Data Science for Mathematicians

Lesson 3: Probabilistic Foundations of Modeling

Department of Mathematics and Computer Science

Outline

From Geometry to Probability

Motivation: Moving beyond deterministic geometry

- Previous approach: Data as fixed vectors, models as subspaces
- Reality: Data contains **uncertainty**, **measurement error**, **randomness**
- Need: Models that can **quantify uncertainty** and make **predictions**

Key transition:

$$\boxed{\text{Deterministic Geometry} \longrightarrow \text{Probabilistic Framework}}$$

σ-Algebras

Definition: σ-Algebra

Let Ω be a non-empty set. A collection $\mathcal{F}$ of subsets of Ω is a σ-**algebra** if:

1. $\Omega \in \mathcal{F}$ (contains whole space)
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ (closed under complement)
3. $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions)

Consequences: $\emptyset \in \mathcal{F}$, closed under countable intersections, set differences

Example

For $\Omega = \{1, 2, 3, 4\}$ and $A = \{1, 2\}$:

$$\mathcal{F} = \{\emptyset, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$$

Borel σ-Algebra

Definition: Borel σ-Algebra

The **Borel σ-algebra** on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra containing all open intervals (a, b).

$\mathcal{B}(\mathbb{R})$ **contains:**

- All open sets, closed sets
- All intervals: (a, b), $[a, b]$, $[a, b)$, $(a, b]$
- Countable sets: singletons $\{x\}$, $\mathbb{Z}$, $\mathbb{Q}$

Example: Common Borel Sets

- $[0, 1]$ (closed interval)
- $\mathbb{Q} = \bigcup_{q \in \mathbb{Q}} \{q\}$ (countable union)
- Cantor set

Measures and Lebesgue Measure

Definition: Measure

A function $\mu : \mathcal{F} \to [0, \infty]$ is a **measure** if:

1. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$
2. $\mu(\emptyset) = 0$
3. $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ for pairwise disjoint A_i

Definition: Lebesgue Measure

The **Lebesgue measure** λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

- $\lambda([a, b]) = b - a$ (length of interval)
- Translation-invariant: $\lambda(A + t) = \lambda(A)$

Example

$\lambda([2, 5]) = 3, \quad \lambda(\{x\}) = 0, \quad \lambda(\mathbb{Q} \cap [0, 1]) = 0$

The Lebesgue Integral

Definition: Simple Function

$\phi : \Omega \to \mathbb{R}$ is **simple** if it takes finitely many values:

$$\phi = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}.$$

Definition: Lebesgue Integral

Let $\phi : \Omega \to \mathbb{R}$ be **simple**, denote

$$\int_{\Omega} \phi \, \mathrm{d}\mu = \sum_{i=1}^{n} a_i \cdot \mu(A_i).$$

Example

For $\phi(x) = 2 \cdot \mathbf{1}_{[0,1)} + 5 \cdot \mathbf{1}_{[1,2)} + 1 \cdot \mathbf{1}_{[2,3]}$:

$$\int_{[0,3]} \phi \, \mathrm{d}\lambda = 2(1) + 5(1) + 1(1) = 8$$

Riemann vs. Lebesgue Integration

Theorem: Riemann vs. Lebesgue

Let $f : [a, b] \to \mathbb{R}$ be bounded. Then:

1. If f is Riemann integrable, then it is Lebesgue integrable and the integrals agree
2. f is Riemann integrable $\Leftrightarrow$ f is continuous almost everywhere

Example: Dirichlet Function – Lebesgue but not Riemann

$$\mathbf{1}_{\mathbb{Q}}(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

Not Riemann integrable (discontinuous everywhere)
Lebesgue integrable:

$$\int_{[0,1]} \mathbf{1}_{\mathbb{Q}} \, \mathrm{d}\lambda = 1 \cdot \lambda(\mathbb{Q} \cap [0,1]) + 0 \cdot \lambda([0,1] \setminus \mathbb{Q}) = 0$$

Probability Space

Definition: Probability Space

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω: **sample space** (all possible outcomes)
- $\mathcal{F}$: σ-**algebra** (collection of events)
- $\mathbb{P}$: **probability measure** with $\mathbb{P}(\Omega) = 1$

Example: Fair Coin Toss

- $\Omega = \{H, T\}$
- $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ (power set)
- $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 0.5$

Example: Fair Die

$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad \mathbb{P}(\{k\}) = 1/6$ for each k

Random Variables

Definition: Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random variable** is a function $X : \Omega \to \mathbb{R}$ such that for every Borel set $B \subset \mathbb{R}$:

$$\{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$$

This property is called **measurability**.

