Introduction

What is Machine Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

程序通过利用经历E，在T的任务中获得了性能改善，就说关于T和P，对E进行了学习。

Example: playing checkers.

E = the experience of playing many games of checkers

T = the task of playing checkers.

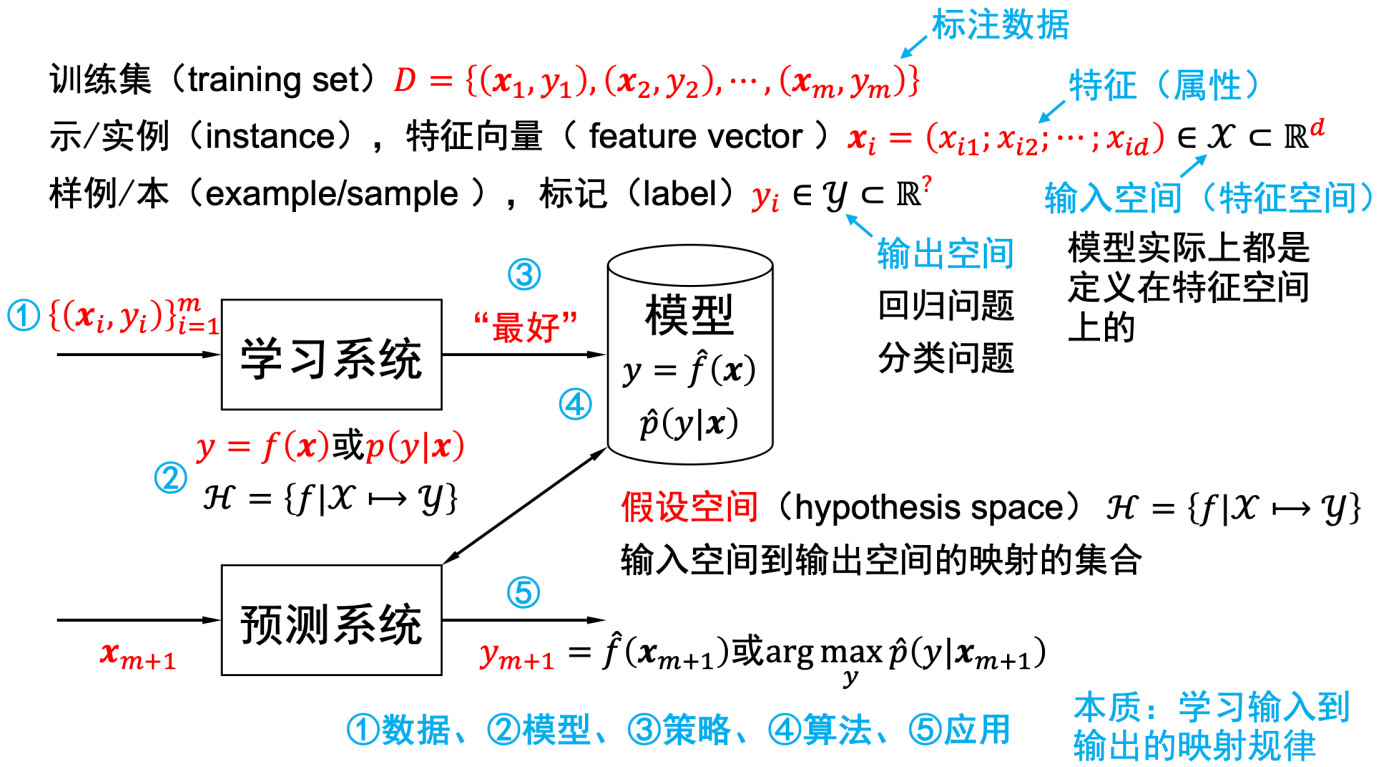
P = the probability that the program will win the next game.

