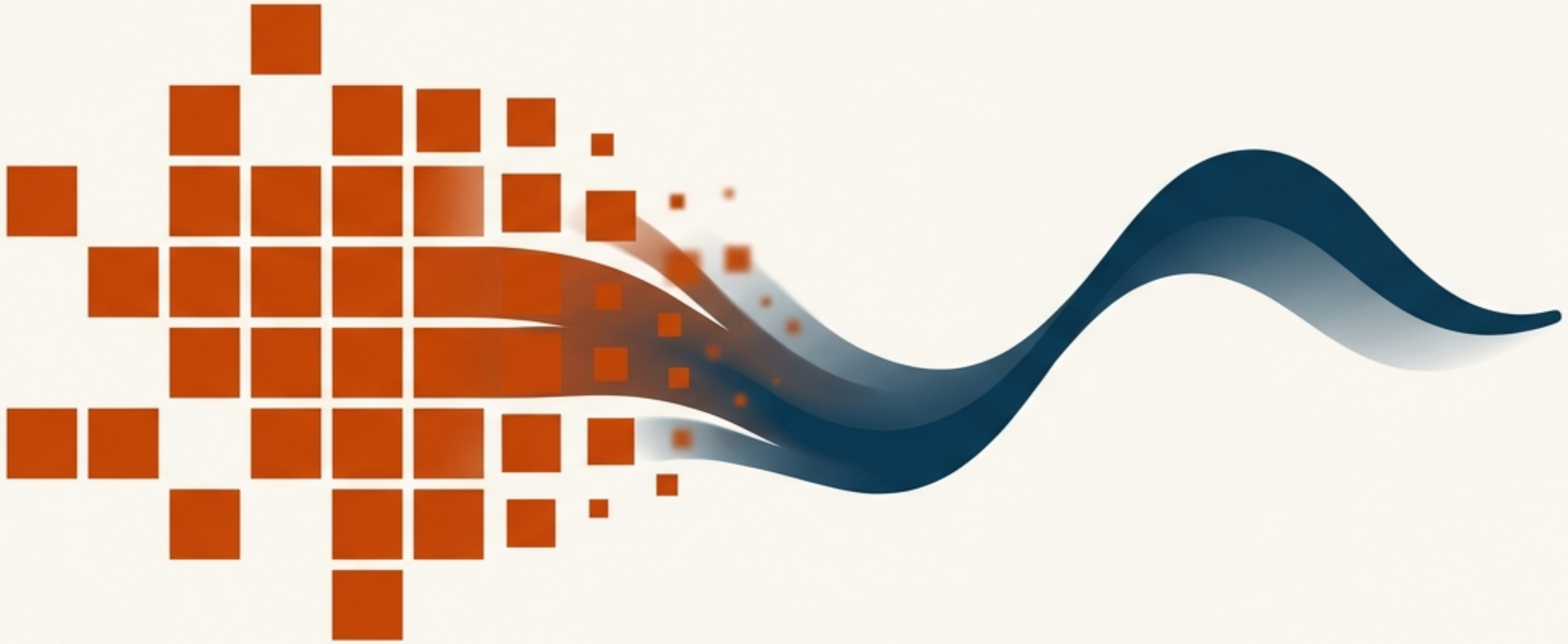
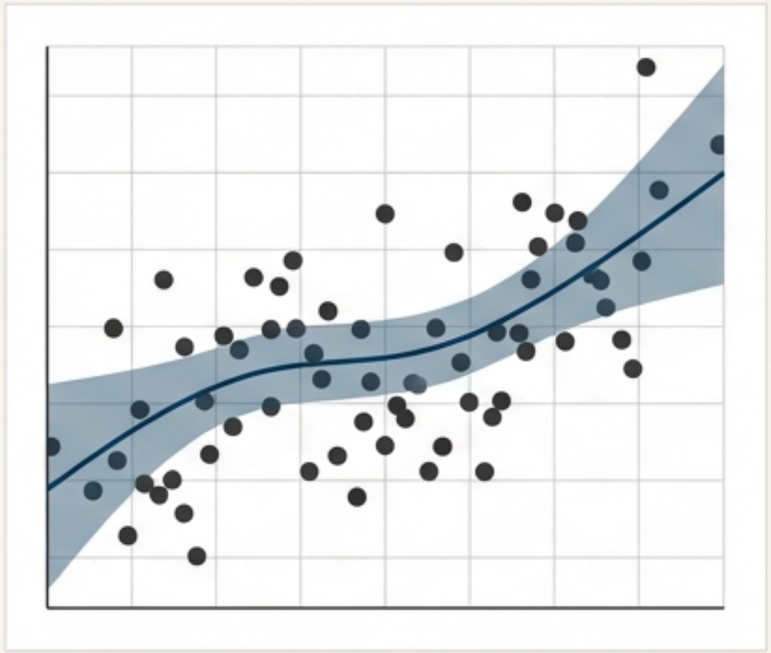
Certainty in Uncertainty

