Lecture 1

谢丹
清华大学数学系

September 15, 2025

Core Components of Machine Learning

The basic ingredients for training ML models:

1. **Model**: Typically probabilistic - reflects the probabilistic nature of reality
2. **Data**: Represented as vectors, matrices, or tensors
3. **Training**: Optimization process to find function minima (using computer)
4. **Inference**: a): Making predictions on new data. b): Generative AI: generates new data

Machine learning essentially involves careful parameter tuning by using the computer (parallel computing and GPU)!

Machine Learning Applications

ML methods can solve diverse problems:

1. Regression analysis (linear and nonlinear curve fitting)
2. Classification tasks
3. Clustering problems
4. Generative AI:
   - Translation
   - Text, Image, audio, video generation

Learning Paradigms

- Supervised learning: Regression and classification
- Unsupervised learning: Clustering
- Reinforcement learning.

The New Era of Machine Learning

The Rise of Large Language Models

The rapid advancement of Large Language Models (LLMs) has unlocked unprecedented capabilities in artificial intelligence, creating a transformative shift in the field.

The Goal of This Course

This course is designed to provide a deep and thorough understanding of the mathematical foundations essential for:

- Developing and understanding traditional machine learning models.
- **Demystifying the core principles behind large language models.**: New ideas beyond traditional machine learning models.

We will build the toolkit to understand the revolution.

Course Information

Primary References

a: C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

b: C. Bishop, H. Bishop. *Deep Learning: Foundations and Concepts*. Springer, 2024.

c: D.J. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Additional Resources

Advanced Large Language Models (e.g., DeepSeek, Qwen3, GLM4.5, Kimi2, GPT-4/5) may be used for exploration and code assistance.

Essential Python Libraries for Machine Learning

Core Machine Learning Libraries

- **Scikit-learn** - Comprehensive general-purpose ML algorithms
- **PyTorch** - Flexible deep learning research framework
- **Transformers** - State-of-the-art natural language processing

When to Use Each Library

- Scikit-learn: Traditional ML tasks (classification, regression, clustering)
- PyTorch: Custom neural networks, research prototypes
- Transformers: NLP tasks, text generation, translation

Practical Details

- **Office Hours:** Thursday 2:00–4:00 PM
  Location: 理科楼 (Science Building) A-114 (Need change after October)
- **Grade Distribution:**
  40% Homework
  60% Final Exam and projects

CHAPTER 1: Probability I

Probability Fundamentals: Part I

A Brief Review

Our core assumption is that observed data is generated from an underlying probability distribution. Let's begin by reviewing some fundamental concepts.

Basic Probability Concepts

A one-dimensional probability model is described by a **probability density function (pdf)** or **probability mass function (pmf)** $p(x)$ satisfying:

$$p(x) \geq 0 \quad \text{for all } x \qquad \text{and} \qquad \int p(x)\,dx = 1$$

These functions can be classified into two main types:

- **Continuous**: $p(x)$ is defined for a continuous variable x.
- **Discrete**: $p(x)$ takes non-zero values only for a finite or countable set of points.

Multivariate Distributions

The concepts of probability extend naturally to higher dimensions.

- **Joint Probability:** $p(x, y)$
- **Marginal Density:** The probability of one variable, ignoring the other.

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

- **Conditional Probability:** $p(x|y)$ (probability of x given y) or $p(y|x)$.

$$p(x|y) = \frac{p(x, y)}{p(y)} \quad \text{(provided } p(y) > 0\text{)}$$

These densities are fundamentally related by the **Product Rule**:

$$\boxed{p(x, y) = p(x|y)\, p(y) = p(y|x)\, p(x)}$$

Relevance to Physics

- Statistical physics

Statistical distribution

$$\rho(p, q) = \frac{\exp(-E(p, q))}{Z}$$

p, q are the momentum and position variables. Z is the partition function. Famous model involves the Ising model, etc. Many insights of machine learning comes from physics.