Example: Coin Toss as Random Variable

Define $X : \{H, T\} \to \mathbb{R}$ by $X(H) = 1$, $X(T) = 0$

$$\mathbb{P}(X = 1) = \mathbb{P}(\{H\}) = 0.5$$

Example: Sum of Two Dice

$S((i, j)) = i + j, \quad \mathbb{P}(S = 7) = 6/36 = 1/6$

Types of Random Variables

Discrete:

- Finite or countably infinite range
- Can ask $\mathbb{P}(X = x)$ for each value

Continuous:

- Uncountably infinite range
- $\mathbb{P}(X = x) = 0$ for any single value

Definition: PMF

Probability Mass Function:

$$p_X(x) = \mathbb{P}(X = x)$$

1. $p_X(x) \geq 0$
2. $\sum_x p_X(x) = 1$

Definition: PDF

Probability Density Function:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$$

1. $f_X(x) \geq 0$
2. $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$

Independent and Identically Distributed (i.i.d.)

Definition: i.i.d.

Random variables $X_1, X_2, \ldots, X_n$ are **i.i.d.** if:

1. **Independence:** For any sets $A_1, \ldots, A_n$:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

2. **Identically Distributed:** All X_i have the same distribution

Example: i.i.d. Example

Rolling a die n times: $X_1, \ldots, X_n$ are i.i.d.

- Each roll independent of others
- Each $X_i \sim \text{Uniform}\{1, 2, 3, 4, 5, 6\}$

Non-example: Drawing cards *without replacement* – not independent

The Bernoulli Distribution

Definition: Bernoulli Distribution

$X \sim \text{Bernoulli}(p)$ if its PMF is:

$$\mathbb{P}(X = x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0, 1\}.$$

- $\mathbb{P}(X = 1) = p$ (success)
- $\mathbb{P}(X = 0) = 1 - p$ (failure)

Example: Biased Coin

Coin lands heads with $p = 0.7$:

$$X \sim \text{Bernoulli}(0.7), \quad \mathbb{P}(X = 1) = 0.7, \quad \mathbb{P}(X = 0) = 0.3$$

Example: Titanic Survival

$X_i = 1$ if passenger i survived, $X_i = 0$ otherwise
Model: $X_i \sim \text{Bernoulli}(p)$ where $p =$ survival probability

The Binomial Distribution

Definition: Binomial Distribution

If $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p) and $Y = \sum_{i=1}^{n} X_i$, then $Y \sim \text{Binomial}(n, p)$:

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, 1, \ldots, n\}$$

Example: Coin Flips

Flip fair coin 5 times, Y = number of heads, $Y \sim \text{Binomial}(5, 0.5)$:

$$\mathbb{P}(Y = 3) = \binom{5}{3} (0.5)^3 (0.5)^2 = 10 \cdot 0.03125 = 0.3125$$

Example: Free Throws

Player has 80% success rate, attempts 10 shots:

$$Y \sim \text{Binomial}(10, 0.8) \qquad \mathbb{P}(Y = 8) = \binom{10}{8} (0.8)^8 (0.2)^2 \approx 0.302$$

The Gaussian (Normal) Distribution

Definition: Gaussian Distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$ if its PDF is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- μ: mean (location parameter)
- σ^2: variance (spread parameter)
- $\mathcal{N}(0,1)$: **standard normal distribution**

Why is Gaussian so important?

$\downarrow$

Central Limit Theorem

Properties of the Gaussian Distribution

Theorem: Symmetry

For $X \sim \mathcal{N}(\mu, \sigma^2)$: $f_X(\mu + h) = f_X(\mu - h)$ for all h

Theorem: Unimodality

The PDF has a unique global maximum at $x = \mu$.

