learning—The road to intelligence

学习
—
通向智能之路

尹慧琳

同济大学电子与信息工程学院控制科学与工程系

同济大学中德智能中心

学习板块安排

第9讲:　学习概述和机器学习

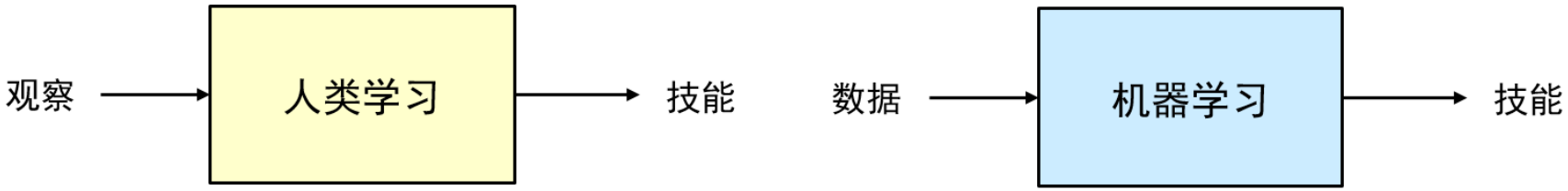
第10讲: 神经网络、深度学习和卷积神经网络CNN

第11讲: CV：分类和目标识别网络

第12讲: NLP：RNN/LSTM　Transformer

什么是学习?

- 学习，是指通过阅读、听讲、思考、研究、实践等途径获得知识和技能的过程。(百度百科)
- 学习，是通过外界教授或从自身经验提高能力的过程。（wiki）
- 对机器而言，学习是达成智能（智慧能力）的途径。
- 通过对世界的观察和探索，能够改进执行未来任务时的性能。



