Learning

吉建民

USTC
jianmin@ustc.edu.cn

2024 年 5 月 7 日

Used Materials

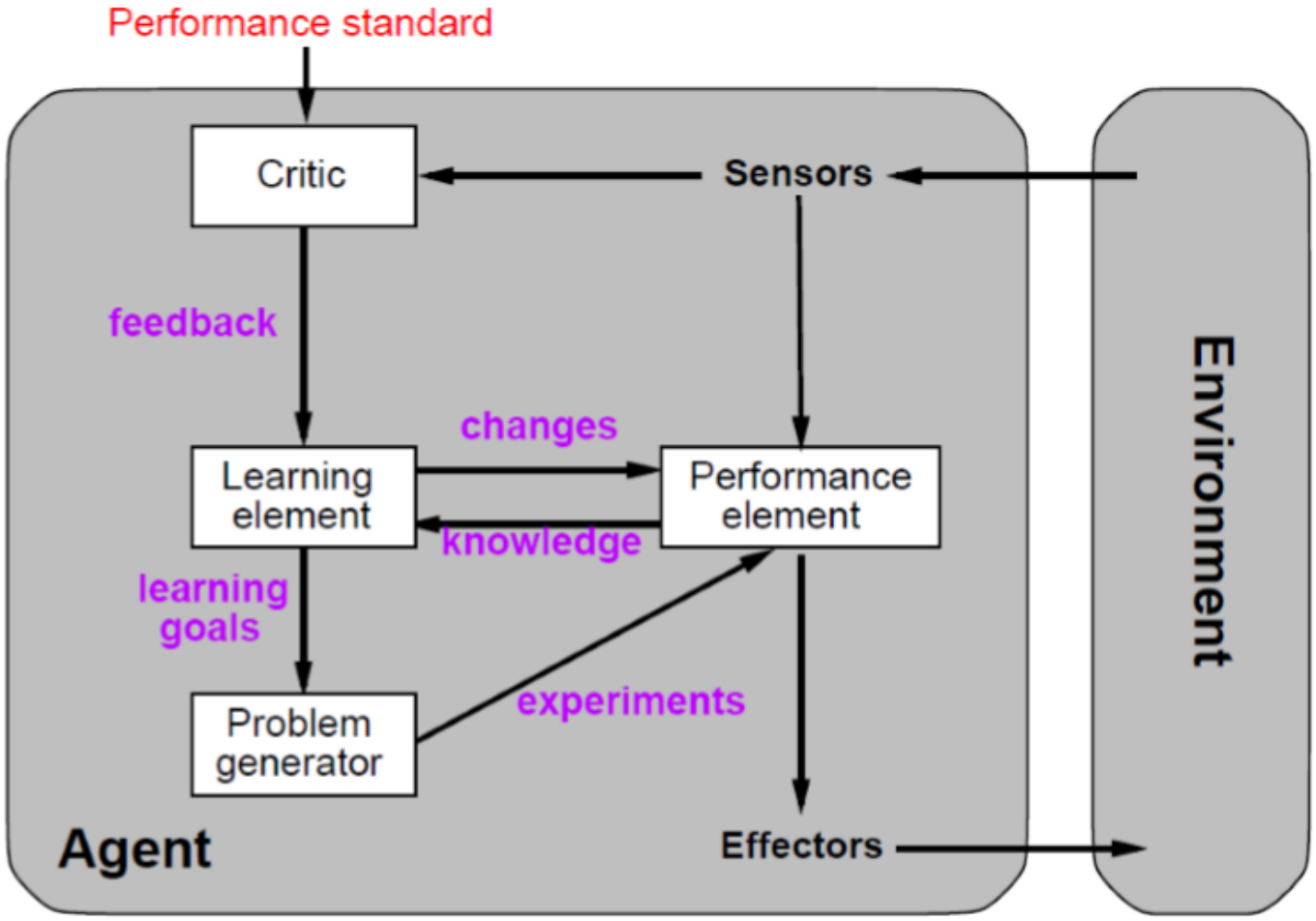
Disclaimer: 本课件采用了 S. Russell and P. Norvig's Artificial Intelligence –A modern approach slides, 徐林莉老师课件和其他网络课程课件，也采用了 GitHub 中开源代码，以及部分网络博客内容

Table of Contents

Learning

- Learning is essential for unknown environments, –i.e., when designer lacks omniscience (全知)
- Learning is useful as a system construction method, –i.e., expose the agent to reality rather than trying to write it down
- Learning modifies the agent's decision mechanisms to improve performance

Learning agents



Learning agents

Design of a learning element is affected by

- Which components of the performance element are to be learned
- What feedback is available to learn these components
- What representation is used for the components

Learning agents

> *Machine learning is an interdisciplinary field focusing on both the mathematical foundations and practical applications of systems that learn, reason and act.*
> **机器学习是一个交叉学科的领域，着重于研究具有学习、推理和行动的系统所需要的数学基础以及实际应用**