A Guide to Discrete & Continuous Probability in Machine Learning

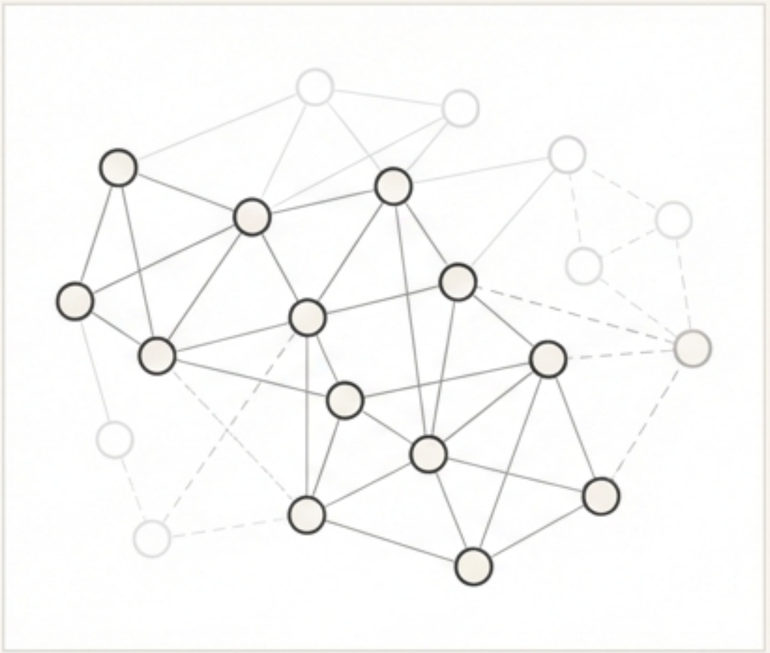


Based on concepts from *Mathematics for Machine Learning* by Deisenroth, Faisal, and Ong.

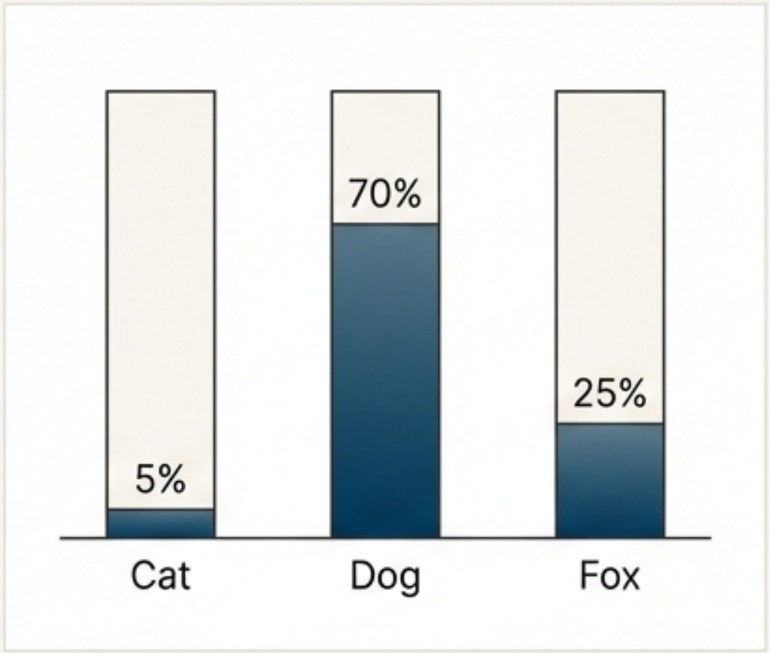
Machine learning models must operate in an uncertain world.



