Contents

Chapter 1

Foundations of Machine Learning from a Probabilistic Perspective

Machine Learning (ML) can be broadly defined as the process by which a computational system improves its own performance or makes more accurate predictions through experience. From a probabilistic perspective, machine learning leverages the principles of probability theory to quantify the uncertainty inherent in data and models.

In this chapter, we focus on the fundamental concepts of machine learning through a probabilistic lens, without delving into the details of any specific models such as linear regression, neural networks, or support vector machines. Instead, we explore the core building blocks—random variables, probability distributions, and the general paradigms used to learn from data.

1.1 Basic Probability Theory

1.1.1 Random Variables and Distributions

A *random variable* X is a mathematical formalization of an uncertain quantity, mapping outcomes in a sample space Ω to real values. For instance, X might represent the measured value of a physical process, or the label associated with a data point. Random variables can be:

- **Discrete**, where X takes values from a countable set (e.g., $\{0, 1, 2, \ldots\}$).
- **Continuous**, where X takes values from an uncountable set (e.g., any real number $\mathbb{R}$).

A probability distribution characterizes the likelihood of different outcomes for a random variable. In the discrete case, the distribution is specified by a probability mass function (PMF):

$$P(X = x) = p(x), \quad x \in \mathcal{X},$$

while in the continuous case, it is typically given by a probability density function (PDF):

$$p(x) = \frac{d}{dx} P(X \leq x), \quad x \in \mathbb{R}.$$

1.1.2 Joint, Marginal, and Conditional Distributions

Many machine learning problems involve multiple random variables (e.g., features X and labels Y). We often encounter:

- **Joint distribution** $p(x, y)$: The probability distribution over pairs (X, Y).
- **Marginal distribution** $p(x) = \sum_y p(x, y)$ (discrete) or $p(x) = \int p(x, y)\, dy$ (continuous).
- **Conditional distribution** $p(y \mid x) = \frac{p(x,y)}{p(x)}$, which describes the distribution of Y given X.

1.1.3 Expectation and Moments

The *expectation* (or mean) of a random variable X is defined by:

- Discrete:
$$\mathbb{E}[X] = \sum_x x\, p(x),$$
- Continuous:
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, p(x)\, dx.$$

Similarly, the *variance* of X is:
$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Higher-order moments (e.g., skewness, kurtosis) capture further properties of the distribution.

1.2 Probabilistic View of Learning

1.2.1 The Data Generating Process

In the probabilistic setting, we assume there is a true—but typically unknown—data generating distribution $p(x, y)$. Machine learning tasks revolve around using a finite sample $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ drawn from $p(x, y)$ to infer properties about this distribution, such as predicting y for a previously unseen x, or understanding the structure within x itself.

1.2.2 Supervised, Unsupervised, and Beyond

- **Supervised learning**: We have labeled data (x, y). The goal is to learn a function f that maps x to y accurately under some measure of success.
- **Unsupervised learning**: We have unlabeled data x. The goal is to discover patterns or structures within x (e.g., clustering, density estimation, dimensionality reduction).
- **Reinforcement learning**: Data arrives sequentially as an agent interacts with an environment. The agent learns a policy to maximize future rewards.

From a probabilistic perspective, each of these can be viewed as working under assumptions about the underlying data distribution, whether that distribution includes labels, has a certain latent structure, or is observed incrementally.

1.3 Hypothesis Spaces and Parameters

1.3.1 Parametric and Non-Parametric Perspectives

1. **Parametric**: We assume a functional form $p_\theta(y \mid x)$ (or simply $f_\theta(x)$) governed by a finite set of parameters θ. Learning involves estimating θ from data.
2. **Non-parametric**: We do not fix the functional form beforehand, allowing the model complexity to grow with the amount of data (e.g., kernel methods, some forms of nearest neighbors).

Both can be cast in probabilistic terms, though parametric methods often lean more directly on probability distributions with explicit likelihoods.

1.3.2 Likelihood and Parameter Estimation

Given a parametric family $\{p_\theta(x, y)\}$, the data sample $\mathcal{D}$ yields the *likelihood*:

$$\mathcal{L}(\theta; \mathcal{D}) = p_\theta(\mathcal{D}) = \prod_{i=1}^{N} p_\theta(x_i, y_i),$$

or in a conditional modeling framework:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} p_\theta(y_i \mid x_i).$$

The *maximum likelihood estimate* (MLE) is the parameter $\hat{\theta}_{\text{MLE}}$ that maximizes $\mathcal{L}(\theta; \mathcal{D})$ or equivalently the log-likelihood.