Supervised Learning

example:线性，神经，支持向量机，决策树，集成学习

- Regression 回归：输出连续值
- Classification 分类：输出离散值{0,1}

从标注数据中学习预测模型的问题。



训练集，特征向量，标记。

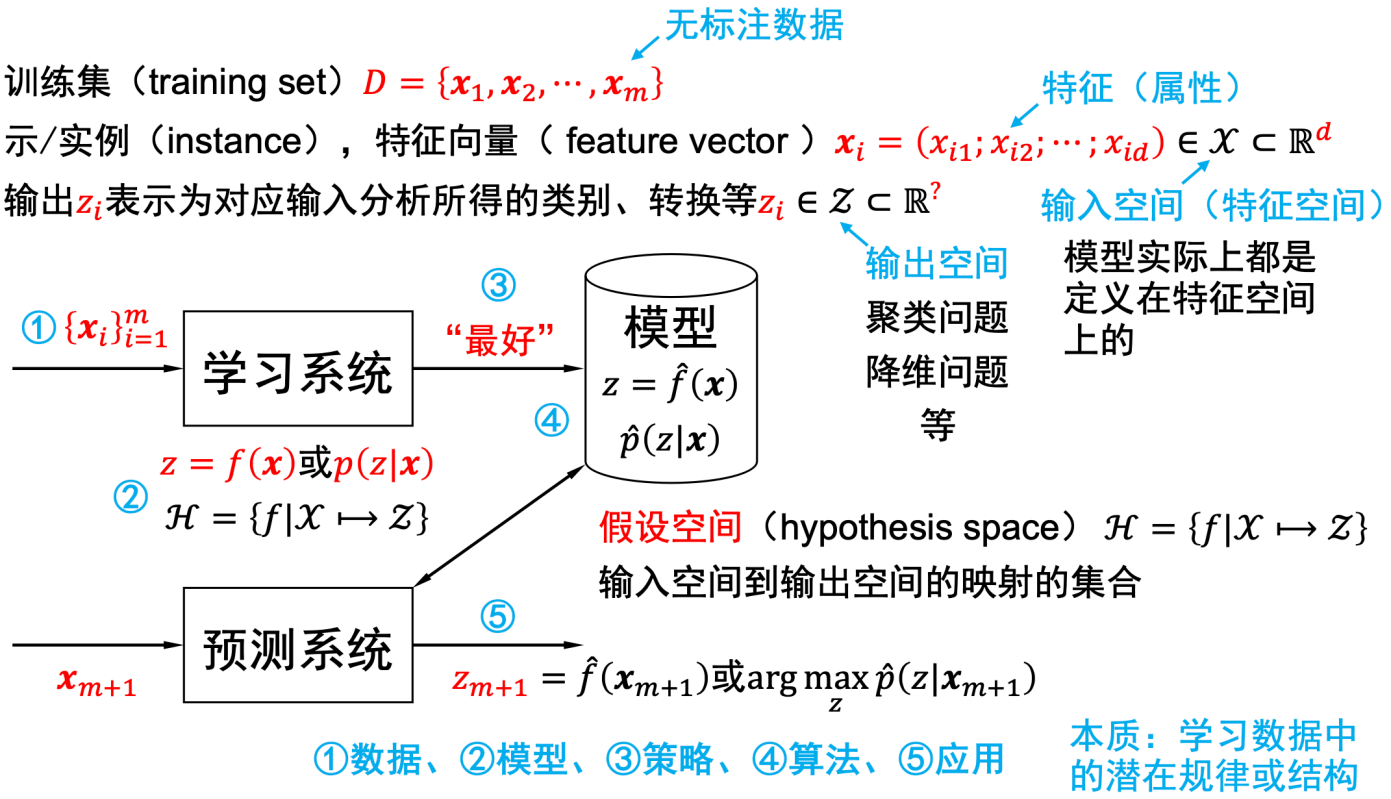
本质：学习输入到输出的映射关系。

Unsupervised Learning

example:聚类，降纬

在脱离标签的情况下，从数据中直接找出数据的结构。

“自然”得到的数据，没有标签。



本质：学习数据潜在的结构。

Reinforcement learning

example:有模型，无，k—摇臂赌博机

智能系统在与环境交互的过程中，通过奖励机制学习最优策略。

长期奖励的最大化为目的，不断的试错，从所有的策略中学得最优策略。

本质：学习如最优的序贯策略。

机器学习关键要素

数据，模型，策略，算法

数据

一般地，令$D = \{x_1, x_2, x_3, \cdots, x_m\}$表示包含$m$个示例的数据集，每个示例由$d$个属性描述，则每个示例$x_i = (x_{i1}; x_{i2}; \cdots; x_{id})$是$d$维样本空间$\mathcal{X}$中的一个（列）向量，$x_i \in \mathcal{X}$，其中$x_{ij}$是$x_i$在第$j$个属性上的取值，$d$称为样本$x_i$的“维数”（dimensionality）。