Change variables

For $Y = g(X)$ with inverse $X = h(Y)$:

$$f_Y(y) = f_X(h(y)) \cdot \left|\frac{dh}{dy}\right|$$

- $f_X(h(y))$: Original PDF evaluated at inverse transformation
- $\left|\frac{dh}{dy}\right|$: Absolute Jacobian (ensures positivity)
- The absolute value handles both increasing/decreasing cases

Example: Linear Transformation

Let $X \sim$ Uniform$(0, 1)$, $f_X(x) = 1$ for $0 < x < 1$
Define $Y = 2X + 5$

1. Find inverse: $X = h(Y) = \frac{Y-5}{2}$
2. Find derivative: $\frac{dh}{dy} = \frac{1}{2}$
3. Apply formula:
$$f_Y(y) = 1 \cdot \left|\frac{1}{2}\right| = \frac{1}{2}$$
4. Find support: $5 < y < 7$

$\therefore Y \sim$ Uniform$(5, 7)$

Multivariate Case

For $\mathbf{Y} = g(\mathbf{X})$ with inverse $\mathbf{X} = h(\mathbf{Y})$:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(h(\mathbf{y})) \cdot |J|$$

Where J is the **Jacobian determinant**:

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} \end{vmatrix}$$

The absolute value $|J|$ represents the **volume scaling factor**.

Probability Characteristics

Key quantities to characterize a one-dimensional probability distribution:

1. **Mean** (expected value):

$$\mu = \mathbb{E}[x] = \int x\, p(x)\, dx$$

2. **Variance** (spread around mean):

$$\sigma^2 = \mathbb{E}[(x-\mu)^2] = \int (x-\mu)^2 p(x)\, dx$$

3. **Expectation value of a function** $f(X)$

$$\mathbb{E}(f(X)) = \int f(x) p(x)\, dx$$

Covariance in Two-Dimensional Distributions

Measuring joint variability

Definition

The covariance between two random variables X and Y is defined as:

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

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$ are the expected values.

Alternative Computation

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

Interpretation

- Positive covariance: X and Y tend to move together
- Negative covariance: X and Y tend to move oppositely
- **Zero covariance**: No linear relationship (but may have nonlinear dependence)

What is Entropy?

- Entropy measures the uncertainty or randomness of a random variable
- For discrete random variables: $H(X) = -\sum_{x \in \mathcal{X}} P(x) \log_2 P(x)$
- Measured in bits (when using base-2 logarithm)
- Higher entropy $=$ more uncertainty
- Lower entropy $=$ more predictability

Example: Fair 6-Sided Die

Problem Setup

- Random variable X: outcome of die roll $\{1, 2, 3, 4, 5, 6\}$
- Uniform distribution: $P(X = i) = \frac{1}{6}$ for $i = 1, \ldots, 6$

Entropy Calculation

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{6} P(i) \log_2 P(i) \\
&= -6 \times \left(\frac{1}{6} \log_2 \frac{1}{6} \right) \\
&= -\log_2 \frac{1}{6} = \log_2 6 \\
&\approx 2.585 \text{ bits}
\end{aligned}
$$

Interpretation

- Entropy of 2.585 bits means we need about 2.585 bits on average to encode each die roll
- This represents maximum uncertainty for a 6-outcome variable
- Compare to biased die: if $P(6) = 0.5$ and others $= 0.1$ each:

$$\begin{aligned} H(X) &= -\left[4 \times 0.1 \log_2 0.1 + 0.5 \log_2 0.5\right] \\ &= -\left[4 \times 0.1 \times (-3.3219) + 0.5 \times (-1)\right] \\ &= -\left[-1.3288 - 0.5\right] = 1.8288 \text{ bits} \end{aligned}$$

Lower entropy due to predictability!