Theorem: Asymptotic Behavior

$$\lim_{x \to \pm\infty} f_X(x) = 0$$

Gaussian Distribution: Inflection Points

Theorem: Inflection Points

The PDF $f_X(x)$ for $X \sim \mathcal{N}(\mu, \sigma^2)$ has exactly two inflection points at $x = \mu \pm \sigma$.

Proof.

From $f'_X(x) = -\frac{x-\mu}{\sigma^2} f_X(x)$, using product rule:

$$f''_X(x) = \frac{f_X(x)}{\sigma^4} \left[(x-\mu)^2 - \sigma^2\right]$$

Setting $f''_X(x) = 0$: $(x-\mu)^2 = \sigma^2 \Rightarrow x = \mu \pm \sigma$ □

Example: Standard Normal $Z \sim \mathcal{N}(0,1)$

- Maximum at $z = 0$: $f_Z(0) = \frac{1}{\sqrt{2\pi}} \approx 0.399$
- Inflection points at $z = \pm 1$
- Symmetric: $f_Z(-2) = f_Z(2)$

Mathematical Expectation

Definition: Expectation – Measure-Theoretic

For random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$:

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, \mathrm{d}\mathbb{P}(\omega)$$

Computational Formulas:

- **Discrete:** $\mathbb{E}[X] = \sum_{x \in S} x \cdot p_X(x)$
- **Continuous:** $\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, \mathrm{d}x$

Example: Fair Die

$\mathbb{E}[X] = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$

Example: Bernoulli(p)

$\mathbb{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p$

Expectation: More Examples

Example: Uniform(a, b)

With PDF $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$:

$$\mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a} \, \mathrm{d}x = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{a+b}{2}$$

Example: Exponential(λ)

With PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$:

$$\mathbb{E}[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x} \, \mathrm{d}x = \frac{1}{\lambda}$$

(Using integration by parts)

Interpretation: Expectation is the "center of mass" of the distribution – the long-run average over infinitely many trials

Linearity of Expectation

Theorem: Linearity of Expectation

For any random variables X, Y (not necessarily independent) and constants $a, b \in \mathbb{R}$:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

Proof (Continuous Case).

$$\begin{aligned}
\mathbb{E}[aX + bY] &= \iint (ax + by) f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\
&= a \int x \underbrace{\left(\int f_{X,Y}(x, y)\, \mathrm{d}y \right)}_{f_X(x)} \mathrm{d}x + b \int y f_Y(y)\, \mathrm{d}y \\
&= a\mathbb{E}[X] + b\mathbb{E}[Y]
\end{aligned}$$

□

Key insight: Works *regardless of independence*!

Linearity of Expectation: Applications

Example: Number of Heads in n Coin Flips

Let $X_i = \mathbf{1}_{\{\text{flip } i \text{ is heads}\}}$, so $X = \sum_{i=1}^{n} X_i$.
By linearity: $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$
(No need to sum over all $n+1$ binomial terms!)

Example: Sum of Two Dice

$S = X_1 + X_2$ where X_i is outcome of die i:

$$\mathbb{E}[S] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 3.5 + 3.5 = 7$$

Example: Sample Mean

For i.i.d. $X_1, \ldots, X_n$ with mean μ:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \cdot n\mu = \mu$$

Variance

Definition: Variance

For random variable X with mean $\mu = \mathbb{E}[X]$:

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$$

Standard deviation: $\sigma_X = \sqrt{\text{Var}(X)}$

Theorem: Computational Formula

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Proof.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Variance: Examples

Example: Fair Die

$$\mathbb{E}[X] = 3.5, \quad \mathbb{E}[X^2] = \frac{1+4+9+16+25+36}{6} = \frac{91}{6}$$

$$\text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} \approx 2.92$$

Example: Uniform(a, b)

$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

Example: Exponential(λ)

$$\mathbb{E}[X^2] = \frac{2}{\lambda^2}, \quad \text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Note: $\sigma_X = 1/\lambda = \mathbb{E}[X]$ (mean equals std dev)