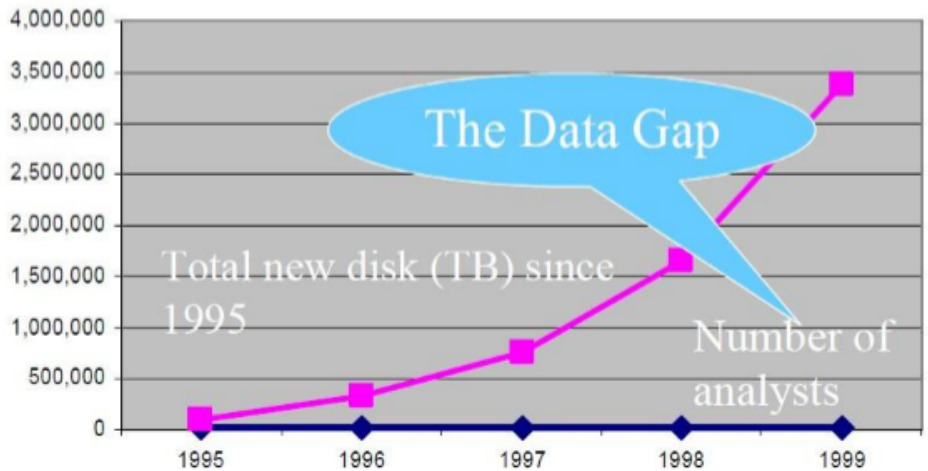
Other related terms: Pattern Recognition (**模式识别**), Neural Networks (**神经网络**), Data Mining (**数据挖掘**), Statistical Modeling (**统计模型**) ...

Using ideas from: Statistics, Computer Science, Engineering, Applied Mathematics, Cognitive Science (**认知科学**), Psychology (**心理学**), Computational Neuroscience (**计算神经学**), Economics

The goal of these lectures: to introduce important concepts, models and algorithms in machine learning.

Why machine learning?

- Large amounts of data
  - Web data
  - Medical data
  - Biological data
- Expensive to analyze by hand
- Computers become cheaper and more powerful



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Why machine learning?

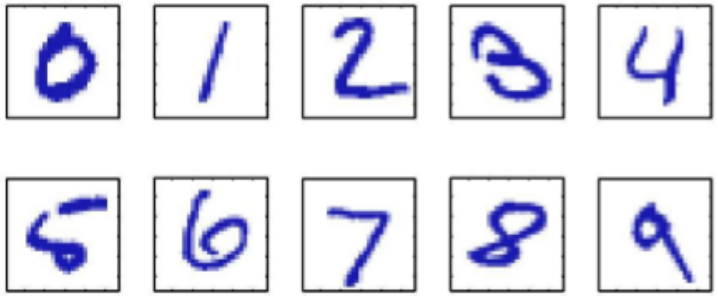
What is machine learning useful for?

Automatic speech recognition (自动语音识别)

Now most Speech Recognizers or Translators are able to learn — the more you play/use them, the smarter they become

Computer vision

e.g., object, face and handwriting recognition



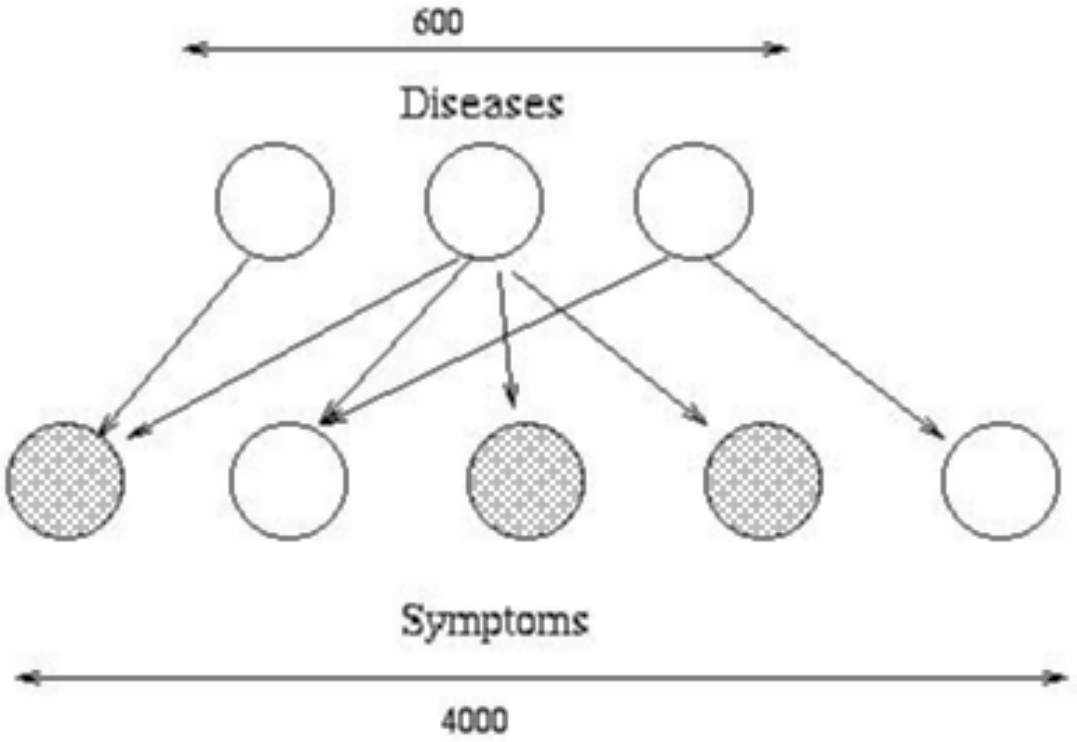
Information retrieval–信息检索

Reading, digesting, and categorizing a vast text database is too much for human

- Web Pages Retrieval (检索)
- Categorization (分类)
- Clustering (聚类)
- Relations between pages