Joint Entropy

Definition

For a two-dimensional random variable (X, Y) with joint distribution $P(x, y)$:

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log_2 P(x,y)$$

Example

Fair Coin Tosses Two fair coins: X = first coin, Y = second coin
$P(H,H) = P(H,T) = P(T,H) = P(T,T) = 0.25$
$H(X,Y) = -4 \times (0.25 \times \log_2 0.25) = -4 \times (0.25 \times -2) = 2$ bits

Conditional Entropy

Definition

Uncertainty of Y given knowledge of X:

$$H(Y|X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) \log_2 P(y|x)$$

Alternative Form

$$H(Y|X) = H(X, Y) - H(X)$$

Example

Dependent Variables If $Y = X$ (perfect correlation): $H(Y|X) = 0$
If X and Y independent: $H(Y|X) = H(Y)$

Chain Rule for Entropy

General Formula

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1})$$

For Two Variables

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Example

Application If $H(X) = 1$ bit, $H(Y|X) = 0.5$ bits
Then $H(X, Y) = 1 + 0.5 = 1.5$ bits

Mutual Information

Definition

Information shared between X and Y:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Relationships

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= D(P(x,y) \| P(x)P(y)) \end{aligned}$$

$D(P|Q)$ is the KL divergence, and is also called relative entropy (see discussion later).

Mutual Information Example I

The Binary Symmetric Channel (BSC) can be represented by the transition matrix:

$$P(Y|X) = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}$$

Where:

- $P(Y=0|X=0) = P(Y=1|X=1) = 1-\epsilon$
- $P(Y=1|X=0) = P(Y=0|X=1) = \epsilon$

Mutual Information Example II

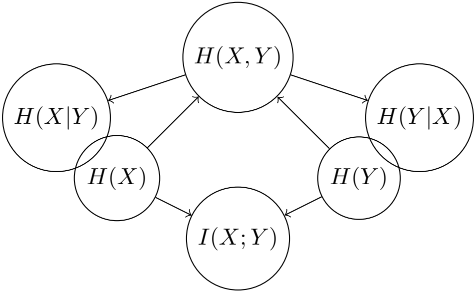
Example

Binary Symmetric Channel X = input, Y = output, error probability = ϵ
$P(X = 0) = P(X = 1) = 0.5$.

$$\begin{aligned} I(X;Y) &= 1 - H(\epsilon) \\ &= 1 + \epsilon \log_2 \epsilon + (1-\epsilon)\log_2(1-\epsilon) \end{aligned}$$

When $\epsilon = 0$ (no errors): $I(X;Y) = 1$ bit
When $\epsilon = 0.5$ (random): $I(X;Y) = 0$ bits

Summary of Relationships



$$H(X,Y) = H(X) + H(Y) - I(X;Y)$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Relative Entropy

Definition

Difference between distributions $P(x,y)$ and $Q(x,y)$:

$$D(P\|Q) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x,y) \log_2 \frac{P(x,y)}{Q(x,y)}$$

Properties

- $D(P\|Q) \geq 0$ (Gibbs' inequality)
- $D(P\|Q) = 0$ iff $P = Q$ almost everywhere
- Not symmetric: $D(P\|Q) \neq D(Q\|P)$

KL Divergence Example

Example

Two Distributions Let $P(x, y)$ be uniform:
$P(0,0) = P(0,1) = P(1,0) = P(1,1) = 0.25$
Let $Q(x, y)$ be:
$Q(0,0) = 0.5, Q(0,1) = 0.2, Q(1,0) = 0.2, Q(1,1) = 0.1$

$$\begin{aligned} D(P\|Q) &= 0.25 \log_2 \frac{0.25}{0.5} + 0.25 \log_2 \frac{0.25}{0.2} \\ &\quad + 0.25 \log_2 \frac{0.25}{0.2} + 0.25 \log_2 \frac{0.25}{0.1} \\ &\approx 0.25(-1) + 0.25(0.3219) + 0.25(0.3219) + 0.25(1.3219) \\ &\approx 0.2414 \text{ bits} \end{aligned}$$

What is the Gibbs Inequality?

The Gibbs inequality establishes a fundamental property of the Kullback-Leibler (KL) Divergence, also known as relative entropy.

