Machine Learning

李帛阳 2300013215

//**The notes are primarily completed based on the class content, but annotations and supplementary information have been added for certain points.**//

Introduction

Definition: **Machine learning** is a technology that enables computers to automatically learn from data and make predictions or decisions through algorithms and models, serving as a subfield of artificial intelligence (AI). Its core objective is to empower computers to identify patterns and rules by analyzing large datasets—without explicit programming instructions—and subsequently build models adaptable to new data. Machine learning encompasses major types such as supervised learning, unsupervised learning, and reinforcement learning. It is widely applied in fields like image recognition, natural language processing (NLP), recommendation systems, and autonomous driving. Characterized by self-adaptive capabilities, automation, and strong generalization abilities, machine learning represents a paradigm of data-driven technological innovation.

Paradigm

1 **Supervised learning** is a machine learning paradigm where models learn patterns from labeled datasets (input-output pairs) to make predictions or decisions. It involves training algorithms on data where the correct answers (labels) are known, enabling the model to map inputs to outputs by minimizing prediction errors. Key tasks include classification (predicting discrete categories, e.g., spam detection) and regression (predicting continuous values, e.g., house prices). Common algorithms include linear regression, logistic regression, support vector machines (SVM), decision trees, and neural networks. Supervised learning is widely applied in fields like image recognition, natural language processing, and medical diagnosis, but its performance heavily relies on the quality and quantity of labeled training data. Evaluation metrics (e.g., accuracy, mean squared error) assess the model's ability to generalize to unseen data.

2 **Unsupervised learning** identifies hidden patterns or structures in unlabeled data without predefined output labels. It focuses on discovering inherent groupings (e.g., clustering like k-means) or reducing dimensionality (e.g., PCA), enabling applications such as customer segmentation, anomaly detection, and data compression. Unlike supervised learning, it relies solely on input features to reveal relationships within the dataset.

3 **Reinforcement learning** trains an agent to make sequential decisions by interacting with an environment to maximize cumulative rewards. The agent learns through trial-and-error, receiving feedback (rewards/punishments) after each action, rather than relying on labeled data. Key to RL are exploration vs. exploitation trade-offs and policy optimization, making it essential for robotics, game-playing (e.g., AlphaGo), and adaptive control systems.

4 **Self-supervised learning** is a machine learning paradigm that enables models to learn representations from unlabeled data by designing pretext tasks that automatically generate supervisory signals. Unlike traditional supervised learning, which depends on human-annotated labels, self-supervised learning exploits intrinsic patterns in the data (e.g., context prediction, data reconstruction) to train models. This approach has gained prominence in domains like natural language processing (e.g., BERT, GPT) and computer vision (e.g., contrastive learning, masked image modeling), where large-scale labeled datasets are impractical. By reducing reliance on manual annotations, self-supervised learning achieves scalability and generalization while maintaining competitive performance in downstream tasks such as classification, segmentation, and transfer learning. However, its effectiveness hinges on the design of pretext tasks and computational resources for pretraining.

Application

- **Classification**: In this task we wanna predict the discrete class label for elements. Here, we propose the concept of a discriminant function, which is a function used to separate different samples. Obviously, a discriminant function cannot directly predict a sample's label; instead, it aims to compute a probability value to measure the optimal label assignment. We design sample-label pairs (hereafter abbreviated as (x, y) , where x generally follows a specific distribution $X \sim \mathcal{D}_X$, and (X,Y) are independent and identically distributed (i.i.d.). The optimal predictor we select should, at a minimum, satisfy the condition that, under the joint distribution $\mathcal{D}_{XY}$ of X and Y, for a training dataset K, the predictor $\mathcal{P}$ takes the simplest possible form while achieving the highest accuracy on K. Detailed content and algorithms will be provided later. Commonly used approaches include Support Vector Machines (SVM), VC theory, and kernel space methods.

- **Regression**: In this task we wanna predict continuous output values (e.g., house price, temperature) by learning a mapping function $f : x \mapsto y \in \mathbb{R}$ from input features $x \in \mathbb{R}^n$ to a real-valued target y.We model the relationship as a linear function below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \ ,$$

where ε represents a Gaussian noise.And the objective usually is to minimize the Mean Squared Error(MSE):

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \beta^\top x_i)^2 \ ,$$

We default to accept that the hypothesis will assume the target y is a linear combination of features plus independent Gaussian noise.The optimal predictor $\hat{y} = \beta^\top x$ minimizes MSE under the conditional distribution $y|x \sim \mathcal{N}(\beta^\top x, \sigma^2)$. In this field we will introduce some common algorithms such as Ordinary Least Squares (OLS), Kernel Ridge Regression, Lasso and Ridge regression(regularized regression).