Financial prediction

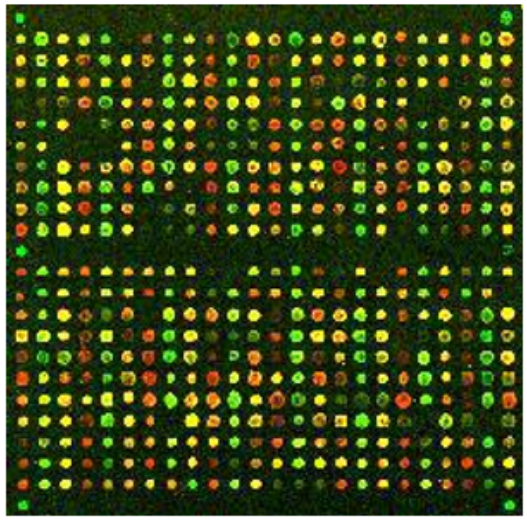
Medical diagnosis (医学诊断)

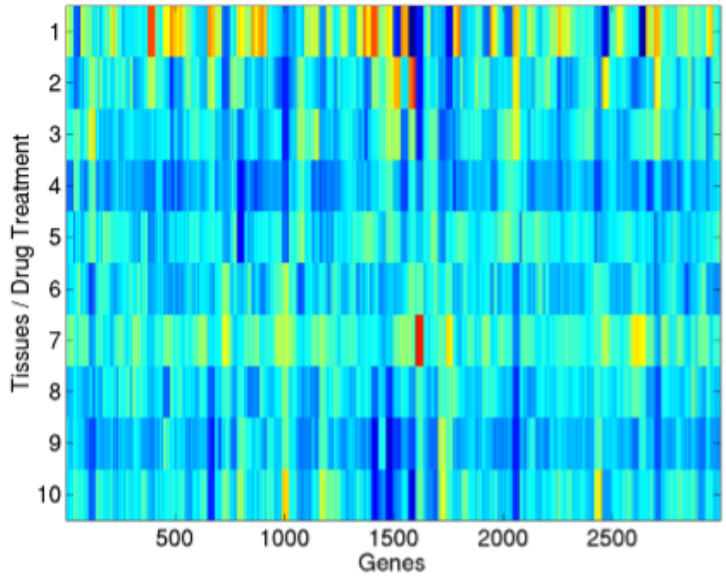


(image from Kevin Murphy)

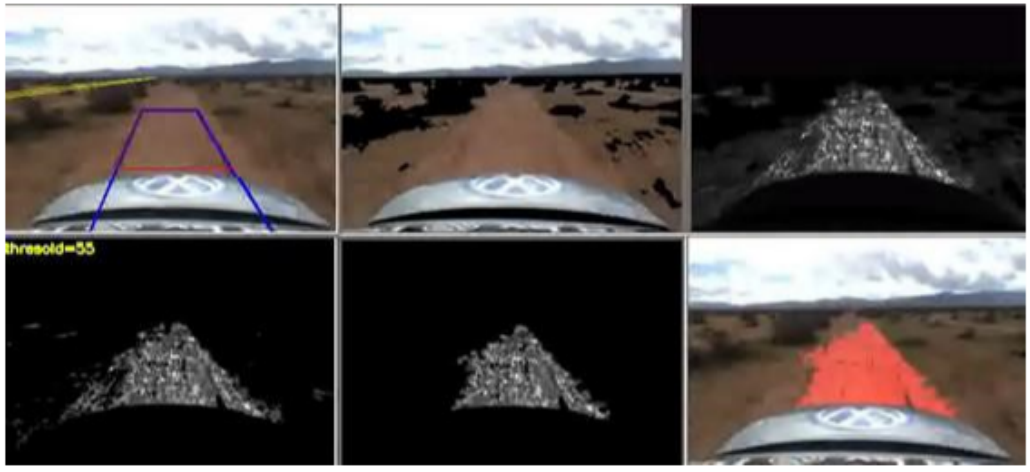
Bioinformatics (生物信息学)

