大模型之机器学习

# 机器学习的概念

目的：通过计算机系统逐步改善特定任务的性能

方法：海量数据 🡪 训练（函数系数或者参数的选择） 🡪 模型（数据表达或者概率表达） 🡪 预测

## 1.1 人工智能、机器学习、深度学习的关系



## 1.2 发展历史

### 早期探索： 20世纪50-70年代

符号学派尝试赋予机器逻辑推理能力；

图灵提出的“图灵测试”；

### 知识驱动与专家系统：20世纪70-80年代

逻辑推理 + 知识驱动；

专家系统、决策树（ID3算法）、贝叶斯定理的应用

### 数据驱动与统计学：20世纪80年代-21世纪

神经网络（连接主义学派）、统计学、互联网的大量数据；

统计学（支持向量机SVM）：决策树、随机森林

行为与环境交互（行为主义学派）

### 深度学习崛起：21世纪

深度学习：数据规模快速增长、GPU的发展、神经网络层数足够的多、CNN卷积神经网络、GAN生成对抗网络、Transformer

### 大模型与通用人工智能：2020年代

自然语言处理 多模态模型 自监督学习 强化学习和深度学习结合

## 1.3 机器学习的应用

图像处理与计算机视觉、自然语言处理、推荐系统、医疗与健康、金融科技、自动驾驶与智能交通；

## 1.4 基本术语

数据集：训练集（模型的训练数据）、测试集（性能的验证）、验证集（超参数的调整）；

样本、特征、特征向量（x）、标签(y)、参数、超参数、模型（数据表达式）

# 2. 机器学习基本理论

## 2.1 机器学习三要素

模型：总结数据的内在规律，用数学语言描述的参数 ；

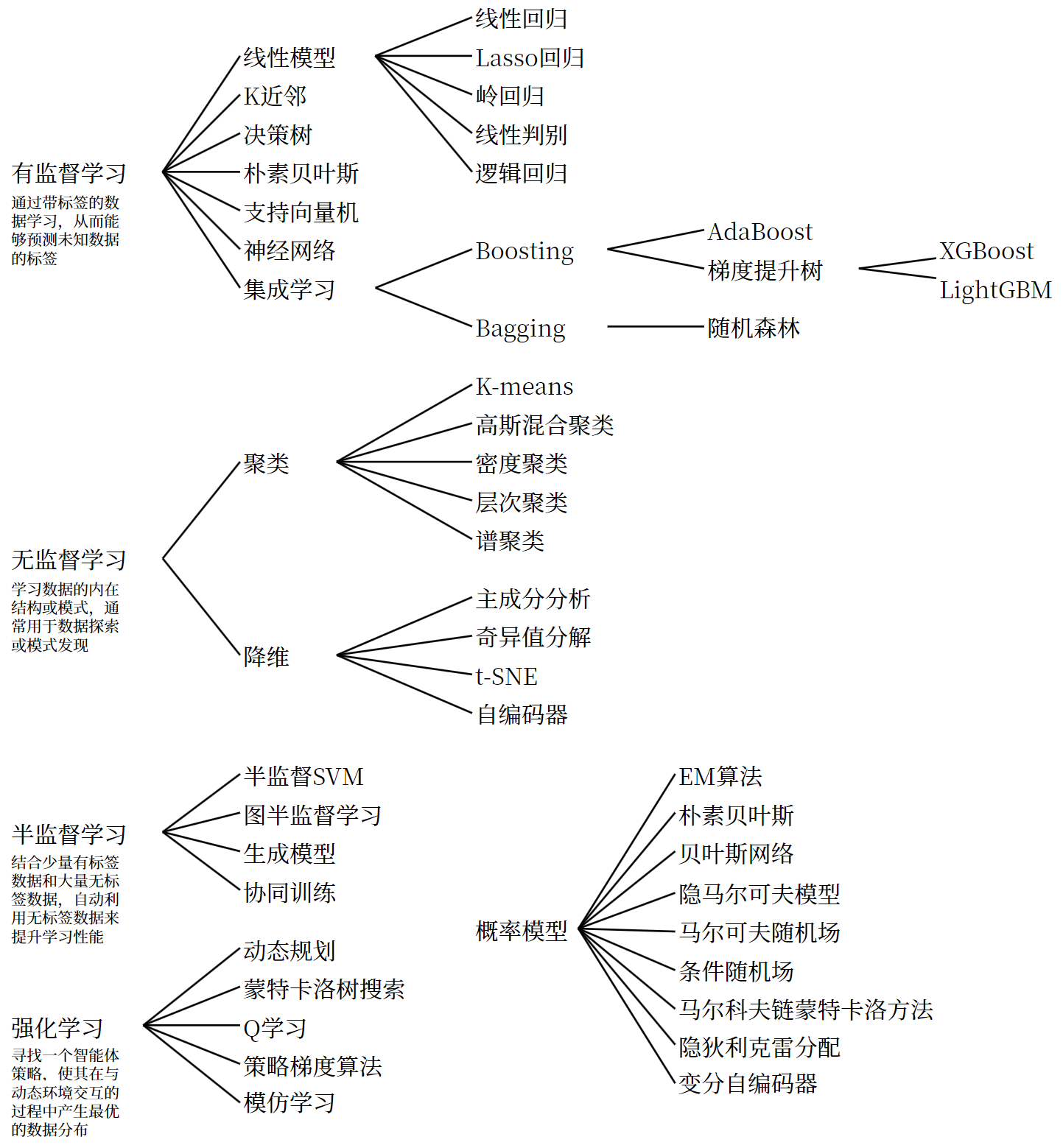
策略：选取最优模型的评价准则；

算法：选取最优模型的具体方法；

## 2.2 机器学习的分类

主要了分类：有监督学习、无监督学习、半监督学习 以及强化学习；

其他的分类：



大模型主要应用的算法是监督学习里的深度神经网络学习（解决问题的大一统的方法），即黑盒中不停地添加神经网络的层级来提高准确度；

## 2.3 建模流程

收集数据、清洗数据、特征工程、选择算法、训练模型、模型评估、模型优化、模型部署

图形用户界面, 应用程序

中度可信度描述已自动生成

## 2.4 特征工程

### 2.4.1 概念

1. 通过对原始数据的处理、转换、和构造，生成新的特征或者选择有效的特征，从而提高模型的性能；

2. 特征工程是一个迭代过程；

3. 本质是数据预处理的阶段；

### 2.4.2 内容

特征选择（减少特征）：

过滤法（低方差），相关系统）、包裹法？、嵌入法？

特征转换：

归一法（缩放到特定范围）、标准化（减去均值并除以方差）、对数变换、类别变量的编码（独热法（二进制编码））；

特征构造：

交互特征、统计特征、日期和时间特征；

特征降维（减少特征）：

低方差过滤法、相关系数法、主成分分析、线性判别分析、t-SNE、自编码器；

### 2.4.3 常用方法

特征选择、特征降维（维度灾难）

#### 低方差过滤

1. 相关代码

F:\MyStudy\LLM\my\_codes\my\_study\_llm\_training\machine\_learning\feature\1\_variance\_filter.ipynb 搜索 “=低方差过滤法=”

1. 该特征方差越小，就越不选择；

#### 相关系数法

##### 皮尔逊相关系数

1. 公式

皮尔逊相关系数（Pearson Correlation）用于衡量两个变量的线性相关性，取值范围。

正相关：值接近1，说明特征随目标变量增加而增加。

负相关：值接近-1，说明特征随目标变量增加而减少。