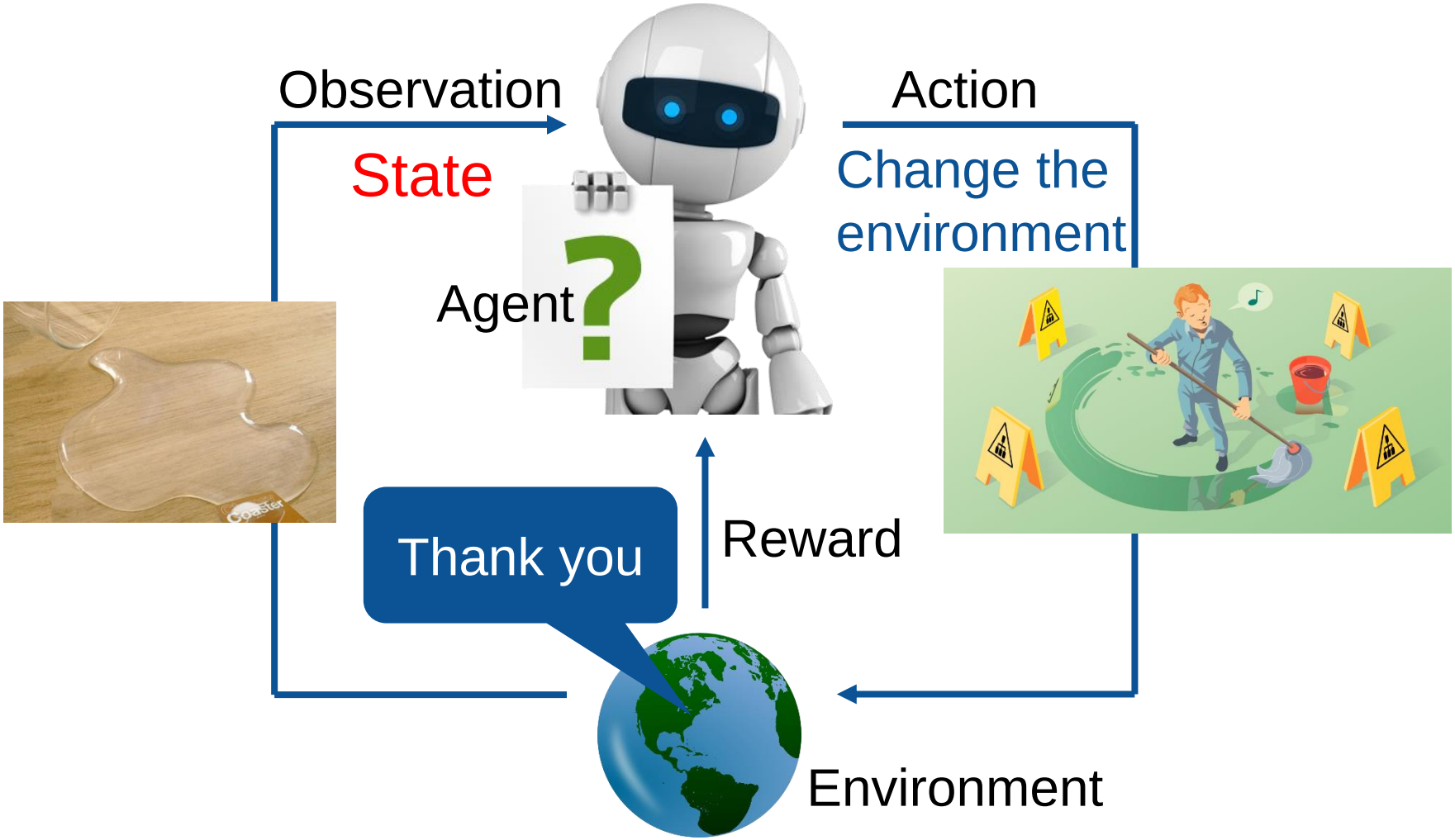
机器学习

- 狭义：从数据中学规律学模型
- 广义：以实现系统的智能能力为目标，不局限于基于数据的学习
  - ✓ 从数据出发的基于样本的学习：统计机器学习 Statistical machine learning，深度学习 Deep learning
  - ✓ 利用人机交互的增强学习，强化学习 Reinforcement Learning, RL
  - ✓ 左右互博向对手学习的对抗学习 Adversarial Learning：生成对抗网络(Generative Adversarial Network, GAN) 、Adversarial example attack and defence

