Probabilistic and Statistical Modeling

Vasilis Gkolemis

ATHENA RC — HUA

June 2025

Contents

Helping Material

- **Primer on Probabilistic Modeling**
  https://www.inf.ed.ac.uk/teaching/courses/pmr/22-23/assets/notes/probabilistic-modelling-primer.pdf

Session 1 – Recap

- **What we covered:**
  - **Probabilistic Modeling:** Model the world using probabilities
  - **Probabilistic Reasoning (Inference):** Use knowns to infer unknowns
  - **Bayesian Analysis:** Modeling and Reasoning with Bayes' rule.
  - **Core Rules of Probability:** The sum, product and Bayes' rule.
  - *Example: Alzheimer's diagnostic test.*

- **What's still to explore:**
  - Our example was simple
    - X: the test result — a $1D$ random variable in $\{0, 1\}$
    - Y: the disease status — a $1D$ random variable in $\{0, 1\}$
  - Real-world problems are more complex
    - Involve high-dimensional random variables
    - Involve complex relationships between variables
  - How can we model these complexities?
    - Session 2 extended our probabilistic toolbox.
    - Session 3 will show how to use it in practice.

Session 2 – Recap

- **What we covered:**
  - **Multivariate Random Variables and Distributions:**
    - PDFs, PMFs and CDFs
    - Key properties: expectation and variance.
    - How to sample from these distributions.
    - Key-distributions: Bernoulli, Normal, Poisson.
  - We now have powerful tools to model complexity!

- **What's still to explore:**
  - A glue to connect our probabilistic tools for performing analysis.
  - A *principled* and *unified* way to:
    - Model complex relationships between variables
    - Infer unknowns from knowns
    - Make predictions about future observations
  - The **Bayesian framework** is (among others) a powerful glue for this.

Session 3 – Overview

- **What we'll explore:**
  - The Bayesian Framework with each key components:
    - ★ **Prior Distribution:** Our belief before seeing the data.
    - ★ **Likelihood:** How compatible is the observed data is with different parameter values.
    - ★ **Posterior Distribution:** Our updated beliefs after observing the data.
    - ★ **Predictive Distribution:** Make predictions about new, unseen data.
  - How to use the Bayesian framework for predictive tasks.

- **Be confident. You already know important stuff:**
  - Session 1:
    - ★ Intuition about Bayesian modeling $\rightarrow$ Alzheimer's test case
    - ★ Core probability rules: sum, product, and Bayes' rule
  - Session 2:
    - ★ Multivariate random variables and distributions
    - ★ Key properties: expectation and variance
    - ★ How to sample from these distributions

Models

- The term "model" has multiple meanings, see e.g. https://en.wikipedia.org/wiki/Model
- Let's distinguish between three types of models:
  - probabilistic model
  - (parametric) statistical model
  - Bayesian model
- Note: the three types are often confounded, and often just called probabilistic or statistical model, or just "model".
- Introduction to Probabilistic Modelling $\rightarrow$ for further reading.

Probabilistic model

- From first lecture:

  A probabilistic model is an abstraction of reality that uses probability theory to quantify the chance of uncertain events.

- Example from the first lecture: cognitive impairment test
  - Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
  - Probabilistic model for presence of impairment ($x = 1$) and detection by the test ($y = 1$):
    - ⋆ $P(x = 1) = 0.11$ (prior)
    - ⋆ $P(y = 1 \mid x = 1) = 0.8$ (sensitivity)
    - ⋆ $P(y = 0 \mid x = 0) = 0.95$ (specificity)

Probabilistic model

- More technically:

  probabilistic model $\equiv$ *probability distribution (pmf/pdf).*

- Probabilistic model was written in terms of the probability P.
- In terms of the pmf it is:
  - $p_x(1) = 0.11$
  - $p_{y|x}(1 \mid 1) = 0.8$
  - $p_{y|x}(0 \mid 0) = 0.95$
- Commonly written as:
  - $p(x = 1) = 0.11$
  - $p(y = 1 \mid x = 1) = 0.8$
  - $p(y = 0 \mid x = 0) = 0.95$
- where the notation for probability measure P and pmf p are confounded.