Noisy Data



Incomplete Information



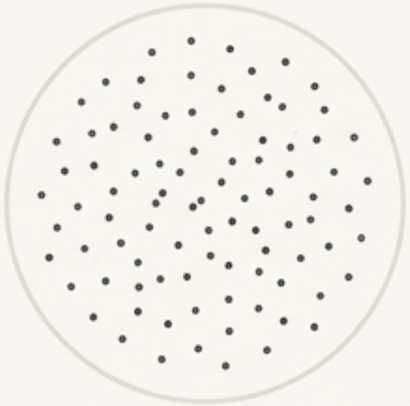
Model Confidence

The world is not deterministic. Data contains noise, information is incomplete, and models must quantify their confidence. To build robust systems, we need a formal language to represent and manipulate uncertainty. Probability theory provides this language.

> Quantification of uncertainty is the realm of probability theory.

All probability begins with a single, rigorous foundation.

The **Probability Space**



`Sample Space (Ω)`

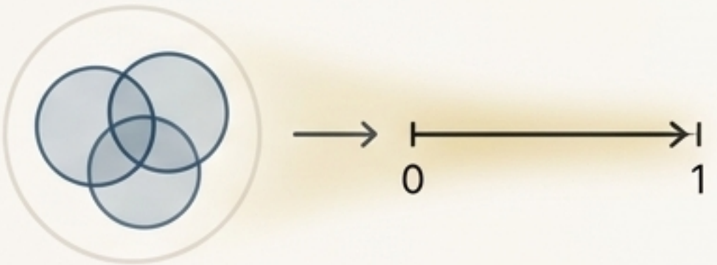
The set of all possible outcomes.

The universe of possibilities.

`Event Space (F)`

The set of “interesting” subsets of Ω about which we can ask probabilistic questions.

The questions we are allowed to ask.



`Probability Measure (P)`

The function that assigns a probability value in [0, 1] to every event in F.

The machine that provides the answers.

From this foundation, two worlds emerge.

The Discrete World

The realm of countable things.

Probability Space

$(\Omega, \mathcal{F}, \mathcal{P})$

While their landscapes differ, they are governed by the same fundamental laws.

The Continuous World

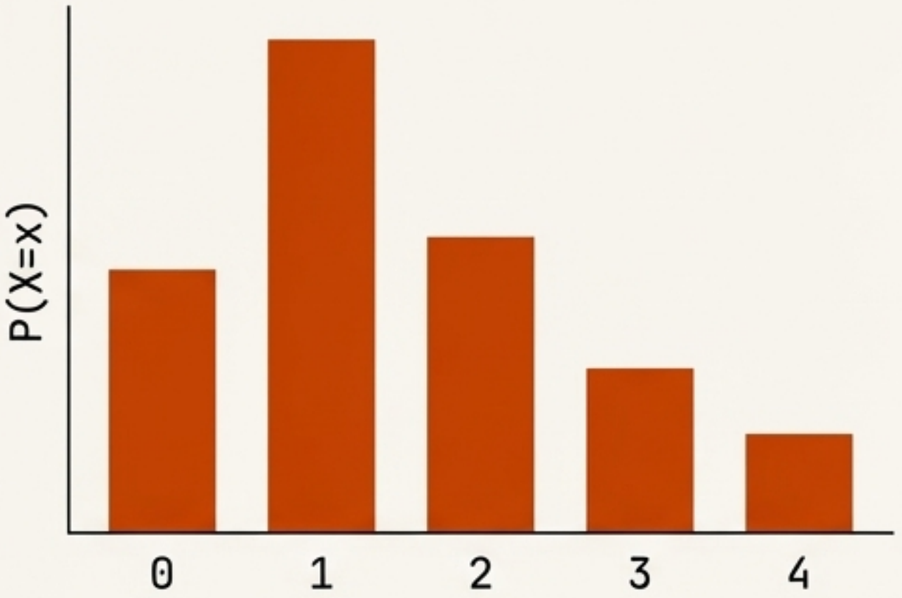
The realm of measurable things.