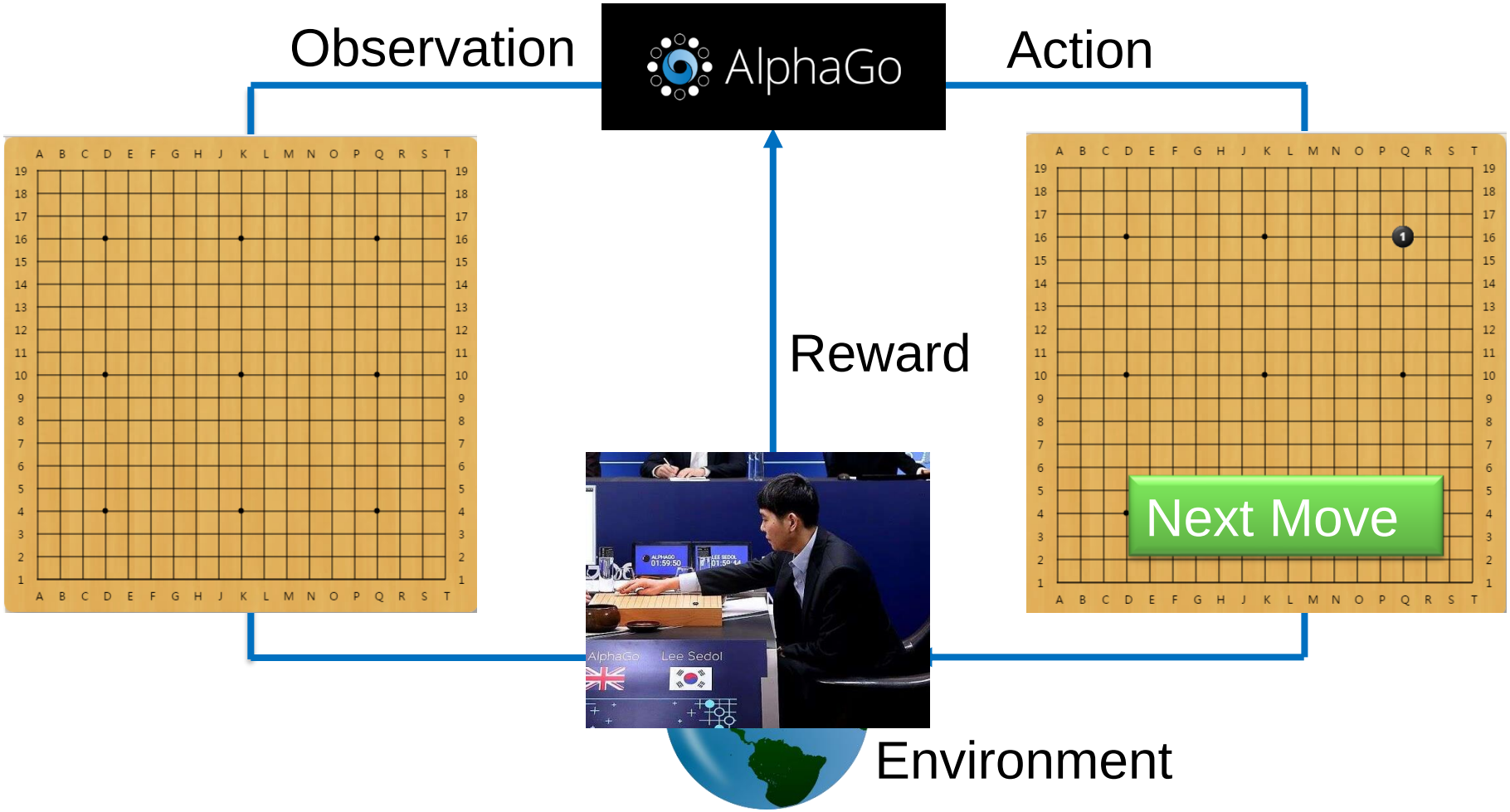
增强学习场景举例


Observation
Action
State
Change the environment
Agent
?
Don’t do that
Reward
Environment