Statistical model

- If we substitute the numbers with parameters, we obtain a (parametric) statistical model:
  - $p(x = 1) = \theta_1$
  - $p(y = 1 \mid x = 1) = \theta_2$
  - $p(y = 0 \mid x = 0) = \theta_3$
- For each value of the θ_i, we obtain a different pmf.
- Dependency highlighted by writing:
  - $p(x = 1; \theta_1) = \theta_1$
  - $p(y = 1 \mid x = 1; \theta_2) = \theta_2$
  - $p(y = 0 \mid x = 0; \theta_3) = \theta_3$
- $p(x, y; \theta)$ where $\theta = (\theta_1, \theta_2, \theta_3)$ is a vector of parameters.
- or $p(x, y \mid \theta)$, for highlighting that θ is considered a random variable.
- A statistical model corresponds to a set of probabilistic models, here indexed by the parameters θ: $\{p(x; \theta)\}_\theta$

What is Bayesian modeling?

A Bayesian model turns a statistical model into a probabilistic one by treating parameters θ as random variables.

Goal: Learn what we believe about θ after seeing data — and use that to make predictions.

- A Bayesian model is a probabilistic model $p(x, y, \theta)$
- In supervised settings, we consider x as observed, so we care about $p(y, \theta \mid x)$.

Bayesian Modeling in Steps

We don't know the full joint distribution $p(x, y, \theta)$.

- If we did, every analysis would be trivial.

What we do have:

- Observed data $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ — i.i.d. samples from $p(x, y)$

What we assume:

- $p(y \mid x, \theta)$ — how data is generated given a specific parameter θ
- $p(\theta)$ — our beliefs about the parameters before seeing data

What we want to learn:

- $p(\theta \mid \mathcal{D})$ — what we believe about the parameters after seeing data
- The predictive distribution $p(y \mid x, \mathcal{D})$ — predictions that account for parameter uncertainty
- Possibly others: e.g. marginal likelihood $p(\mathcal{D})$

Bayesian Modeling for Supervised Tasks

- **Supervised learning:** We observe a dataset of input–output pairs $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ drawn from an unknown joint distribution $p(x, y)$.
- **Bayesian perspective:** Use this data to learn the relationship $x \mapsto y$ while capturing uncertainty in the model parameters.
- **Hypothesis:**
  - (1): assume a parametric family for $p(y \mid x, \theta)$, such as linear regression, neural networks, etc. (Parametric modeling assumption)
  - (2): assume a prior belief over the parameters $p(\theta)$ (prior assumption)

Example: Linear Regression

- **Hypothesis:** The outcome depends linearly on some input plus noise.
- **Example:** Predict house price from size.
  - Y — house price
  - X — house size
  - Model: $Y = wX + b + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $p(Y \mid X; \theta) = \mathcal{N}(Y \mid wX + b, \sigma^2)$
- **Parameters:** $\theta = (w, b)$ – 2 variables.

Example: Linear Regression

- **Hypothesis:** The outcome depends linearly on some input plus noise.
- Generalizes to any number of input features
- **Example:** Predict house price from size, number of rooms, and
  - Y — house price
  - $X = (X_1, X_2, \ldots, X_d)$ — house features (size, number of rooms, etc.)
  - Model: $Y = wX + b + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $p(Y \mid X; \theta) = \mathcal{N}(Y \mid wX + b, \sigma^2)$
- **Parameters:** $\theta = (w, b)$: $d + 1$ variables.

Example: Non-linear Regression

- **Hypothesis:** The relationship between input and output is complex and nonlinear plus noise.
- **Example:** Predict bike rentals from weather data.
  - Y — number of bikes rented per hour
  - X — weather features (temperature, humidity, etc.)
  - Model $Y = f_\theta(X) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
  - $f_\theta(X)$ is a non-linear function parameterized by θ.
- $p(Y \mid X; \theta) = \mathcal{N}(Y \mid f_\theta(X), \sigma^2)$
- what is $f_\theta(X)$?
  - A neural network with weights θ, normally of thousands of variables.
  - A random forest with decision trees where the structure and parameters are defined by θ, normally of hundreds of variables.

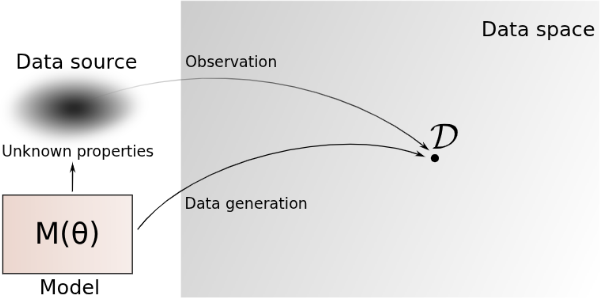
The Flexibility—and Cost—of the Bayesian Framework

- **Bayesian framework is flexible:** We can assume any model for $p(y \mid x, \theta)$.
  - Example: $y = f_\theta(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $f_\theta(x)$ can be a neural network, random forest, etc.
  - The output y can follow any distribution — not just Normal: skewed, heavy-tailed, discrete (e.g., Poisson, Bernoulli), etc.
- **But flexibility comes at a cost**
  - the more complex $p(y \mid x, \theta)$:
    - Complex $\rightarrow$ a high-dimensional θ.
    - Complex $\rightarrow$ a complex $f_\theta(x)$
  - the harder it is to perform inference.