- **Graphic Models**: Some important algorithms contain Belief Propagation (BP), Loopy Belief Propagation, Tree-reweighted Message Passing, Monte Carlo Sampling (including Gibbs Sampling, Metropolis-Hastings), Junction Tree Algorithm (for exact inference) etc. .

Relative Resources

Specialized machine learning journals include Machine Learning and Journal of Machine Learning Research. Journals such as Neural Computation, Neural Networks, and IEEE Transactions on Neural Networks and Learning Systems also publish a large number of papers related to machine learning. In the field of statistics, journals like Annals of Statistics and Journal of the American Statistical Association occasionally carry articles on machine learning. Additionally, many IEEE Transactions journals (e.g., IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Image Processing, and IEEE Transactions on Signal Processing) feature interesting papers covering both machine learning theory and its applications.

Journals focused on artificial intelligence, pattern recognition, and signal processing also include articles about machine learning. For data - mining - oriented journals, examples are Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, and ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations Journal.

Major conferences in machine learning include "Neural Information Processing Systems (NIPS)" , "Uncertainty in Artificial Intelligence (UAI)" , "International Conference on Machine Learning (ICML)" , "European Conference on Machine Learning (ECML)" , and "Computational Learning Theory (COLT)". Conferences in fields like pattern recognition, neural networks, artificial intelligence, fuzzy logic, and genetic algorithms, as well as those centered on applications such as computer vision, speech technology, robotics, and data mining, also have special topics dedicated to machine learning.

High-dimensional Data

The teacher mentioned that the current research in machine learning and artificial intelligence should be **data-driven** in the first class.Therefore, we now begin with some important mathematical theorems as the starting point of our study.

1.[**Def1**] Let P and Q be two discrete probability distributions over the same sample space $\mathcal{X}$, where $P(x) \geq 0,\ Q(x) \geq 0$ and $\sum_{x\in\mathcal{X}} P(x) = \sum_{x\in\mathcal{X}} Q(x) = 1$, the **KL divergence** from Q to P is defined as:

$$D_{KL}(P||Q) = \sum_{x\in\mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right).$$

We can find that the formula is always non-negative and not symmetric about P&Q. In particular, if both P and Q follow a Bernoulli distribution, the above formula can be written as:

$$D_B(p||q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$$

2.[**Def2**] Let P be a discrete probability distribution over a sample space $\mathcal{X}$, where $P(x) \geq 0$ for all $x \in \mathcal{X}$ and $\sum_{x\in\mathcal{X}} P(x) = 1$. The **entropy** of P is defined as:

$$H(P) = -\sum_{x\in\mathcal{X}} P(x)\ln P(x)$$

3.[**Def3**] With the same condition of Def 1, the **cross-entropy** between P and Q is defined as:

$$H(P,Q) = -\sum_{x\in\mathcal{X}} P(x)\ln Q(x)$$

And we can find an important identical function:

$$H(P,Q) = H(P) + D_{KL}(P||Q)$$

All of the above natural logarithms (ln) should be replaced with base-2 logarithms ($\log_2$) when measuring in bits (the single unit of information).

Theorem 1 (**Markov's Inequality**)**.** *Let x be a non-negative random variable. For any $a > 0$, we have:*

$$\mathbb{P}(x \geq a) \leq \frac{\mathbb{E}(x)}{a}$$

Proof: For a continuous non-negative random variable x with probability density p:

$$\begin{aligned}\mathbb{E}(x) &= \int_0^\infty xp(x)dx &(1)\\ &= \int_0^a xp(x)dx + \int_a^\infty xp(x)dx &(2)\\ &\geq \int_a^\infty xp(x)dx &(3)\\ &\geq a\int_a^\infty p(x)dx &(4)\\ &= a\cdot\mathbb{P}(x\geq a) &(5)\end{aligned}$$

This completes the proof for the continuous case. The discrete case follows similarly: □

Corollary 1.

$$\mathbb{P}(x\geq b\mathbb{E}(x))\leq\frac{1}{b}$$

Corollary 2. *if the random variable $X\geq 0$ has finite 1st,2nd,$\cdots$,k-th moments, we have the following formula:*

$$\mathbb{P}(X\geq t)\leq\min_{i\in[k]}\frac{\mathbb{E}(X^i)}{t^i}$$

Proof: For we can reckon on $X\geq t$ as $X^i\geq t^i$. If k tends to the positive infinity , we can consider infimum of the moment sequence(**Chernoff bound**):

$$\mathbb{P}\left(\sum_{i=1}^n X_i - n\mathbb{E}(X)\geq\varepsilon\right)\leq\min_{\lambda\geq 0}\left[\prod_{i=1}^n\left(\mathbb{E}[e^{\lambda(X_i-\mathbb{E}(X_i))}]\right)e^{-\lambda\varepsilon}\right],\quad where\ X_i\sim\mathbb{P}_{X_i}(\cdot)$$