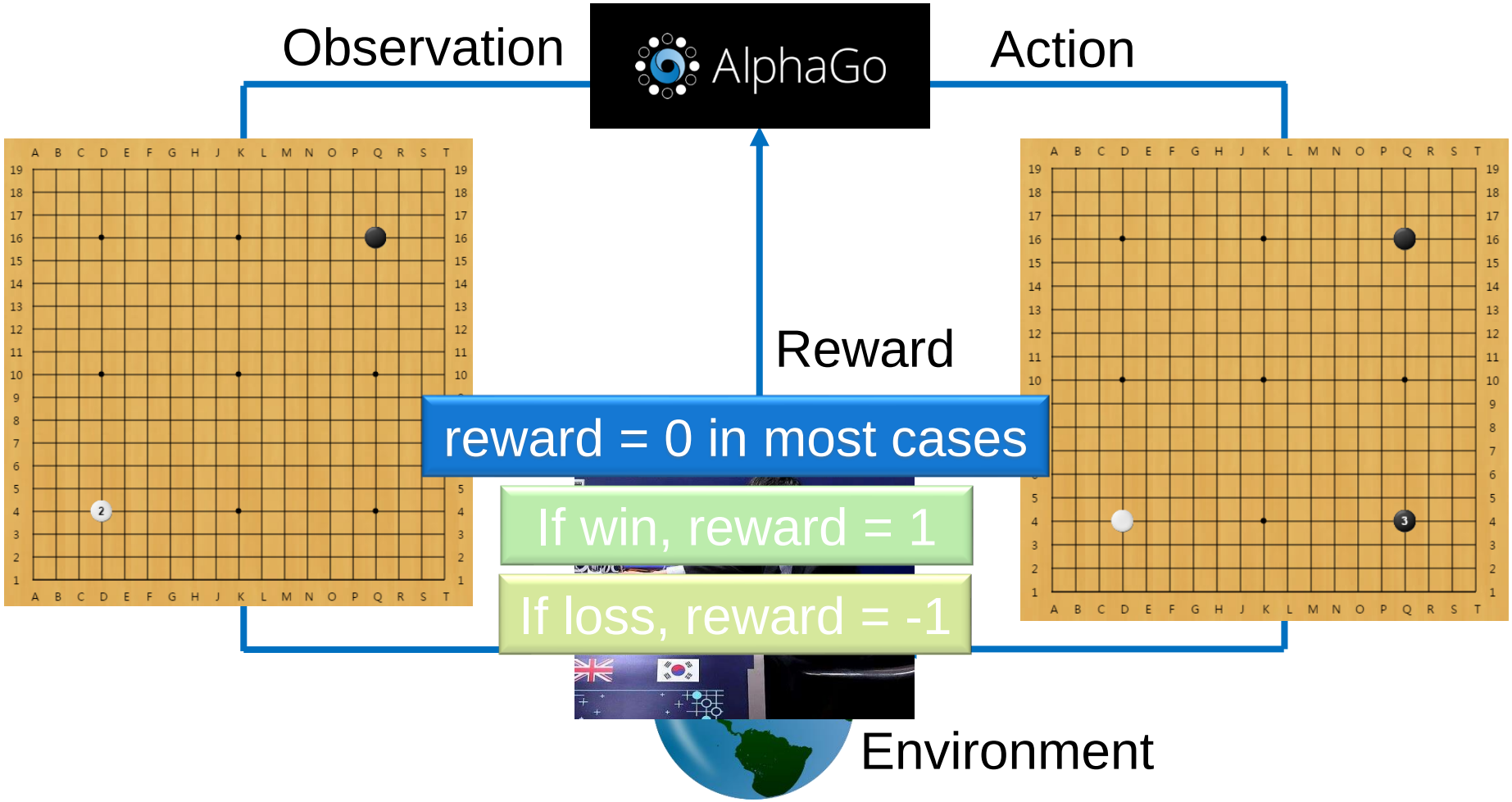
增强学习场景举例


Observation
Action
State
Change the environment
Agent
?
Thank you
Reward
Environment