一般地，用$(\boldsymbol{x}_i, y_i)$表示第i个样例，其中$y_i \in \mathcal{Y}$是示例$\boldsymbol{x}_i$的标记，$\mathcal{Y}$是所有标记的集合，亦称为“标记空间” （ label space）或输出空间。

训练集（ training set）	测试集（ testing set）
训练数据（ training data）	测试数据（ testing data）
训练样本（ training sample）	测试样本（ testing sample）

模型

当获得数据集后，机器学习首要考虑的问题是学习什么样的模型。

- 从数据中学得模型的过程称为“学习” 或“训练”
- 学得模型对应了关于数据的某种潜在的规律，即“假设”（hypothesis）

模型属于由输入空间到输出空间的映射的集合，即假设空间：

- $\mathcal{H} = \{f|\mathcal{X} \mapsto \mathcal{Y}\}$（监督学习）
- $\mathcal{H} = \{f|\mathcal{X} \mapsto \mathcal{Z}\}$（无监督学习）

假设空间可以定义为（非概率）函数的集合：

- $\mathcal{H} = \{f|y = f(\boldsymbol{x})\}$（监督学习）

非概率模型

$y = f(\boldsymbol{x})$

$z = f(\boldsymbol{x})$

假设空间可以定义为条件概率的集合：

- $\mathcal{H} = \{p|p(y|\boldsymbol{x})\}$（监督学习）

概率模型

$p(y|\boldsymbol{x})$

$p(z|\boldsymbol{x})$

策略

ERM

在训练数据集、假设空间、损失函数都确定的情况下， ERM策略认为，经验风险最小的模型是最优的模型：

$$\hat{f} = \min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\big(y_i, f(\boldsymbol{x}_i)\big)$$

当样本容量足够大时，经验风险最小化能够保证有很好的学习效果，在现实中被广泛采用。

SRM

样本容量较小时，ERM模型产生过拟合现象，SRM就是预防这种结果出现的。
结果风险小的模型对训练数据、未知的测试数据都有良好的预测。

> **在训练数据集、假设空间、损失函数都确定的情况下，SRM策略在经验风险的基础上加上表示模型复杂度的正则化项（regularization）形成结构风险，并认为结构风险最小的模型是最优的模型：**

$$\hat{f} = \min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(y_i, f(\boldsymbol{x}_i)\right) + \lambda \Omega(f) \right\}$$

> **其中，$\Omega(f)$ 为模型的复杂度，模型f 越复杂$\Omega(f)$ 就越大；$\lambda \geq 0$ 是系数，用以权衡经验风险和模型复杂度。**

名次解析

泛化能力

模型对未知数据的预测能力

过拟合

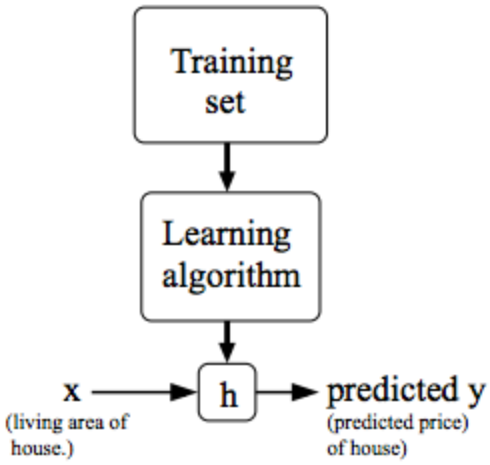
学习能力太强；克服方法：优化目标中增加正则项，增加训练数据量，减少特征量，进行模型选择

机器学习器把训练样本学的太好，将训练样本本身的特点当作所有样本的一般性质，导致泛化能力下降。

欠拟合

学习能力太弱；克服方法：增加模型复杂度，训练轮数

通过训练算法得出一个h(hypothesis)



Cost Function

衡量模型的函数，通常用$J(\theta)$表示