e.g. modeling gene microarray (微阵列) data, protein structure prediction

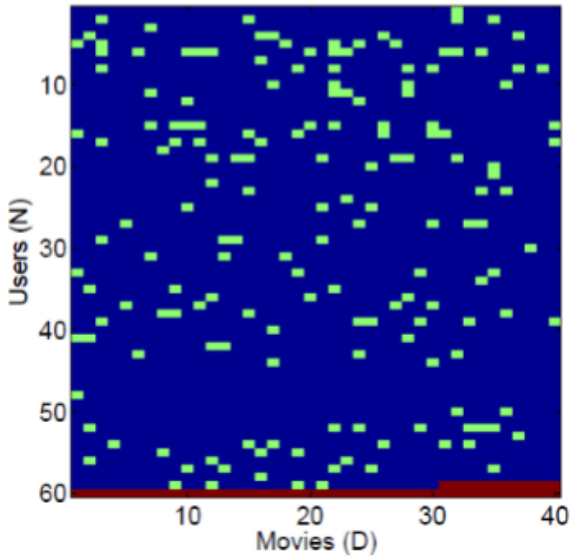




Robotics



Movie recommendation systems

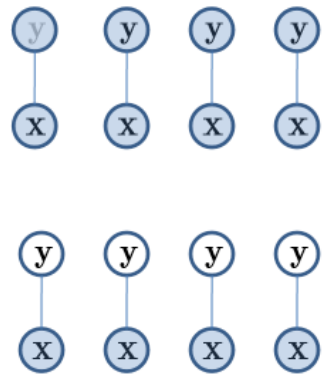


Challenge: to improve the accuracy of movie preference predictions Netflix $1m Prize.

Three Types of Learning

Imagine an agent or machine which experiences a series of sensory inputs: $x_1, x_2, x_3, x_4, ...$

- Supervised learning (监督学习): The machine is also given desired outputs $y_1, y_2, ...$, and its goal is to learn to produce the correct output given a new input.
- Unsupervised learning (无监督学习): outputs $y_1, y2, ...$Not given, the agent still wants to build a model of x that can be used for reasoning, decision making, predicting things, communicating etc



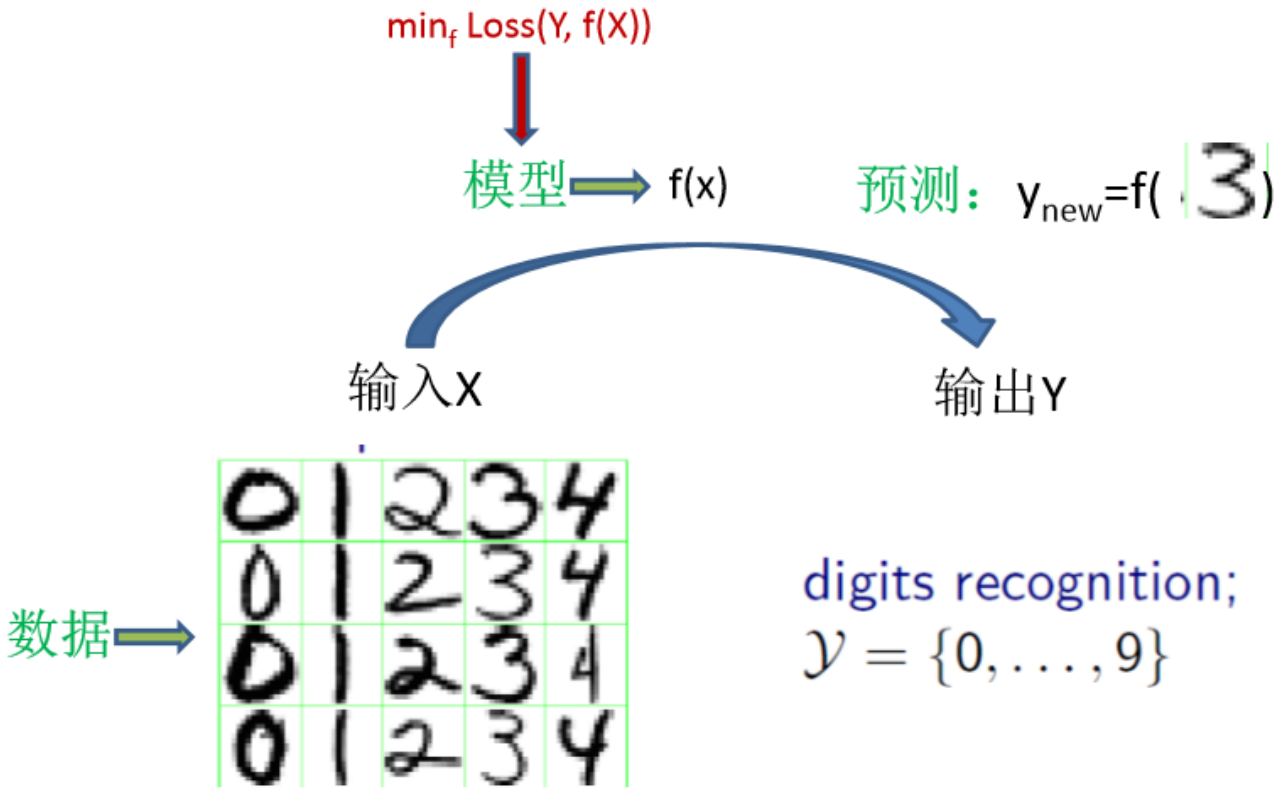
- Reinforcement learning (强化学习): The machine can also produce actions $a_1, a_2, ...$ which affect the state of the world, and receives rewards (or punishments) $r_1, r_2, ...$ Its goal is to learn to act in a way that maximizes rewards in the long term.