ALPHAGO 增强学习

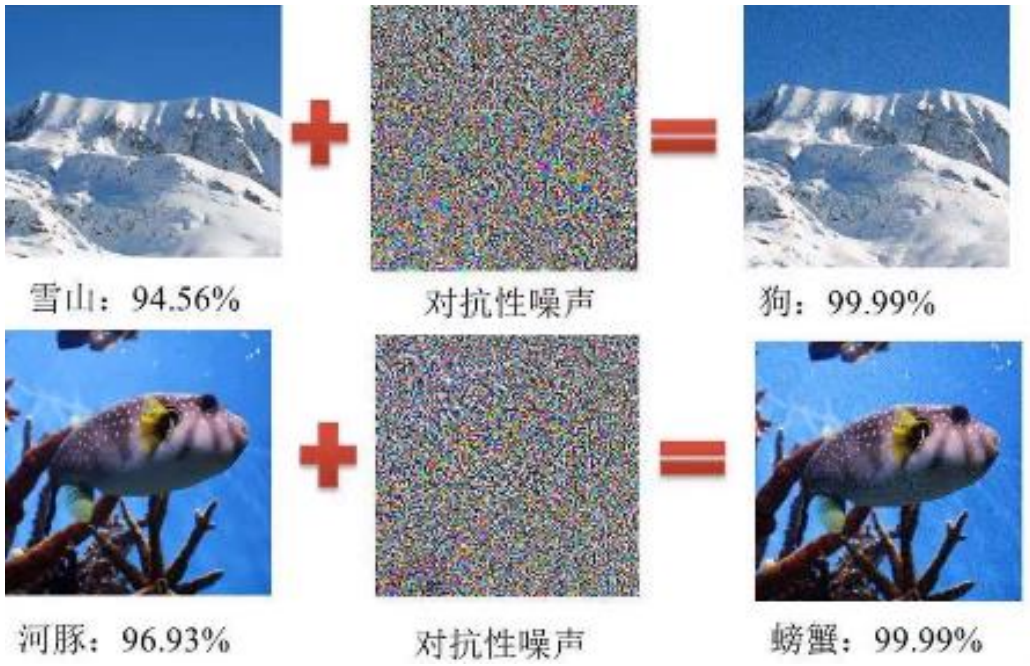


ALPHAGO 增强学习



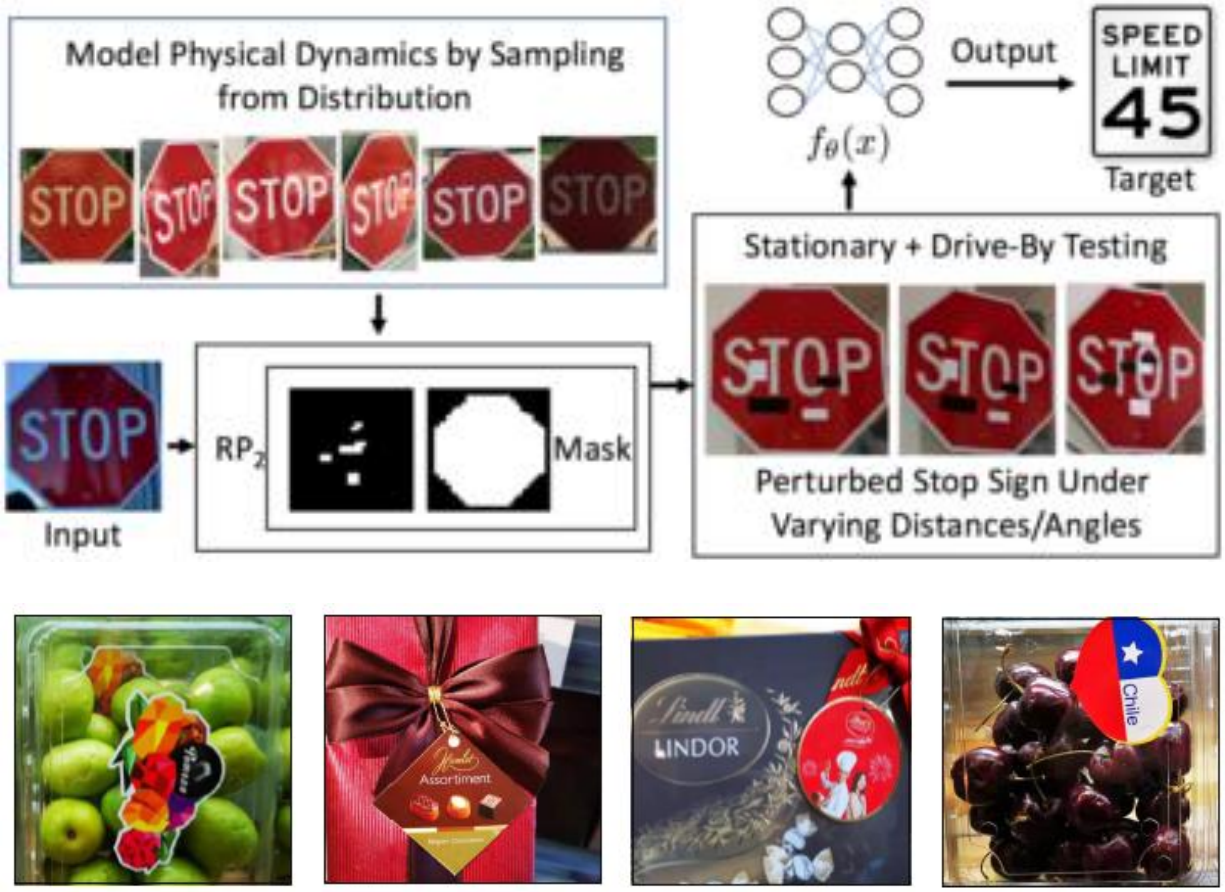
对抗样本攻击举例

对抗样本（Adversarial examples）是指在数据集中通过故意添加细微的干扰所形成的输入样本，会导致模型以高置信度给出一个错误的输出。



对抗样本攻击举例

物理攻击 physical-world attack



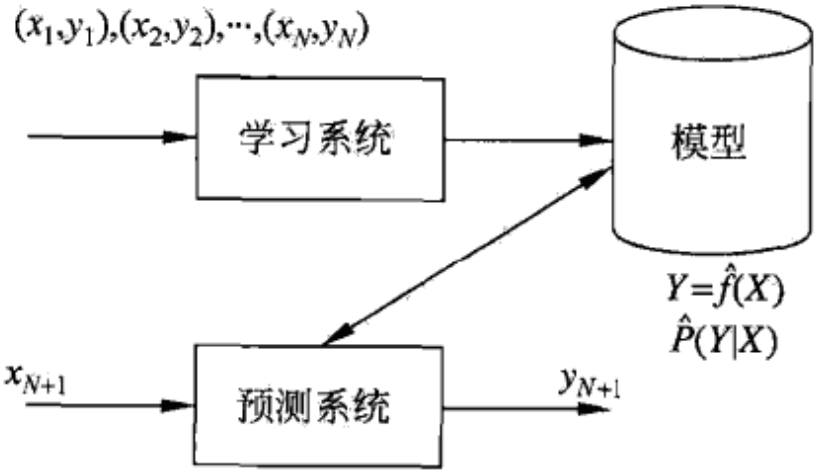
- 统计机器学习是基于数据构建统计模型从而对数据进行预测与分析
- 对象：数据 data (->特征)
- 前提基本假设：同类数据具有一定的统计规律性
- 建模
- 预测或判断
- 数据包括：数字，文字，图像，视频，音频等

学习的分类

- 监督学习(supervised learning)
- 非监督学习(unsupervised learning)
- 半监督学习(semi-supervised learning)
- 强化学习(reinforcement learning)