Properties of Variance

Theorem: Scaling Property

For constants $a, b \in \mathbb{R}$:

$$\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X)$$

Proof.

$$\begin{aligned}\mathrm{Var}(aX + b) &= \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] = \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\ &= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] = a^2 \,\mathrm{Var}(X)\end{aligned}$$

□

Example: Temperature Conversion

C = temperature in Celsius with $\mathrm{Var}(C) = 25$
$F = \frac{9}{5}C + 32$ (Fahrenheit)
$\mathrm{Var}(F) = \left(\frac{9}{5}\right)^2 \cdot 25 = \frac{81}{25} \cdot 25 = 81$
(Adding 32 doesn't affect variance; scaling by $9/5$ squares the effect)

Covariance

Definition: Covariance

For random variables X, Y with means μ_X, μ_Y:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Interpretation:

- $\text{Cov}(X, Y) > 0$: X and Y tend to move together
- $\text{Cov}(X, Y) < 0$: X and Y tend to move oppositely
- $\text{Cov}(X, Y) = 0$: X and Y are **uncorrelated**

Important: Independence $\Rightarrow$ Uncorrelated, but not vice versa!

Note: $\text{Cov}(X, X) = \text{Var}(X)$

Covariance: The Geometric Connection

Fundamental insight: Covariance $\approx$ Dot Product

For data vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, define centered vectors:

$$\tilde{\mathbf{x}} = \mathbf{x} - \bar{x}\mathbf{1}, \quad \tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}$$

Sample covariance:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}(\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}})$$

Sample covariance = scaled **dot product** of centered vectors

Statistical correlation $\equiv$ Geometric alignment

The Covariance Matrix

Definition: Covariance Matrix

For random vector $\mathbf{X} = [X_1, \ldots, X_p]^T$ with mean $\boldsymbol{\mu}$:

$$\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]$$

where $\Sigma_{ij} = \text{Cov}(X_i, X_j)$.

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \cdots & \cdots & \text{Var}(X_p) \end{pmatrix}$$

Key Properties:

1. **Symmetric:** $\Sigma = \Sigma^T$
2. **Positive semi-definite:** $\mathbf{a}^T \Sigma \mathbf{a} \geq 0$ for all $\mathbf{a}$
3. **Diagonal entries:** variances

Chebyshev's Inequality

Lemma: Chebyshev's Inequality

For random variable X with mean μ and variance σ^2:

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof.

Define indicator $\mathbf{1}_{|X-\mu|\geq\epsilon}$. When $|X - \mu| \geq \epsilon$:

$$\epsilon^2 \cdot \mathbf{1}_{|X-\mu|\geq\epsilon} \leq (X - \mu)^2$$

Taking expectations:

$$\epsilon^2 \cdot \mathbb{P}(|X - \mu| \geq \epsilon) \leq \mathbb{E}[(X - \mu)^2] = \sigma^2$$

Dividing by ϵ^2 gives the result. □

Significance: Universal bound using only mean and variance – works for *any* distribution!

The Law of Large Numbers

Theorem: Weak Law of Large Numbers

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

Proof.

First: $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$.

By Chebyshev: $\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$ as $n \to \infty$. □

Interpretation: Sample mean converges to true mean as $n \to \infty$

Justification: Why sample statistics estimate population parameters

The Central Limit Theorem

Theorem: Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance $\sigma^2 > 0$. Then:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Equivalently, for large n: $\bar{X}_n \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Profound implication:

Distribution of sample means approaches Gaussian **regardless** of the underlying distribution

Why Gaussian is everywhere: Many phenomena arise from aggregating small independent effects

CLT: Proof Sketch

Proof Outline using Characteristic Functions.