1.3.3 Bayesian Perspective

In a Bayesian framework, we treat θ itself as a random variable with a *prior distribution* $p(\theta)$. The *posterior distribution* is obtained via Bayes theorem:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})} = \frac{\mathcal{L}(\theta; \mathcal{D})\, p(\theta)}{\int \mathcal{L}(\theta'; \mathcal{D})\, p(\theta')\, d\theta'}.$$

By incorporating a prior, Bayesian methods can encode domain knowledge or constraints, and they quantify uncertainty through the posterior distribution rather than a single parameter estimate.

1.4 Loss Functions and Risk

1.4.1 Defining a Loss

A *loss function* $L(y, \hat{y})$ measures the cost of predicting $\hat{y}$ when the true outcome is y. Common examples include:

- Squared loss: $L(y, \hat{y}) = (y - \hat{y})^2$.
- Zero-one loss: $L(y, \hat{y}) = \mathbb{I}[y \neq \hat{y}]$.
- Absolute loss: $L(y, \hat{y}) = |y - \hat{y}|$.

1.4.2 Expected Risk and Empirical Risk

Under a probability distribution $p(x, y)$, the *expected risk* for a hypothesis h (or model) is:

$$R(h) = \mathbb{E}_{(x,y)\sim p}\big[L(y, h(x))\big].$$

However, since $p(x, y)$ is unknown in practice, we approximate $R(h)$ using the *empirical risk*:

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} L\big(y_i, h(x_i)\big),$$

based on a finite sample $\mathcal{D}$. Learning often involves finding a hypothesis h^* that minimizes $\hat{R}(h)$, subject to various regularization or capacity constraints.

1.5 Generalization and the Role of Probability

1.5.1 Overfitting and Underfitting

Because we rely on finite data, a model might *overfit*, capturing noise or peculiarities in $\mathcal{D}$ that do not generalize to new data, or *underfit*, failing to capture the essential structure of $\mathcal{D}$. Probabilistic arguments help us understand how to control these errors by analyzing, for example, confidence intervals, bounds on deviations of empirical means from expectations, and complexity penalties.

1.5.2 Law of Large Numbers and Central Limit Theorem

- **Law of Large Numbers (LLN)**: As the number of samples N grows, the empirical average converges to the true expectation:
$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} X_i = \mathbb{E}[X].$$
- **Central Limit Theorem (CLT)**: Under mild conditions, the distribution of the sample mean around the true mean converges to a normal distribution for large N. This theorem is fundamental for constructing confidence intervals and quantifying uncertainty.

Such probabilistic results give us insight into how well the empirical risk approximates the true risk and how sample size influences reliability of our estimates.

1.6 Learning Frameworks and Beyond

1.6.1 Frequentist vs. Bayesian Interpretations

- **Frequentist interpretation**: Parameters are fixed but unknown quantities, and variability arises from different possible samples of data.
- **Bayesian interpretation**: Parameters themselves have a distribution, capturing prior knowledge and uncertainty.

Both interpretations find their place in machine learning and can often be complementary.

1.6.2 PAC Learning

In the Probably Approximately Correct (PAC) framework, a learner aims to find a hypothesis h such that with high probability (over the choice of the training set), h has a low generalization error. Formally, for any $\epsilon > 0$ and $\delta > 0$, with probability at least $1 - \delta$, the hypothesis h satisfies $R(h) \leq \epsilon$. The framework provides bounds on the number of samples needed for such guarantees.

1.7 Conclusion

In this chapter, we examined the fundamentals of machine learning from a purely probabilistic standpoint, introducing the core principles of probability theory, likelihood-based approaches, Bayesian inference, and the idea of measuring success through loss and risk. We did not focus on any particular model; rather, we underscored the overarching concepts that shape a wide range of learning algorithms. Understanding the probabilistic foundation is crucial for:

- Designing new models that capture the uncertainties of real-world data.
- Interpreting the behavior of algorithms as more data becomes available.
- Balancing model complexity with the risk of overfitting.

These ideas form the basis upon which nearly all modern machine learning methods are built. By grounding learning in probability theory, we gain a principled way to handle uncertainty, optimize performance metrics, and assess how well a hypothesis will generalize beyond the observed data.

Further Reading

- *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman.
- *Pattern Recognition and Machine Learning* by Christopher M. Bishop.
- *Machine Learning: A Probabilistic Perspective* by Kevin P. Murphy.

Key Takeaways

1. Probability theory provides the language to describe uncertainty in data and model parameters.
2. Learning can be seen as a process of fitting a model $p_\theta(y \mid x)$ or estimating a function $f_\theta(x)$ to minimize some form of expected risk.
3. Bayesian methods incorporate prior knowledge and produce posterior distributions over model parameters.
4. Generalization is at the heart of machine learning, and probabilistic tools guide how to build models that generalize well.