Key Ingredients

- Data The data set D consists of N data points: $D = \{x_1, x_2 \ldots, x_N\}$
- Predictions We are generally interested in predicting something based on the observed data set. Given D what can we say about x_{N+1}?
- Model To make predictions, we need to make some assumptions. We can often express these assumptions in the form of a model, with some parameters (参数)

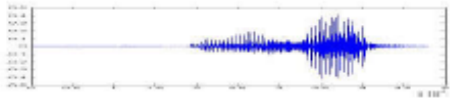
Given data D, we learn the model parameters, from which we can predict new data points.

Key Ingredients



Machine Learning ≈ Looking for a Function

- Speech Recognition

$f($  $)=$ “How are you”

- Image Recognition

$f($  $)=$ “Cat”

- Playing Go

$f($  $)=$ “5-5” (next move)

- Dialogue System

$f($ “Hi” $)=$ “Hello”

(what the user said) (system response)

Framework

Image Recognition:

$f($  $)=$ "cat"

A set of function

Model

$f_1, f_2 \cdots$

$f_1($  $)=$ "cat"

$f_2($  $)=$ "money"

$f_1($  $)=$ "dog"

$f_2($  $)=$ "snake"

Framework

Image Recognition:

$f($  $) =$ "cat"

A set of function

Model

$f_1, f_2 \cdots$

Goodness of function f

Training Data

$f_1($  $) =$ "cat"

$f_2($  $) =$ "money"

Better!

$f_1($  $) =$ "dog"

$f_2($  $) =$ "snake"

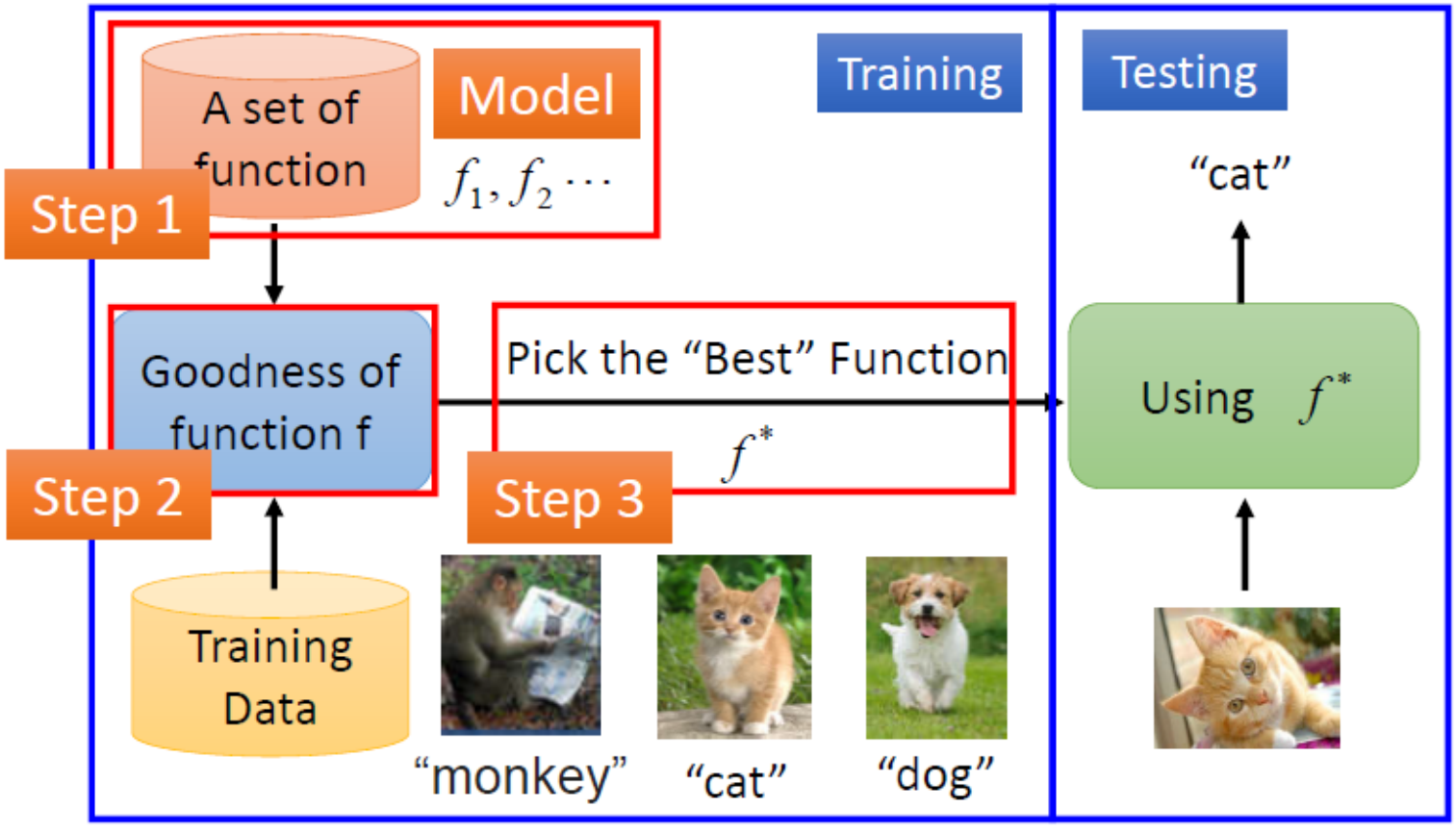
Supervised Learning

function input:   

function output: "monkey" "cat" "dog"