Step 1: Let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ (standardized, assuming $\mu = 0$, $\sigma = 1$)

Step 2: Characteristic function: $\varphi_{Z_n}(t) = \left[\varphi_X(t/\sqrt{n})\right]^n$

Step 3: Taylor expand: $\varphi_X(s) = 1 - \frac{s^2}{2} + O(s^3)$

Step 4: Substitute $s = t/\sqrt{n}$:

$$\varphi_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + O(n^{-3/2})\right]^n$$

Step 5: Take limit using $(1 + a/n)^n \to e^a$:

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = e^{-t^2/2}$$

This is the characteristic function of $\mathcal{N}(0,1)$. By Levy's theorem: $Z_n \xrightarrow{d} \mathcal{N}(0,1)$. □

CLT: Example

Example: Averaging Dice Rolls

Single die roll: $\mu = 3.5$, $\sigma^2 = 35/12 \approx 2.917$
Distribution: discrete uniform (far from Gaussian!)

Roll $n = 100$ dice and compute $\bar{X}_{100}$. By CLT:

$$\bar{X}_{100} \overset{\text{approx}}{\sim} \mathcal{N}\left(3.5, \frac{2.917}{100}\right) = \mathcal{N}(3.5, 0.0292)$$

Standard deviation of $\bar{X}_{100}$: $\sigma/\sqrt{n} \approx 0.171$

95% of the time:

$$\bar{X}_{100} \in [3.5 - 1.96(0.171), 3.5 + 1.96(0.171)] \approx [3.16, 3.84]$$

The Estimation Problem

Setting:

- Observed data: $D = \{x_1, \ldots, x_n\}$
- Model: Parametric distribution $f(x|\theta)$
- Goal: Estimate θ from data

Examples:

- Coin flips $\rightarrow$ estimate bias p
- Heights $\rightarrow$ estimate mean μ, variance σ^2
- Linear regression $\rightarrow$ estimate coefficients β

Approach: Maximum Likelihood Estimation (MLE)

The Likelihood Function

Definition: Likelihood Function

Given data $D = \{x_1, \ldots, x_n\}$ from i.i.d. distribution $f(x|\theta)$:

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

Key distinction:

- $f(x|\theta)$: probability of data given fixed θ
- $L(\theta|D)$: same expression, but data is fixed, θ varies

Example: Bernoulli Likelihood

Observe $D = \{1, 0, 1\}$ (H, T, H):

$$L(p|D) = p \cdot (1-p) \cdot p = p^2(1-p)$$

$L(0.5) = 0.125, \quad L(0.7) = 0.147 \quad (p = 0.7$ more likely!)

Maximum Likelihood Estimation

Definition: MLE

The **maximum likelihood estimate** is:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta|D) = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i|\theta)$$

Practical approach: Maximize **log-likelihood** instead

$$\ell(\theta|D) = \log L(\theta|D) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Why?

- Products $\rightarrow$ sums (easier calculus)
- Avoids numerical underflow
- log is monotonic, so same maximizer

Method: Set $\frac{\partial \ell}{\partial \theta} = 0$ and solve for θ

MLE for Bernoulli Parameter

Example: MLE for Bernoulli(p)

Data: n trials with n_1 successes, $n_0 = n - n_1$ failures

Step 1: Likelihood

$$L(p|D) = p^{n_1}(1-p)^{n_0}$$

Step 2: Log-likelihood

$$\ell(p|D) = n_1 \log p + n_0 \log(1-p)$$

Step 3: Differentiate

$$\frac{d\ell}{dp} = \frac{n_1}{p} - \frac{n_0}{1-p} = 0$$

Step 4: Solve

$$n_1(1-\hat{p}) = n_0\hat{p} \quad \Rightarrow \quad \boxed{\hat{p}_{\text{MLE}} = \frac{n_1}{n}}$$

Result: MLE = sample proportion (intuitive!)

Connecting OLS and MLE

Linear model: $y_i = \mathbf{x}_i^T \beta + \epsilon_i \quad$ with $\quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

This implies: $y_i | \mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i^T \beta, \sigma^2)$

Log-likelihood:

$$\begin{aligned}\ell(\beta, \sigma^2 | D) &= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \beta)^2}{2\sigma^2} \right) \right] \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2\end{aligned}$$

To maximize w.r.t. β**:**

$$\hat{\beta}_{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 = \arg\min_{\beta} \|y - X\beta\|^2$$

OLS = MLE Under Gaussian Errors

Key Result:

Under assumption $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$\hat{\beta}_{\text{MLE}} = \hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T y$$

Significance:

- **Geometric view** (Lesson 2): OLS = orthogonal projection
- **Probabilistic view** (Lesson 3): OLS = MLE
- Same answer, different perspectives!