无关：值接近0，说明特征和目标变量无明显关系。

注：

1. 重点关注的是衡量两个变量的线性相关性
2. 相关代码

F:\MyStudy\LLM\my\_codes\my\_study\_llm\_training\machine\_learning\feature\2\_pearson\_filter.ipynb

##### 斯皮尔曼相关系数

1. 公式

斯皮尔曼相关系数（Spearman’s Rank Correlation Coefficient）的定义是等级变量之间的皮尔逊相关系数。用于衡量两个变量之间的单调关系，即当一个变量增加时，另一个变量是否总是增加或减少（不要求是线性关系）。适用于非线性关系或数据不符合正态分布的情况。

其中：

是两个变量的等级之差

是样本数

斯皮尔曼相关系数的取值范围为：

注：

1. 等级也就是排序;
2. 重点是等级变量之间的皮尔曼祥光系数；

相关代码

F:\MyStudy\LLM\my\_codes\my\_study\_llm\_training\machine\_learning\feature\3\_spearman.ipynb

##### 主成分分析（PCA）

主成分分析（Principal Component Analysis，PCA）是一种常用的降维技术，通过线性变换将高维数据投影到低维空间，同时保留数据的主要变化模式。

图表, 散点图

描述已自动生成

使用sklearn.decomposition.PCA进行主成分分析。参数n\_components若为小数则表示保留多少比例的信息，为整数则表示保留多少个维度。

注：

1. 高维数据投影到低维空间；
2. sklearn.decomposition.PCA分析，一般会用剩余成分的百分比作为降维的结果，这里用参数n\_components作为演示，正太分布方差越小的会被降掉（方差贡献率）

相关代码

F:\MyStudy\LLM\my\_codes\my\_study\_llm\_training\machine\_learning\feature\4\_pca.ipynb

## 模型评估和模型选择（重点）

### 损失函数

概念：是关于模型参数的函数，用来度量预测偏差的程度；

公式（*L(Y,f(X))*）：

常见的损失函数有：

* + - * 1. 0-1损失函数

注：类比向量范数0

* + - * 1. 平方损失函数

注释：类比向量范数2

* + - * 1. 绝对损失函数

注：类比向量范数1

* + - * 1. 对数似然损失函数

注：概率计算是使用似然函数，取log相乘转相加，取负号为了求极小值

### 经验误差（经验风险）

模型f(X)在训练集上的平均误差（损失函数的平均误差）；

公式：

### 泛化误差

上述相同的公式，模型f(X)在测试集数据上的平均误差（（损失函数的平均误差））

### 经验风险最小化

当经验风险最小时，模型最优的策略被称为经验风险最小化（empirical risk minimization ERM）

### 欠拟合与过拟合

**拟合：**

找到数据的规律

**欠拟合：**

高偏差；

原因是：模型复杂度不够、特征不足、训练补充分、过强的正则化；

**过拟合：**

模型过于复杂，高方差；

原因：模型复杂度过高（函数复杂或者参数太多）、特征过多、训练过长；

解决办法：减少模型复杂度、增加训练数据、使用正则化（弱化特征）、交叉验证

**泛化能力：**

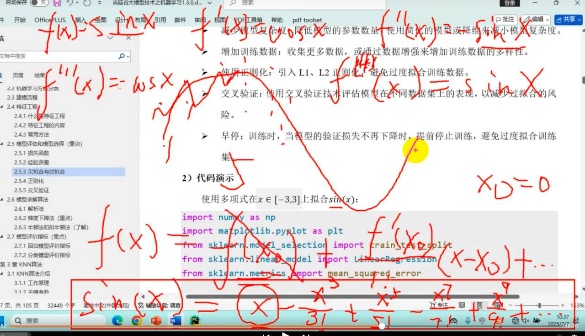
通用能力

图片包含 图示

描述已自动生成

#### 案例：

使用多项式在上拟合，如下使用的是泰勒展开，且用的是5阶的多项式即可



具体代码可看：