Framework

Image Recognition:



Learning = Representation + Evaluation + Optimization

- 学习 = 表示 + 评价 + 优化
  - 表示（Representation）：确定假设空间（hypothesis space）
    - Pick some class of functions $f(x)$ (decision trees, linear functions, etc.)
  - 评价（Evaluation）：评价函数（目标函数、打分函数）来判断优劣
    - Pick some loss function, measuring how we would like $f(x)$ to perform on test data.
  - 优化（Optimization）：需要一个搜索方法，能够在假设空间中找到评价函数得分最高的函数
    - Fit $f(x)$ so that it has a good average loss on training data.

为什么机器学习是可能的

- 为什么机器学习是可能的？
  - 什么样的问题能够被有效率的学习 (What can be learned efficiently)？
  - 什么样的问题天生无法有效地被学习 (What is inherently hard to learn)？
  - 成功的学习需要多少样本 (How many examples are needed to learn successfully)？
  - 学习有没有一个综合性的模型指导 (Is there a general model of learning)？
- PAC 学习框架 (Probably Approximately Correct learning framework) 可以解释上述问题

Table of Contents

Leslie Valiant （莱斯利 · 瓦朗特）

Leslie Valiant

PAC learning was invented by Leslie Valiant in 1984, and it birthed a new subfield of computer science called computational learning theory and won Valiant some of computer science's highest awards.

基本概念

为了叙述 PAC 模型，先引进一些记号：

- 输入空间 (input space) 记作 $\mathcal{X}$ ，表示所有样本 (examples) 或实例 (instances) 的所有可能取值的集合
- 所有可能的标签 (labels) 或目标值 (target values) 的集合记作 $\mathcal{Y}$
- 一个概念 (concept) 记作 c : $\mathcal{X} \to \mathcal{Y}$，是从 $\mathcal{X}$ 到 $\mathcal{Y}$ 的一个映射（规则）
- 一个概念类 (concept class) 是一些我们希望学习到的概念的集合，记作 $\mathcal{C}$
- 我们假设，所有的样本都是独立同分布的 (i.i.d.)，满足一个固定的但是未知的分布 $\mathcal{D}$ (fixed but unknown distribution)

监督学习问题

我们可以将学习问题进行如下叙述：

- 首先，给定学习器 (learner) 一个由可能概念构成的固定集合，即假说集 (hypothesis set)，记作 $\mathcal{H}$，这个假说集不一定必须和 $\mathcal{C}$ 有重合
- 接着，给予学习器样本 $S = (x_1, \ldots, x_m)$ 与对应的标签 $(c(x_1), \ldots, c(x_m))$，其中每个样本都是根据分布 $\mathcal{D}$ 来 i.i.d. 得到的，c 是需要学到的目标概念，属于概念类 $\mathcal{C}$
- 学习器需要根据带有标签的数据集 S，从假说集 $\mathcal{H}$ 中选择一个假说 h_s，该 h_s 有着相对于目标概念 c 很小的泛化误差 (generalization error)。

两种误差

- 泛化误差 (generalization error)：给定 $h \in \mathcal{H}$, c$\in \mathcal{C}$，分布 $\mathcal{D}$，h 的泛化误差为