标注样本 labeled examples/samples

监督学习问题的形式化



首先给定一个训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

建模过程中，学习系统利用训练集，通过学习得到一个模型，

决策函数 $Y=f(X)$ 或条件概率分布 $P(Y|X)$

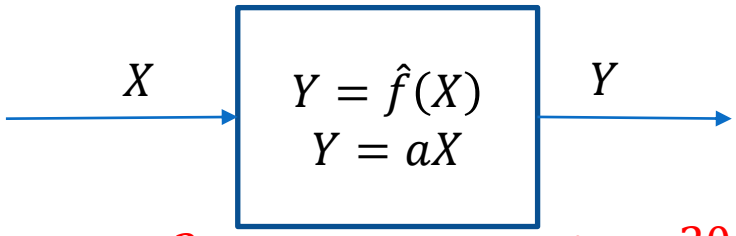
预测过程中，预测系统对于给定的测试样本，由模型得出相应的输出，

$y_{N+1}=\hat{f}(x_{N+1})$ 或 $y_{N+1}=argmax\hat{P}((y_{N+1}|x_{N+1}))$

举例: {(1,10),(3,30),(4,40)}, N=3, y=10x

$x_1 = 1$
$x_2 = 3$
$x_3 = 4$

$y_1 = 10$
$y_2 = 30$
$y_3 = 40$



$x_4 = 2$

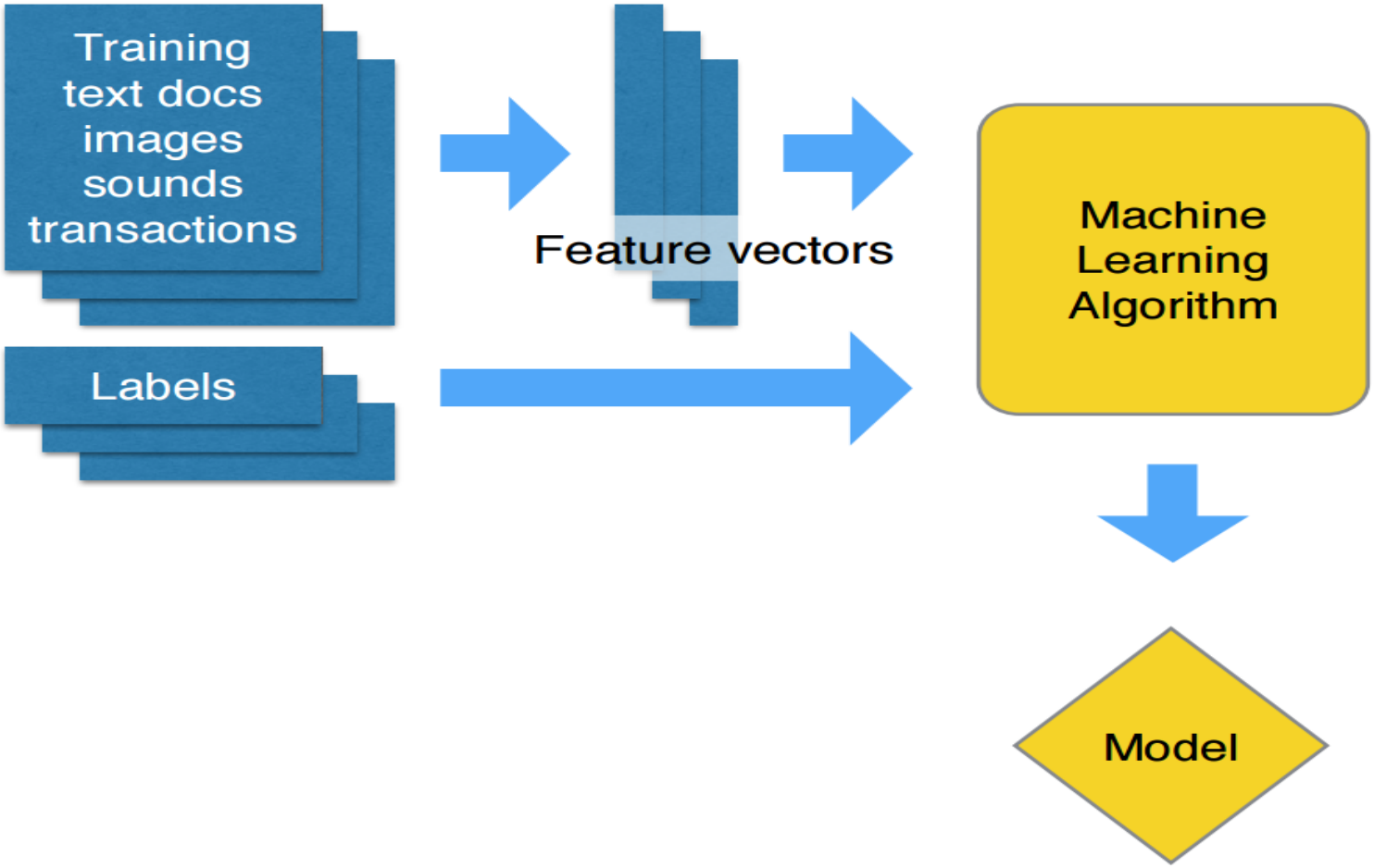
$y_4 = 20$

$k = 0: a = 0,\ loss = (10 - 0) + (30 - 0) + (40 - 0)$
$k = 1: a = 3,\ loss = (10 - 3) + (30 - 9) + (40 - 12)$
$k = 2: a = 5$
$k = 3: a = 6.5$