Exploring the Discrete World: Probability of Countable Outcomes

The **Probability Mass Function (PMF)** assigns a specific probability to each possible discrete outcome. For a random variable X, the PMF gives $\mathrm{P}(X = x)$.

The probability of any specific outcome *can* be greater than zero: $\mathrm{P}(X = x) > 0$.

The sum of probabilities over all possible outcomes must equal 1.



Starring Example: The Bernoulli Distribution

The simplest discrete distribution, modeling a single trial with two outcomes (e.g., success/failure, 1/0, click/no-click).

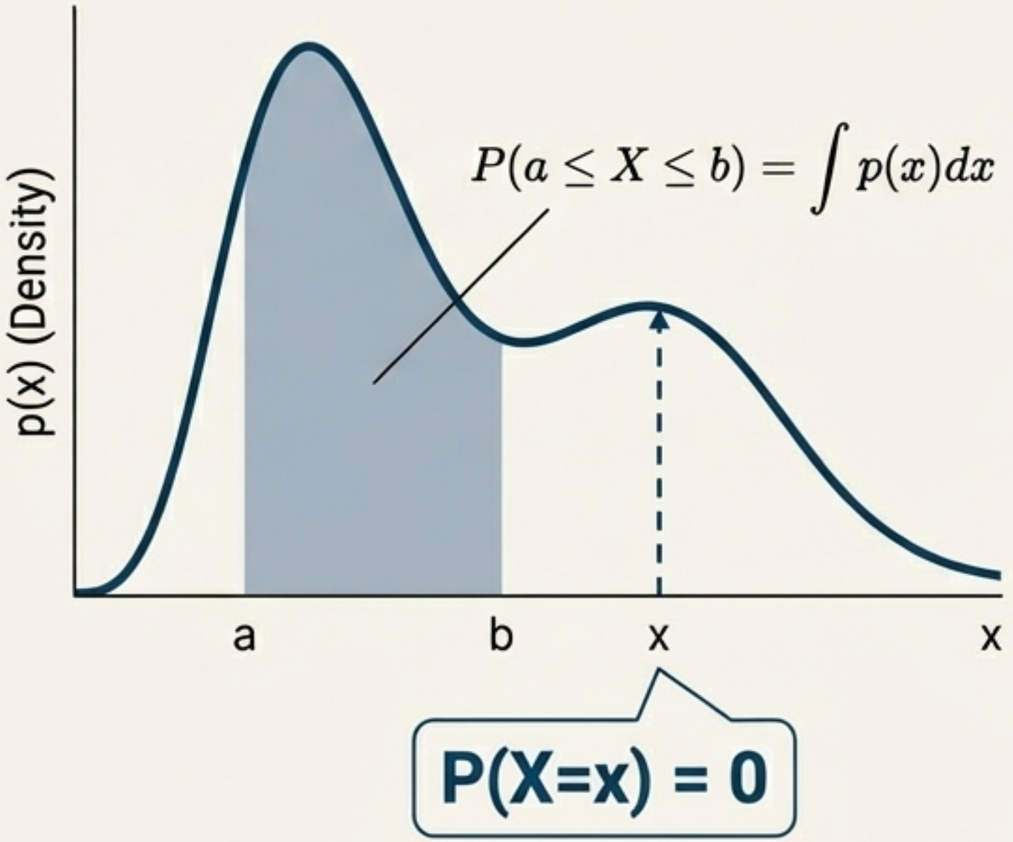
$X \sim \mathrm{Ber}(\mu)$, where μ is the probability of success.

