Bayesian Analysis Course

A quick recap

Vasilis Gkolemis

ATHENA RC — HUA

June 2025

Program

1. Course Recap
2. Real-World Applications
3. Tools and Libraries

Course objective

Be able to apply probabilistic models to real-world problems.

Course objective

Be able to apply ~~probabilistic~~ models to real-world problems.
Be able to apply Bayesian models to real-world problems.

Session 1: Probabilistic Modeling and Reasoning

Key Topics

- The Role of Uncertainty in Machine Learning
- Probabilistic Modeling and Reasoning
- Application on the Altzheimer Test
- Key Rules of Probability

Modeling Uncertainty

Key Topics

- Random Variables and Probability Distributions
- Basic Properties of Random Variables
  - Expectation $\mathbb{E}[X]$
  - Variance $\mathrm{Var}(X)$
  - Sampling to approximate them
    - $\mathbb{E}[X] \approx \frac{1}{N}\sum_{i=1}^{N} x_i$
    - $\mathrm{Var}(X) \approx \frac{1}{N}\sum_{i=1}^{N}(x_i - \mathbb{E}[X])^2$
- Common Probability Distributions
  - Bernoulli, Binomial, Poisson, Gaussian
  - Multivariate Gaussian: $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Random Variables are the Building Blocks

Session 3: Bayesian Modeling — A Unified Framework

Key Topics

- Probabilistic vs. Statistical vs. Bayesian Models
  - Probabilistic Model: a known probability distribution $p(x, y)$
  - Statistical Model: a model with unknown parameters $\boldsymbol{\theta}$, e.g., $p(x, y; \boldsymbol{\theta})$
    - Defines a set of probabilistic models: $\{p(x, y; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$
  - Bayesian Model: a statistical model with a prior distribution $p(\boldsymbol{\theta})$
- Prior $p(\boldsymbol{\theta})$
  - Our beliefs about the parameters before observing data
- Likelihood $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
  - How likely is the data given the parameters
- Posterior $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$
  - Our updated beliefs about the parameters after observing data
- Predictive Posterior Distribution $p(y|\boldsymbol{x}, \mathcal{D})$
  - Predict on new inputs $\boldsymbol{x}$

Session 4: Bayesian Linear Regression

Key Topics

- Linear Regression as a Probabilistic Model
  - Model: $y = \boldsymbol{x}^T\boldsymbol{\theta} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$
  - Likelihood: $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$
- Exact Inference with Conjugate Priors
  - Some prior–likelihood pairs lead to closed-form solutions
  - Conjugate Prior: A prior that, when combined with a likelihood, results in a posterior of the same family
    - Example: Gaussian likelihood with Gaussian prior
- Posterior Distribution
  - Posterior: $p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is Gaussian
  - Close-form solutions are available:
    - $\boldsymbol{\Sigma} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^{-2}\boldsymbol{I})^{-1}$
    - $\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{X}^T\boldsymbol{y}$

Session 5: Bayesian Logistic Regression

Key Topics

- Logistic Regression as a Probabilistic Model
  - Model: $p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = \sigma(\boldsymbol{x}^T\boldsymbol{\theta})$, where $\sigma(z) = \frac{1}{1+e^{-z}}$
- Approximate Inference Methods:
  - Laplace Approximation
    - Gaussian centered at the MAP estimate
    - Hessian used to approximate the curvature
  - Importance Sampling
    - Samples from a proposal distribution $q(\boldsymbol{\theta})$
    - Weights are computed as $w(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})}{q(\boldsymbol{\theta})}$
  - Markov Chain Monte Carlo (MCMC)
    - Samples from the posterior distribution directly
    - Examples: Metropolis-Hastings, Gibbs Sampling

Session 6: Putting It All Together

Key Topics

- Let's solve a Real-World Problem with Bayesian Modeling
  - Application of all concepts learned in the course
- Two main datasets:
  - Bike Sharing Dataset
    - Predict the number of bike rentals based on weather and time features
  - Boston Housing Dataset
    - Predict house prices based on various features

From Theory to Practice

Program

1. Course Recap
2. Real-World Applications
3. Tools and Libraries

Dataset: Bike Sharing

What is it?

- Collected by Capital Bikeshare in Washington, D.C.
- Contains daily and hourly counts of bike rentals.
- Includes contextual and weather information.

Key Features

- **datetime**: Date and hour.
- **season**: Winter, spring, summer, fall.
- **holiday**: Whether the day is a holiday.
- **workingday**: Is it a workday?
- **weather**: Clear, mist, rain, snow.
- **temp**, **atemp**: Temperature, perceived temperature.
- **humidity**, **windspeed**

Bike Sharing: Prediction Goal

Prediction Task

- Predict the **total rental count** (`count`) for a given day or hour.
- Understand how weather, seasonality, and holidays affect demand.

Apply a Bayesian linear regression model.

Dataset: Boston Housing

What is it?

- Classic dataset from the 1970s housing data for Boston suburbs.
- Collected by the U.S. Census Service.
- Widely used for regression tasks.

Key Features:

- Socio-economic indicators such as crime rate, income levels, and education.
- Housing characteristics like the average number of rooms and age of buildings.
- Environmental factors including air pollution levels and proximity to the Charles River.
- Accessibility measures such as distances to employment centers and highways.
- Local taxation and zoning information.

Boston Housing: Prediction Goal

Prediction Task

- Predict **MEDV**: Median value of owner-occupied homes ($1000s).
- Explore how socio-economic and environmental factors affect house prices.

Apply a Bayesian linear regression model.

Program

1. Course Recap
2. Real-World Applications
3. Tools and Libraries

Useful R Packages for Bayesian Modeling

Key R packages for Bayesian modeling and inference:

- **rstan** — Interface to Stan for Bayesian inference using MCMC.
- **brms** — Bayesian regression models using Stan; user-friendly formula syntax.
- **bayesplot** — Flexible plotting of posterior distributions and diagnostics.
- **tidybayes** — Tidy data tools for working with Bayesian models.
- **coda** — Tools for MCMC output analysis and diagnostics.