$$R(h) = P_{x \sim \mathcal{D}}[h(x) \neq c(x)]$$

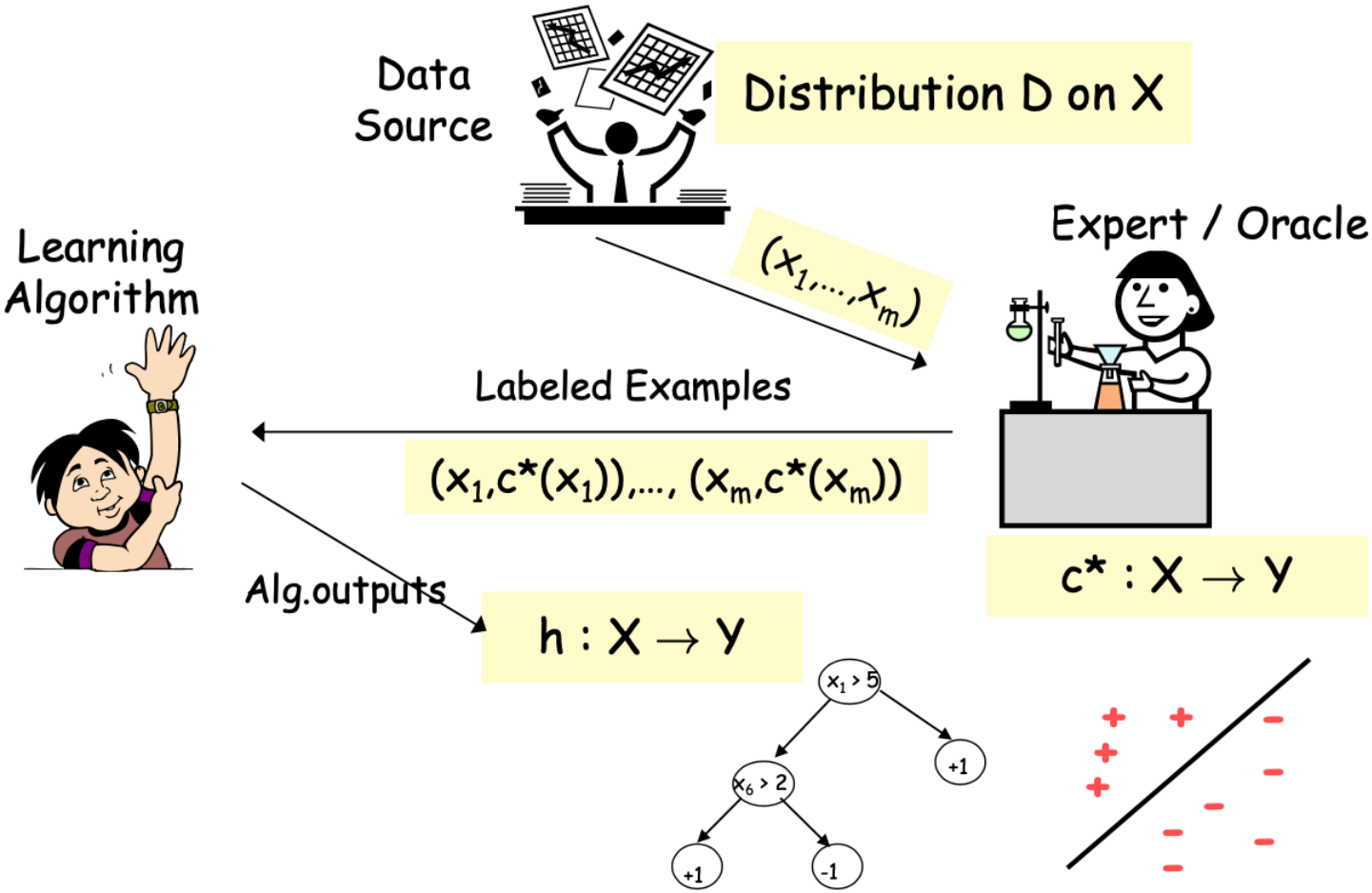
- 经验误差 (empirical error)：给定样本集合 $S = (x_1, \ldots, x_m)$

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} [h(x_i) \neq c(x_i)]$$

- 比较经验误差与泛化误差：
  - 经验误差是 h 在训练数据 S 上的平均误差
  - 泛化误差是 h 在分布 $\mathcal{D}$ 上的期望误差
  - 两者存在联系，因为 S 是根据 $\mathcal{D}$ 独立同分布产生的

$$E_{S \sim \mathcal{D}^m}[\hat{R}_S(h)] = R(h)$$

PAC models for Supervised Learning



PAC Model

1. Generate instances from *unknown* distribution p^*

$$\mathbf{x}^{(i)} \sim p^*(\mathbf{x}),\ \forall i \tag{1}$$

2. Oracle labels each instance with *unknown* function c^*

$$y^{(i)} = c^*(\mathbf{x}^{(i)}),\ \forall i \tag{2}$$

3. Learning algorithm chooses hypothesis $h \in \mathcal{H}$ with low(est) training error, $\hat{R}(h)$

$$\hat{h} = \operatorname*{argmin}_h \hat{R}(h) \tag{3}$$

4. Goal: Choose an h with low generalization error $R(h)$

Two Types of Error

True Error (aka. **expected risk**)

$$R(h) = P_{\mathbf{x} \sim p^*(\mathbf{x})}(c^*(\mathbf{x}) \neq h(\mathbf{x}))$$

This quantity is always **unknown**

Train Error (aka. **empirical risk**)

$$\begin{aligned}\hat{R}(h) &= P_{\mathbf{x} \sim \mathcal{S}}(c^*(\mathbf{x}) \neq h(\mathbf{x})) \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(c^*(\mathbf{x}^{(i)}) \neq h(\mathbf{x}^{(i)})) \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y^{(i)} \neq h(\mathbf{x}^{(i)}))\end{aligned}$$

We can **measure** this on the training data

where $\mathcal{S} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})\}_{i=1}^{N}$ is the training data set, and $\mathbf{x} \sim \mathcal{S}$ denotes that $\mathbf{x}$ is sampled from the empirical distribution.

