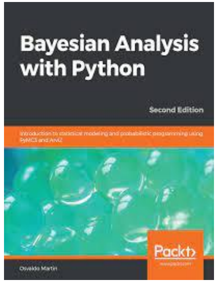
Beyond *just* fitting a line

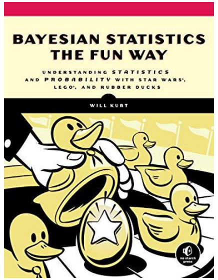
- Outlier Detection
- Fitting a plane (or an n-dimensional hyperplane) to data
- Non-Linear Regression
- Gaussian Process Regression

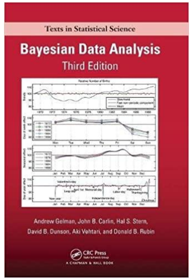
Resources

- Andrew Gelman, Jennifer Hill, Aki Vehtari - Regression and Other Stories (2020)
- Oswaldo Martin - Bayesian Analysis with Python (2018)
- Will Kurt - Bayesian Statistics the Fun Way: Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks
- Andrew Gelman et. al. - Bayesian Data Analysis 3rd ed. (2020)
- Thomas Wiecki - An Intuitive Guide to Bayesian Statistics
- David W. Hogg et. al. - Data analysis recipes: Fitting a model to data
- Jake VanderPlas - Frequentism and Bayesianism: A Practical Introduction









Fin.

Thank you.