Source reference: Symbol table, p. 7.

Exploring the Continuous World: Probability over Intervals

The **Probability Density Function (PDF)**, denoted $p(x)$, describes the relative likelihood of a random variable taking on a certain value.

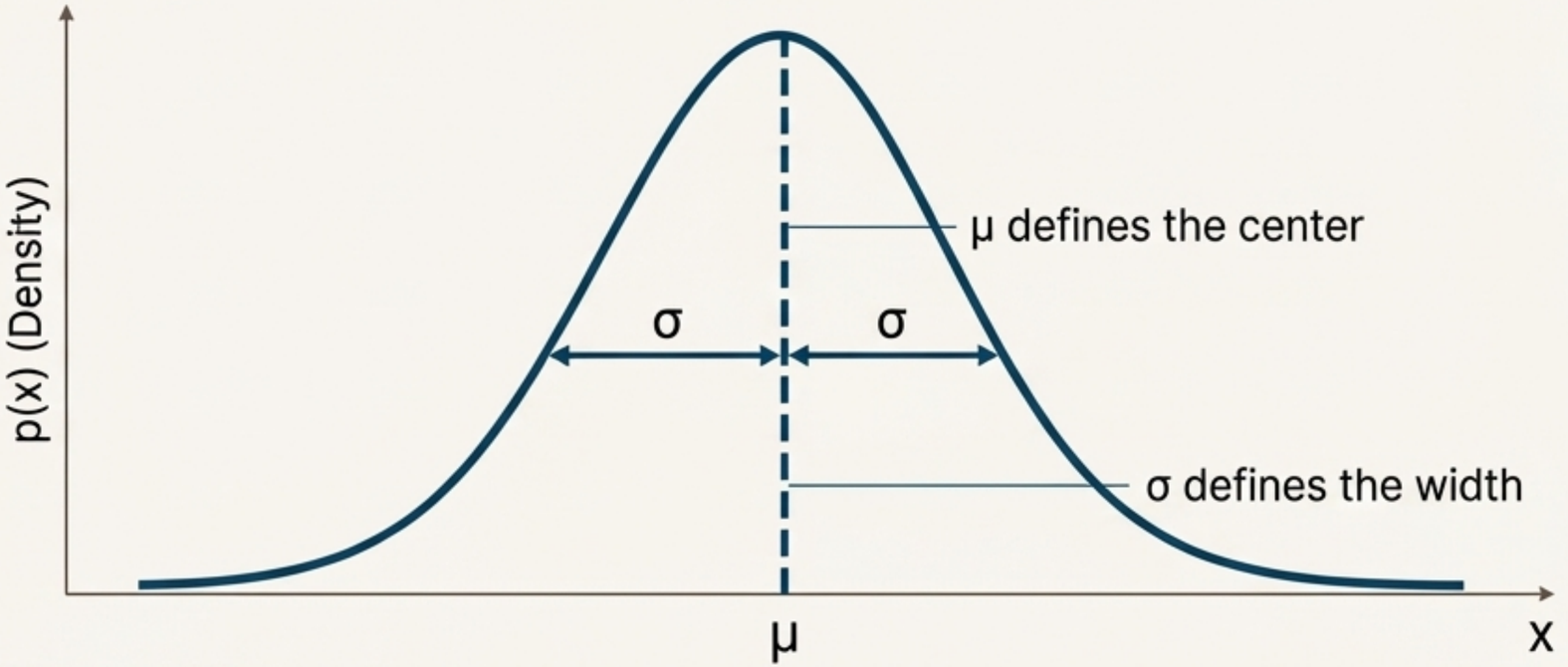
The probability of any single point is zero: $P(X = x) = 0$. Probability is only defined over an interval, calculated as the *area under the curve* of the PDF. The total area The total area must be 1: $\int p(x)dx = 1$.