Three Hypotheses of Interest

The **true function** c^* is the one we are trying to learn and that labeled the training data:

$$y^{(i)} = c^*(\mathbf{x}^{(i)}),\ \forall i \tag{1}$$

The **expected risk minimizer** has lowest true error:

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} R(h) \tag{2}$$

The **empirical risk minimizer** has lowest training error:

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}(h) \tag{3}$$

PAC 学习理论

- PAC 学习理论考虑，能否从假设空间 $\mathcal{H}$ 中学习一个好的假设 h
- “好的假设” 需要满足两个条件 (PAC 辨识条件)：
  - 近似正确 (Approximately Correct)：泛化误差 $R(h)$ 足够小
    - $R(h)$ 越小越好，最好泛化误差能能于 0，但一般是不可能的。那我们就把 $R(h)$ 限定在一个很小的数 ϵ 之内，即只要假设 h 满足 $R(h) \leq \epsilon$，我们就认为 h 是近似正确的
  - 可能正确 (Probably Correct)：h 在很大概率上近似正确
    - 不指望选择的假设 h 百分之百是近似正确的，即 $R(h) \leq \epsilon$，只要很可能是近似正确的就可以，即我们给定一个值 δ，假设 h 满足 $P(R(h) \leq \epsilon) \geq 1 - \delta$
- 同时学习所需的样本数量不能太大，样本数量是关于 $1/\epsilon$, $1/\delta$, 样本大小，$size(c)$ 的多项式函数

PAC Learnable

定义

A concept class $\mathcal{C}$ is said to be *PAC-learnable* if there exists an algorithm $\mathcal{A}$ and a polynomial function function $poly(\cdot, \cdot, \cdot, \cdot)$ such that for any $\epsilon > 0$ and $\delta > 0$, for all distribution $\mathcal{D}$ on $\mathcal{X}$ (containing instances of length n) and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m \geq poly(1/\epsilon, 1/\delta, n, size(c))$:

$$P_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

If $\mathcal{A}$ further runs in $poly(1/\epsilon, 1/\delta, n, size(c))$, then $\mathcal{C}$ is said to be *efficiently PAC-learable*. When such an algorithm $\mathcal{A}$ exists, it is called a *PAC-learning algorithm* for $\mathcal{C}$.

Guarantees for finite hypothesis sets – consistent case

- 一致情形 (consistent)： $c \in \mathcal{H}$
  - target concept c 在 learner 的（有限）假设集合（hypothesis set) $\mathcal{H}$ 中

定理 (Learning bound – finite $\mathcal{H}$, consistent case)

Let $\mathcal{H}$ be a finite set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in \mathcal{H}$ and i.i.d. sample S returns a consistent hypothesis h_S: $\hat{R}_S(h_S) = 0$. Then, for any $\epsilon, \delta > 0$, the inequality $P_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$ holds if

$$m \geq \frac{1}{\epsilon}\left(\log |\mathcal{H}| + \log \frac{1}{\delta}\right).$$

This sample complexity result admits the following equivalent statement as a generalization bound: for any $\epsilon, \delta > 0$, with probability at least $1 - \delta$,

$$R(h_S) \leq \frac{1}{m}\left(\log |\mathcal{H}| + \log \frac{1}{\delta}\right).$$

霍夫丁不等式（Hoeffding's Inequality）

定理 (Hoeffding's Inequality)

Let $X_1, \ldots, X_m$ be i.i.d. random variable in [0, 1], for any $\epsilon > 0$

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E(X_i)\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}.$$

- 该定理给出了位于区间 [0,1] 的两两随机变量其期望与均值之间满足的关系，在任意分布 $\mathcal{D}$ 下
- 由泛化误差 $R(h)$ 与经验误差 $\hat{R}_S(h)$ 的定义易知 $E(\hat{R}_S(h)) = R(h)$，由此得到

$$P_{S\sim\mathcal{D}^m}\left[\left|\hat{R}_S(h) - R(h)\right| \geq \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

- 令 $\delta = 2e^{-2m\epsilon^2}$，则 Fix a hypothesis h: $\mathcal{X} \to \{0, 1\}$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

Guarantees for finite hypothesis sets – inconsistent case

- **不一致情形** (inconsistent)：$c \notin \mathcal{H}$，**更通常的情况**

定理 (Learning bound – finite $\mathcal{H}$, inconsistent case)

Let $\mathcal{H}$ be a finite hypothesis set. Then, for any $\delta > 0$ with probability at least $1 - \delta$, the following inequality holds:

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2m}}.$$

Guarantees for finite hypothesis sets – inconsistent case

- Thus, for a finite hypothesis set $\mathcal{H}$,

$$R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\log_2 |\mathcal{H}|}{m}}\right)$$

- For a fixed $|\mathcal{H}|$, to attain the same guarantee as in the consistent case, a quadratically larger labeled sample is needed
- This can also be viewed as an instance of the so-called Occam's Razor principle
  - Plurality should not be posited without necessity, also rephrased as, the simplest explanation is best.

Summary: Learning

- Machine Learning $\approx$ Looking for a Function
- Learning = Representation + Evaluation + Optimization
  - Representation: hypothesis set
  - Evaluation: loss function
  - Optimization: search approach
- PAC Learning: **机器学习为什么是可能的?**

$$P_{S\sim\mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1-\delta$$

  - **近似正确** (Approximately Correct): **泛化误差** $R(h)$ **足够小**
  - **可能正确** (Probably Correct): h **在很大概率上近似正确**
- PAC criterion: the learner produces a high accuracy learner with high probability

$$P_{S\sim\mathcal{D}^m}\left[\left|R(h)-\hat{R}_S(h)\right| \leq \epsilon\right] \geq 1-\delta$$