Implication: Statistical inference (confidence intervals, hypothesis tests) now possible via probability theory

Conditional Probability

Definition: Conditional Probability

For events A, B with $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Intuition: Restrict sample space to B, rescale probabilities

Example: Rolling a Die

$A =$ "even" $= \{2, 4, 6\}$, $\quad B =$ "> 3" $= \{4, 5, 6\}$

$A \cap B = \{4, 6\}$, so:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(\{4, 6\})}{\mathbb{P}(\{4, 5, 6\})} = \frac{2/6}{3/6} = \frac{2}{3}$$

Product Rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$

Conditional Probability: Examples

Example: Drawing Cards Without Replacement

Probability of two Aces in a row?
Let A_1 = first Ace, A_2 = second Ace

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1) = \frac{3}{51} \times \frac{4}{52} = \frac{1}{221}$$

Example: Medical Testing

- Disease prevalence: $\mathbb{P}(D) = 0.01$
- Sensitivity: $\mathbb{P}(T^+|D) = 0.95$
- Specificity: $\mathbb{P}(T^-|\neg D) = 0.90$, so $\mathbb{P}(T^+|\neg D) = 0.10$

$$\mathbb{P}(D \cap T^+) = \mathbb{P}(T^+|D)\mathbb{P}(D) = 0.95 \times 0.01 = 0.0095$$

If test positive, what's probability of disease? $\rightarrow$ Bayes' Theorem

Bayes' Theorem

Theorem: Bayes' Theorem

For events A, B with $\mathbb{P}(A), \mathbb{P}(B) > 0$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Proof.

From the product rule:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$$
$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

Equating and dividing by $\mathbb{P}(B)$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Bayes' Theorem: Medical Testing

Example: Medical Test Revisited

Given: $\mathbb{P}(D) = 0.01$, $\mathbb{P}(T^+|D) = 0.95$, $\mathbb{P}(T^+|\neg D) = 0.10$
Question: If test positive, what is $\mathbb{P}(D|T^+)$?

Step 1: Total probability of positive test

$$\begin{aligned}\mathbb{P}(T^+) &= \mathbb{P}(T^+|D)\mathbb{P}(D) + \mathbb{P}(T^+|\neg D)\mathbb{P}(\neg D) \\ &= 0.95(0.01) + 0.10(0.99) = 0.0095 + 0.099 = 0.1085\end{aligned}$$

Step 2: Apply Bayes

$$\mathbb{P}(D|T^+) = \frac{0.95 \times 0.01}{0.1085} \approx \boxed{0.088}$$

Surprising: Only 8.8% chance of disease despite positive test!
(Low prevalence $\Rightarrow$ false positives dominate)

Bayesian Inference

Recast Bayes for parameters and data:

$$\underbrace{\mathbb{P}(\theta|D)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(D|\theta)}^{\text{Likelihood}} \cdot \overbrace{\mathbb{P}(\theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(D)}_{\text{Evidence}}}$$

- **Prior** $\mathbb{P}(\theta)$: Belief about θ *before* data
- **Likelihood** $\mathbb{P}(D|\theta)$: Same as MLE's $L(\theta|D)$
- **Posterior** $\mathbb{P}(\theta|D)$: Updated belief *after* data
- **Evidence** $\mathbb{P}(D)$: Normalizing constant

$$\boxed{\text{Prior Belief} + \text{Data} \longrightarrow \text{Updated Posterior Belief}}$$

Often: $\mathbb{P}(\theta|D) \propto \mathbb{P}(D|\theta)\mathbb{P}(\theta)$

Bayesian Inference: Example

Example: Bayesian Coin Flipping

Prior: $\theta \sim \text{Beta}(\alpha, \beta)$ with $\mathbb{P}(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
Data: n flips, h heads, $t = n - h$ tails
Likelihood: $\mathbb{P}(D|\theta) \propto \theta^h(1-\theta)^t$