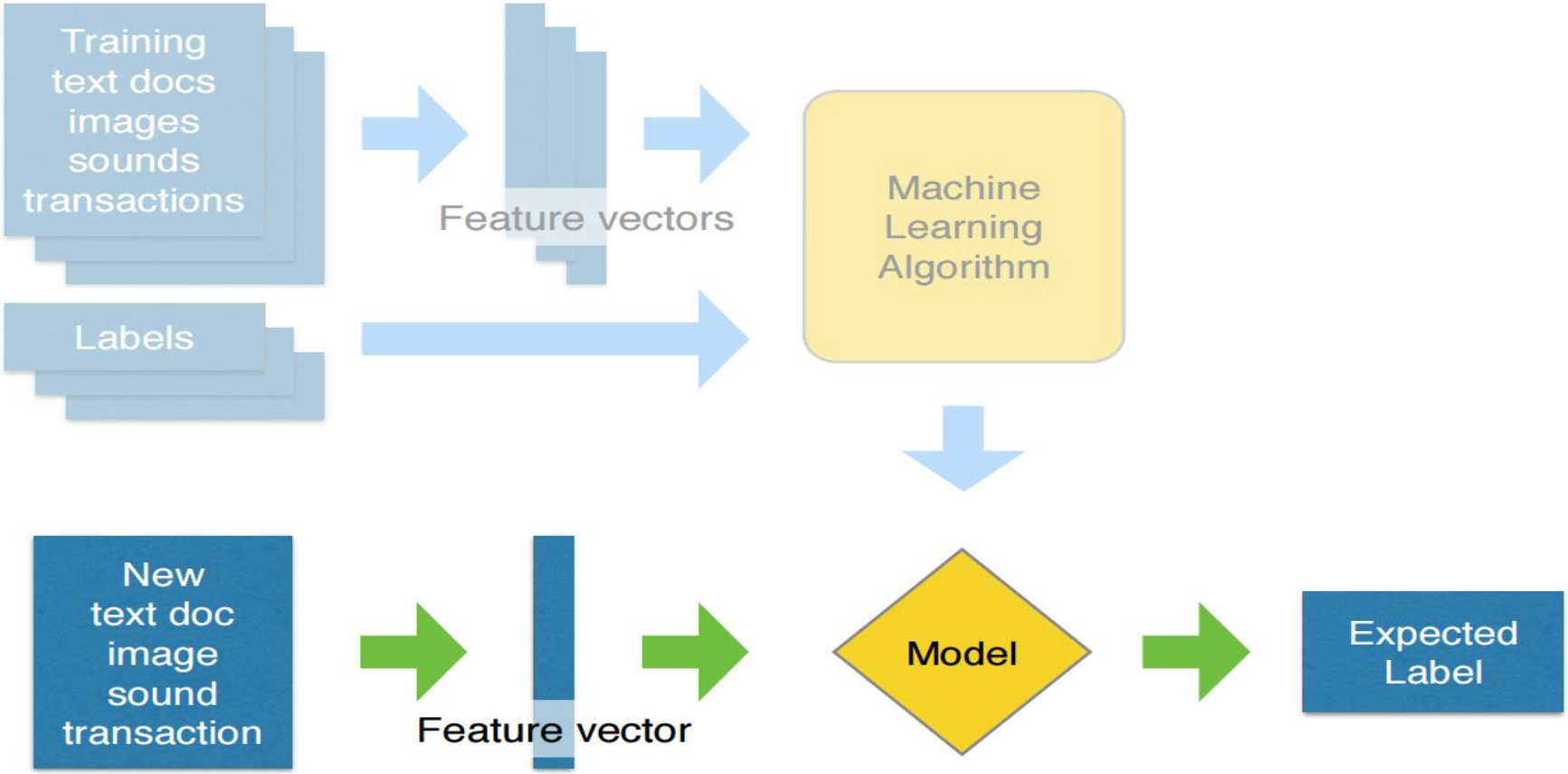
$k = 6: a = 9.8$

$k = 9: a = 10,\ loss = 0$

监督学习建模


Training
text docs
images
sounds
transactions
Labels
Feature vectors
Machine
Learning
Algorithm
Model