□

Theorem 2 (**Chebyshev's inequality**). *Let x be a random variable, then for $c>0$:*

$$\mathbb{P}(|x-\mathbb{E}(x)|\geq c)\leq\frac{Var(x)}{c^2} \tag{6}$$

Proof: Since:

$$\mathbb{P}(|x-\mathbb{E}(x)|\geq c)=\mathbb{P}\left((x-\mathbb{E}(x))^2\geq c^2\right)$$

Let $y=|x-\mathbb{E}(x)|^2$ be a non-negative random variable with $\mathbb{E}(y)=Var(x)$. By Markov's inequality:

$$\mathbb{P}(|x-\mathbb{E}(x)|\geq c)\leq\frac{\mathbb{E}\left(|x-\mathbb{E}(x)|^2\right)}{c^2}=\frac{Var(x)}{c^2}$$

□

By advantage of Chebyshev's inequality and the property $(x)=(x_i)$ we can prove Khinchin's law of large numbers. □

Corollary 3. *If $X_1, X_2, \cdots, X_n$ i.i.d. and $X_i \sim B(1,p)$, then:*

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i - p| \geq \delta\right)\left\{\sim e^{-\mathcal{O}(n)}(Using\ \ global\ \ information \to CLT)\right\} \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

More generally, we can apply the Chernoff bound to the previous inequality without the absolute value constraint:

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum X_i - p \geq \delta\right) &\leq \inf_{t\geq 0} e^{-n(p+\delta)t}\mathbb{E}(e^{t\sum X_i}) \\ &= \inf_{t>0}\{exp(-nt(p+\delta)\cdot(pe^t+1-p)^n\} \\ &= e^{-n\cdot D_B(p+\delta||p)}\end{aligned}$$

If the X_i are independent and identically distributed random variables in [0,1], with $\mathbb{E}(X_i) = p$ and all moments existing, then the following holds:

$$\mathbb{P}\left(\frac{1}{n}\sum X_i - p \geq d\right) \leq e^{-nD_B(p+\delta||p)}$$

This is because of Chernoff bound and the following inequality:

$$\begin{aligned}pe^t + 1 - p &= \mathbb{E}(xe^{t\times 1} + (1-x)e^{t\times 0}) \\ &\geq \mathbb{E}(e^{t(x\times 1 + 0\times(1-x))}) \\ &= \mathbb{E}(e^{tX})\end{aligned}$$

Theorem 3 (**Moment Genarating Function(MGF)**). *The **moment generating function (MGF)** of a random variable X is defined as:*

$$M_X(t) = \mathbb{E}[e^{tX}]$$

provided the expectation exists for t in a neighborhood of 0. And The n-th raw moment (about the origin) of X is given by the n-th derivative of $M_X(t)$ evaluated at $t = 0$:

$$\mathbb{E}[X^n] = M_X^{(n)}(0) = \frac{d^n}{dt^n}M_X(t)\bigg|_{t=0}$$

***Example**: Normal Distribution*

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Theorem 4 (**Riemann-Euler's Identity**). *For $s > 1$, $s \in \mathbb{R}$, the Riemann zeta function has the Euler product representation:*

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s} = \prod_{p\in\ Prime}\frac{1}{1-p^{-s}} \tag{7}$$

where p runs over all prime numbers.

Law of Large Numbers

- **Chebyshev's Law of Large Numbers:**
  Let $\{x_i\}_{i=1}^{\infty}$ be a sequence of independent random variables with finite expectations $\mathbb{E}(x_i)$ and variances $\mathrm{Var}(x_i)$. If $\{\mathrm{Var}(x_i)\}$ is bounded, then for all $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(x_k)\right| < \varepsilon\right) = 1$$

- **Bernoulli's Law of Large Numbers:**
  Let μ_n be the number of occurrences of event A in n independent trials, with success probability p in each trial. Then for all $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) = 1$$

- **Khinchin's Law of Large Numbers:**
  Let $\{x_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed (i.i.d.) random variables with $\mathbb{E}(x_i) = \mu$. Then for all $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} x_k - \mu\right| < \varepsilon\right) = 1$$

  Note: This is a direct consequence of Chebyshev's inequality:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^{n} x_i}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(x)}{n\varepsilon^2}$$