From $p(y \mid x, \theta)$ to the likelihood function $L(\boldsymbol{\theta})$:

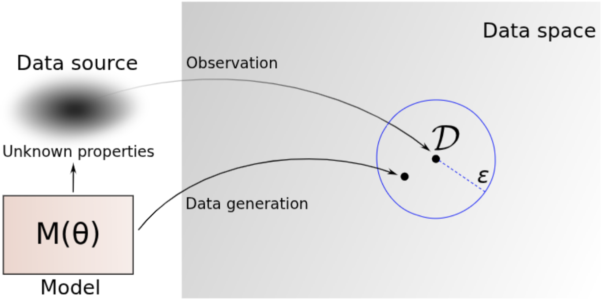
The likelihood function $L(\boldsymbol{\theta})$

- Measures agreement between $\boldsymbol{\theta}$ and the observed data $\mathcal{D}$
- Probability that sampling from the model with parameter value $\boldsymbol{\theta}$ generates data like $\mathcal{D}$
- Exact match for discrete random variables



The likelihood function $L(\boldsymbol{\theta})$

- Measures agreement between $\boldsymbol{\theta}$ and the observed data $\mathcal{D}$
- Probability that sampling from the model with parameter value $\boldsymbol{\theta}$ generates data like $\mathcal{D}$
- Small neighbourhood for continuous random variables



The likelihood function $L(\boldsymbol{\theta})$

- Probability that the model generates data like $\mathcal{D}$ for parameter value $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = p(\mathcal{D}; \boldsymbol{\theta})$$

  where $p(\mathcal{D}; \boldsymbol{\theta})$ is the parameterised model pdf/pmf.
- The likelihood function indicates the likelihood of the parameter values, and not of the data.
- For iid data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

$$L(\boldsymbol{\theta}) = p(\mathcal{D}; \boldsymbol{\theta}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\theta})$$

- Log-likelihood function $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. For iid data:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{x}_i; \boldsymbol{\theta})$$

Different Perspectives on the Likelihood Function

There are different modeling mindsets

- Modeling Mindsets by Christoph Molnar

- **Frequentist perspective:**
  - Premise: The world is best approached through probability distributions with fixed but unknown parameters.
  - one set of parameters $\boldsymbol{\theta}$ is correct, we just don't know which one.
  - Consequence: Find the best parameter values $\boldsymbol{\theta}^*$ — our uncertainty is about whether the parameters are correct.
- **Bayesian perspective:**
  - Premise: The world is best approached through probability distributions with probabilistic parameters.
  - Parameters $\boldsymbol{\theta}$ are random variables with a prior distribution $p(\boldsymbol{\theta})$.
  - Consequence: Update the prior parameter distributions using data to obtain the posterior distribution and draw conclusions.

If we were not Bayesians

- We would use the likelihood function $L(\boldsymbol{\theta})$ to find the best parameter values $\boldsymbol{\theta}^*$.
- Intuition: There is one model that is correct, the one that makes the observed data most probable.
- This is called **maximum likelihood estimation (MLE)**:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}; \boldsymbol{\theta})$$

- MLE does not account for uncertainty in the parameters.

MLE - Example

- lets return to the linear gaussian example:

$$p(y \mid x; \theta) = \mathcal{N}(y \mid wx + b, \sigma^2)$$

- The likelihood function is:

$$L(\theta) = p(\mathcal{D}; \theta) = \prod_{i=1}^{N} p(y^{(i)} \mid x^{(i)}; \theta)$$

- The log-likelihood function is:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{N} \log p(y^{(i)} \mid x^{(i)}; \theta)$$

- MLE maximizes the likelihood

MLE - Example

- MLE finds the parameters $\theta^* = (w, \sigma^2)$ that minimize the negative log-likelihood:

$$\theta^* = \arg \min_{\theta=(w,\sigma^2)} -\ell(\theta)$$

- For $y_i = wx_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the likelihood is:

$$p(y_i \mid x_i, w, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - wx_i)^2}{2\sigma^2}\right)$$