Posterior:

$$\begin{aligned}\mathbb{P}(\theta|D) &\propto \theta^h(1-\theta)^t \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{h+\alpha-1}(1-\theta)^{t+\beta-1}\end{aligned}$$

$$\therefore \quad \theta|D \sim \text{Beta}(\alpha + h, \beta + t)$$

Numerical: Prior $= \text{Beta}(1,1)$, observe 7 heads in 10 flips:

- Posterior $= \text{Beta}(8,4)$, mean $= 8/12 \approx 0.667$
- MLE $= 7/10 = 0.7$

The Naive Bayes Classifier

Goal: Classify observation $\mathbf{x} = (x_1, \ldots, x_p)$ into class C_k

Bayesian approach:

$$\mathbb{P}(C_k|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)}{\mathbb{P}(\mathbf{x})}$$

Decision rule:

$$\hat{C} = \arg\max_k \mathbb{P}(\mathbf{x}|C_k)\mathbb{P}(C_k)$$

Definition: Naive Bayes Assumption

Features are **conditionally independent** given class:

$$\mathbb{P}(\mathbf{x}|C_k) = \prod_{j=1}^{p} \mathbb{P}(x_j|C_k)$$

Final classifier:

$$\hat{C} = \arg\max_k \left[\log \mathbb{P}(C_k) + \sum_{j=1}^{p} \log \mathbb{P}(x_j|C_k) \right]$$

Naive Bayes: Spam Detection Example

Example: Email Spam Detection

Classes: Spam (C_1), Not Spam (C_2)
Features: presence of words "free", "meeting", "winner"

	Spam	Not Spam
Prior $\mathbb{P}(C_k)$	0.40	0.60
$\mathbb{P}(\text{"free"} = 1 \mid C_k)$	0.80	0.10
$\mathbb{P}(\text{"meeting"} = 1 \mid C_k)$	0.10	0.70
$\mathbb{P}(\text{"winner"} = 1 \mid C_k)$	0.70	0.05

New email: contains "free" and "winner", not "meeting"
Spam: $0.40 \times 0.80 \times 0.90 \times 0.70 = 0.202$
Not Spam: $0.60 \times 0.10 \times 0.30 \times 0.05 = 0.0009$
$\mathbb{P}(\text{Spam} \mid \mathbf{x}) \approx 0.996 \Rightarrow$ **Classify as Spam**

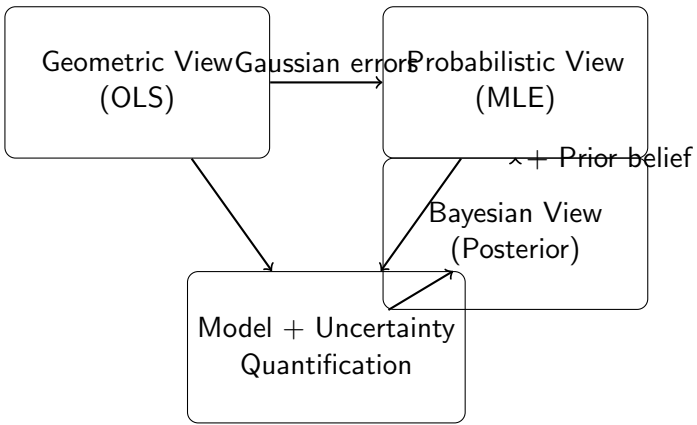
Summary: Three Pillars

1. **Distributions as Models**
   - Data = realizations of random variables
   - Bernoulli/Binomial for classification
   - Gaussian for continuous variables (CLT justification)

2. **Parameter Estimation via MLE**
   - Likelihood function: $L(\theta|D) = \prod_i f(x_i|\theta)$
   - MLE: maximize likelihood (or log-likelihood)
   - **OLS = MLE** under Gaussian errors

3. **Moments and Relationships**
   - Expectation (center), Variance (spread)
   - Covariance $\approx$ Dot product of centered vectors
   - LLN and CLT: foundations of statistical inference

The Big Picture



Questions?

"Probability theory is nothing but common sense reduced to calculation."

— Pierre-Simon Laplace