Definition (KL Divergence)

For discrete probability distributions P and Q on a set $\mathcal{X}$:

$$D_{\mathsf{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

Gibbs Inequality

$$D_{\mathsf{KL}}(P \parallel Q) \geq 0$$

With equality **if and only if** $P(x) = Q(x)$ for all x.

A Useful Inequality

The entire proof rests on a simple inequality for the logarithm.

Inequality of the Logarithm

For any $t > 0$,

$$\log t \leq t - 1$$

Equality holds **if and only if** $t = 1$.

Step 1: Apply the Inequality

We want to relate the ratio $\frac{P(x)}{Q(x)}$ to the logarithm. Let's instead consider its inverse and apply our key inequality.
For any x where $P(x) > 0$, set $t = \frac{Q(x)}{P(x)}$. Since $Q(x) \geq 0$, we have $t > 0$.

$$\log\left(\frac{Q(x)}{P(x)}\right) \leq \frac{Q(x)}{P(x)} - 1$$

Step 2: Multiply and Sum

Multiply both sides of the inequality by $P(x)$ (a non-negative quantity, so the inequality is preserved):

$$P(x)\log\left(\frac{Q(x)}{P(x)}\right) \leq P(x)\left(\frac{Q(x)}{P(x)} - 1\right) = Q(x) - P(x)$$

Now, sum this inequality over all $x \in \mathcal{X}$:

$$\sum_x P(x)\log\frac{Q(x)}{P(x)} \leq \sum_x (Q(x) - P(x))$$

Step 3: Simplify the Right-Hand Side

The right-hand side (RHS) is a difference of sums:

$$\sum_x Q(x) - \sum_x P(x) = 1 - 1 = 0$$

Therefore, we have:

$$\sum_x P(x) \log \frac{Q(x)}{P(x)} \leq 0$$

Step 4: Rearrangement

Note that $\log \frac{Q(x)}{P(x)} = -\log \frac{P(x)}{Q(x)}$. Let's substitute this:

$$\sum_x P(x) \left(-\log \frac{P(x)}{Q(x)} \right) \leq 0$$

This is equivalent to:

$$-\sum_x P(x) \log \frac{P(x)}{Q(x)} \leq 0$$

Multiplying both sides by -1 (which reverses the inequality):

$$\sum_x P(x) \log \frac{P(x)}{Q(x)} \geq 0$$

Which is precisely:

$$\boxed{D_{\mathsf{KL}}(P \parallel Q) \geq 0}$$

When Does Equality Hold?

Recall our chain of inequalities. Equality in the Gibbs inequality holds **if and only if**:

$$\begin{aligned}
\log \frac{Q(x)}{P(x)} = \frac{Q(x)}{P(x)} - 1 \quad & \text{for all } x \text{ with } P(x) > 0 \\
\Rightarrow \quad \frac{Q(x)}{P(x)} = 1 \quad & \text{for all } x \text{ with } P(x) > 0 \\
\Rightarrow \quad P(x) = Q(x) \quad & \text{for all } x \text{ with } P(x) > 0
\end{aligned}$$

For x where $P(x) = 0$, the term in the sum is 0 by convention and does not affect the equality. Thus, overall equality holds **if and only if** $P(x) = Q(x)$ for all $x \in \mathcal{X}$.

Important Probability Distributions

1. **Gaussian (Normal) Distribution**:

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

   Mean μ, variance σ^2, which are the parameters.

2. **Bernoulli Distribution** (discrete):

$$\mathrm{Ber}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

   - $P(x=0) = 1-\mu$, $P(x=1) = \mu$
   - Mean μ, variance $\mu(1-\mu)$

Multivariate Distributions

Multivariate (n dimensional) Gaussian

$$P(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Mean vector $\boldsymbol{\mu}$ (n dimensional vector), covariance matrix Σ ($n \times n$ matrix).

Categorical Distribution

For K classes:

$$P(t=i) = p_i \quad (i = 1, \ldots, K), \quad \sum_{i=1}^{K} p_i = 1$$

Compact representation:

$$P(\mathbf{t}) = \prod^{K} p_i^{t_i}$$