Based on Section 6.2, Discrete and Continuous Probabilities.

The Star of the Continuous World: The Gaussian Distribution

The **Gaussian** (or **Normal**) distribution is the most important continuous distribution in statistics and machine learning, often used to model noise or the distribution of natural phenomena.



It is completely specified by two parameters:

- **Mean (μ):** The center of the distribution (the peak of the bell curve).
- **Covariance (Σ):** The spread or width of the distribution.
- **Notation:** $X \sim N(\mu, \Sigma)$

In many models, the mean and covariance are the very parameters the machine learning algorithm aims to *learn* from the data.

The Unifying Laws of Probability

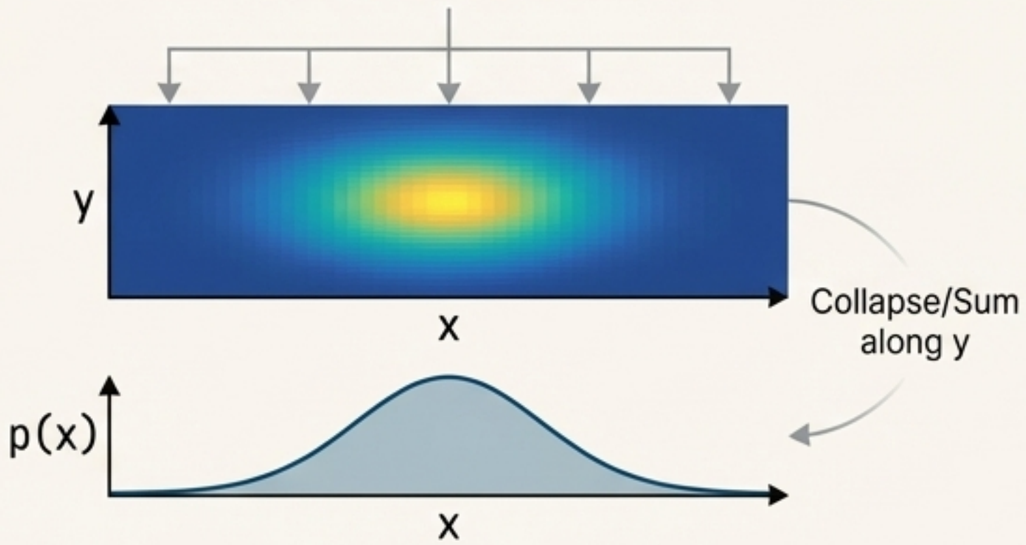
Despite their differences, both discrete and continuous worlds are governed by two fundamental rules for combining probabilities.

Rule 1: The Sum Rule (The Law of Marginalization)

How to remove a variable you don't care about from a joint distribution.

$$p(x) = \int p(x,y)\,dy \quad \text{(Continuous)}$$

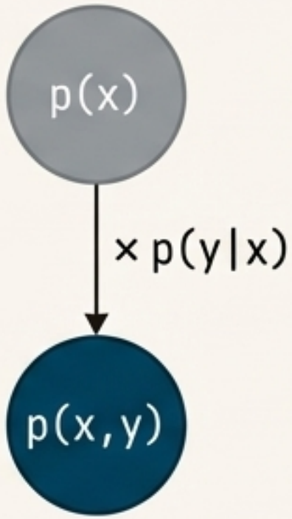
$$p(x) = \sum_y p(x,y) \quad \text{(Discrete)}$$



Rule 2: The Product Rule (The Chain Rule of Probability)

How to build complex joint distributions from simpler conditional and marginal parts.

$$p(x,y) = p(y|x)p(x)$$



Section 6.3, Sum Rule, Product Rule, and Bayes' Theorem.