- The negative log-likelihood is:

$$-\ell(w, \sigma^2) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - wx_i)^2$$

MLE - Example

- Minimizing w.r.t. w gives:

$$w^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

- Plugging w^* back and minimizing w.r.t. σ^2 gives:

$$\sigma^{2*} = \frac{1}{n} \sum_{i=1}^{n} (y_i - w^* x_i)^2$$

- w^* are the ones that minimize the squared error between the predicted and observed values.
- σ^{2*} is the variance of the residuals, i.e., the noise in the data.
- for new predictions, we can use:

$$p(y \mid x; w^*, \sigma^{2*}) = \mathcal{N}(y \mid w^* x, \sigma^{2*})$$

Why MLE is Not Enough

- **MLE** finds the parameter θ^* that makes the data most likely.

$$\theta^* = \arg\max_{\theta} L(\theta)$$

- But MLE treats θ^* as *the truth* — no room for doubt.
- **Problem:** It ignores **epistemic uncertainty** — our uncertainty about θ.
- It only models **aleatory uncertainty** — randomness in the data.

Why MLE is Not Enough

- **MLE** finds the parameter θ^* that makes the data most likely.

$$\theta^* = \arg\max_\theta L(\theta)$$

- But MLE treats θ^* as *the truth* — no room for doubt.
- **Problem:** It ignores **epistemic uncertainty** — our uncertainty about θ.
- It only models **aleatory uncertainty** — randomness in the data.
- What if:
  - We have limited data?
  - The model is overly complex?
  - Multiple θ values explain the data almost equally well?

Why We Are Bayesians: Embracing Uncertainty

- MLE ranks parameter values via the likelihood $L(\theta)$:

$$L(\theta^*) = \max_{\theta} L(\theta)$$

- But many θ may be almost as plausible!
- Especially when:
  - data is scarce,
  - the model is complex,
  - or the model is mis-specified.

- **Bayesian modeling** treats θ as a *random variable*, not a fixed value.
- We don't commit to one model — we reason over a **distribution of plausible models**.
- This gives us a posterior distribution:

$$p(\theta \mid \mathcal{D})$$

  capturing our full uncertainty given the data.

Prior and Posterior

- **Prior distribution** $p(\theta)$: Our beliefs about the parameters before seeing data.
- **Posterior distribution** $p(\theta \mid \mathcal{D})$: Our updated beliefs after observing data $\mathcal{D}$.
- The posterior is computed using Bayes' rule:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} = \frac{L(\theta)p(\theta)}{p(\mathcal{D})}$$

  where:
  - $p(\mathcal{D} \mid \theta)$ is the likelihood function.
  - $p(\mathcal{D})$ is the marginal likelihood, a normalizing constant.
- we often write $p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$

Predictive Posterior

- **Predictive posterior** $p(y \mid x, \mathcal{D})$: Our predictions about new data y given input x and observed data $\mathcal{D}$.
- It accounts for uncertainty in the parameters θ:

$$p(y \mid x, \mathcal{D}) = \int p(y \mid x, \theta) p(\theta \mid \mathcal{D}) d\theta$$

  where:
  - $p(y \mid x, \theta)$ is the model likelihood for a specific parameter θ.
  - $p(\theta \mid \mathcal{D})$ is the posterior distribution of the parameters.
- This integral averages over all plausible parameter values, weighted by their posterior probability.
- Normally, it is impossible to compute analytically, so we use approximations and sampling methods.

Predictive Posterior using samples

- If we have samples from the posterior:

$$\theta^m \sim p(\theta \mid \mathcal{D}) \text{ for } m = 1, \ldots, M$$

- we can make predictions by sampling from the predictive posterior:

$$y^m \sim p(y \mid x, \theta^m) \text{ for } m = 1, \ldots, M$$

- This gives us a set of predictions $\{y^m\}_{m=1}^{M}$ with:
  - Expectation:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} y^m$$

  - Variance:

$$\hat{\sigma}^2 = \frac{1}{M} \sum_{m=1}^{M} (y^m - \hat{y})^2$$

- This approach captures uncertainty in the predictions by averaging over all plausible parameter values.

Conclusion

- We have seen how to use the Bayesian framework for probabilistic modeling.
- We have learned how to:
  - Define a prior distribution over parameters.
  - Compute the likelihood function from observed data.
  - Update our beliefs using Bayes' rule to obtain the posterior distribution.
  - Make predictions using the predictive posterior.
- The Bayesian framework allows us to reason about uncertainty in a principled way.
- Next, we will explore practical applications and tools for Bayesian modeling.