The Master Key: Bayes' Theorem

Bayes' Theorem is the engine of probabilistic learning. It tells us how to systematically update our beliefs in the face of new evidence. It is a direct consequence of the product rule.

$$[\text{Posterior}] \propto [\text{Likelihood}] \times [\text{Prior}]$$

$$p(\theta|D) \propto p(D|\theta)\, p(\theta)$$

Prior ($p(\theta)$)

What we believe about our model parameters *before* seeing data.

Likelihood ($p(D|\theta)$)

A function that measures how well our model (with parameters θ) explains the observed data D.

Posterior ($p(\theta|D)$)

Our updated belief about the parameters *after* observing the data. This is often the goal of learning.

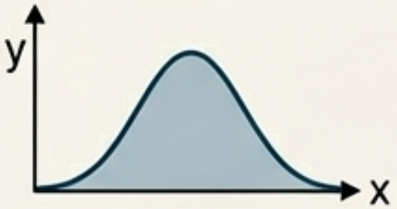
Evidence ($p(D)$)

The denominator, $p(D) = \int p(D|\theta)p(\theta)d\theta$, acts as a normalizer. It is often the most difficult part to compute.

From the Product Rule

$$p(D,\theta) = p(D|\theta)p(\theta) \qquad p(D,\theta) = p(\theta|D)p(D)$$

The Two Worlds at a Glance

	Discrete Probability	Continuous Probability
Outcomes	Countable (e.g., integers, categories)	Uncountable (e.g., real numbers)
Key Function	Probability Mass Function (PMF), $P(x)$	Probability Density Function (PDF), $p(x)$
Interpretation	$P(X=x)$ is the probability of outcome x.	$p(x)$ is a density; probability is the area under the curve, $\int_a^b p(x)dx$.
Key Property	$P(x) \geq 0$, and $\sum_x P(x) = 1$	$p(x) \geq 0$, and $\int p(x)dx = 1$. Crucially, $P(X=x) = 0$.
Example Plot		
Common Distributions	Bernoulli, Binomial, Categorical, Poisson	Gaussian (Normal), Uniform, Exponential, Beta

The Bridge to Central Machine Learning Problems

Understanding these probabilistic foundations is essential for deriving and interpreting the four pillars of machine learning.

Pillar 1: Linear Regression

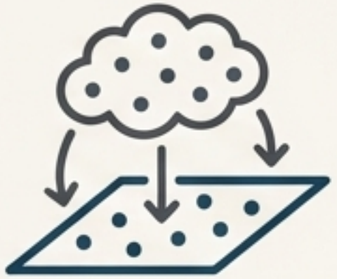
Often assumes observation noise is Gaussian (a continuous distribution). Bayesian Linear Regression uses Bayes' Theorem to find a posterior distribution over model parameters. (Chapter 9)

Pillar 2: Density Estimation

The goal is to model the probability distribution of data. Gaussian Mixture Models (GMMs) explicitly model the data density as a sum of multiple continuous Gaussian distributions. (Chapter 11)

Pillar 3: Classification

Probabilistic classifiers like Naive Bayes use the rules of probability and Bayes' Theorem to calculate the probability of a class given a set of features. (Related to concepts in Chapter 12)



Pillar 4: Dimensionality Reduction

Probabilistic PCA (PPCA) frames dimensionality reduction from a latent-variable perspective, modeling the data with a continuous Gaussian latent variable model. (Chapter 10)

“Probability is the language we use to build models that learn, reason, and decide in an uncertain world.

It is not just a prerequisite; it is the operating system for modern machine learning, transforming data into beliefs, and beliefs into action.

For tutorials, errata, and additional materials, visit:

https://mml-book.com