监督学习预测


Training
text docs
images
sounds
transactions
Labels
Feature vectors
Machine
Learning
Algorithm
New
text doc
image
sound
transaction
Feature vector
Model
Expected
Label

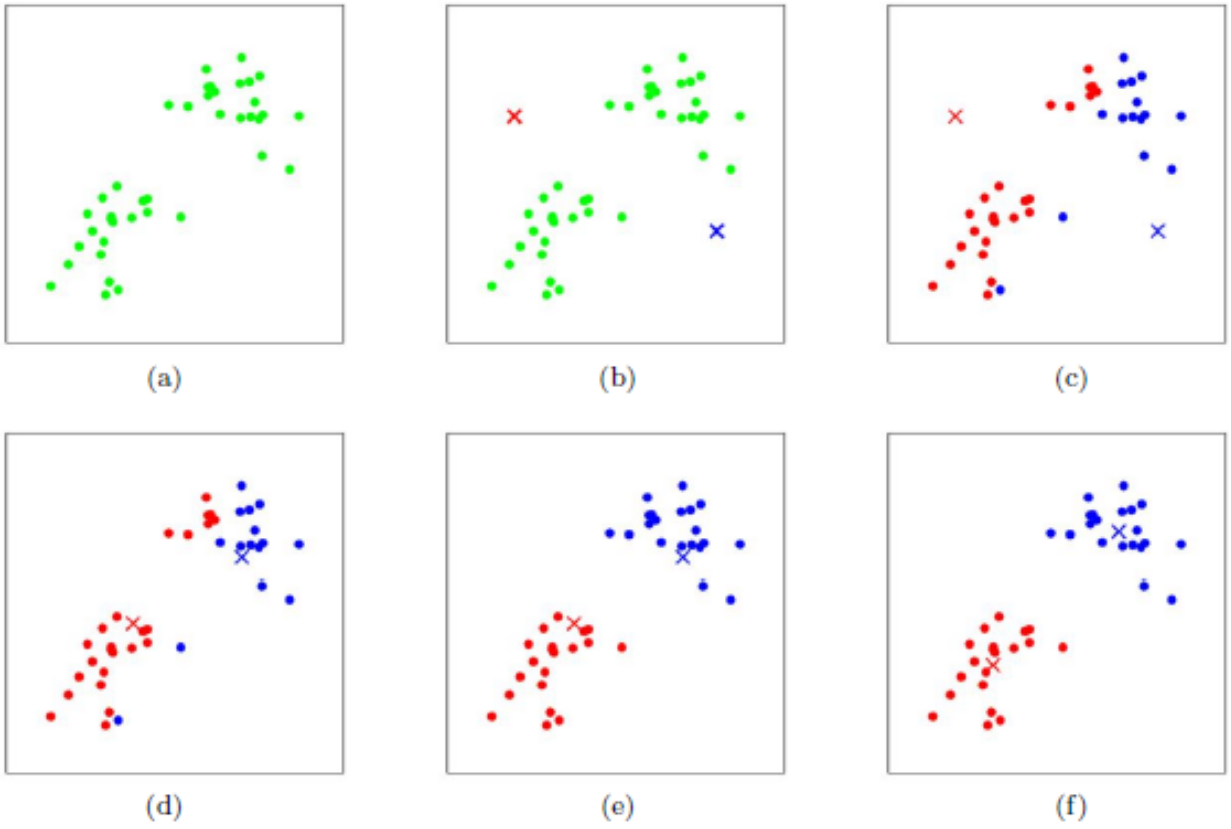
非监督学习

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

Yann LeCun：于NIPS (NeurIPS) 2016： 如果说智能是一块蛋糕，无监督学习就是这块蛋糕，监督学习只是蛋糕上的糖霜，而强化学习仅是蛋糕上的樱桃。

非监督学习-举例K-MEANS聚类

对于给定的样本集，按照样本之间的距离大小，将样本集划分为K个簇；



统计机器学习三要素：模型

模型+策略+算法

假设空间（hypothesis space）：模型的集合，参数向量决定的决策函数族或条件概率分布族

$$\mathcal{F} = \{f | Y = f_\theta(X), \theta \in \boldsymbol{R}^n\}$$

$$\mathcal{F} = \{P | P_\theta(Y|X), \theta \in \boldsymbol{R}^n\}$$

e.g. $Y = a_0 + a_1 X \qquad \theta = (a_0, a_1)$

$$Y \sim N(a_0 + a_1 X, \sigma^2) \qquad \theta = (a_0, a_1)$$

统计机器学习三要素：策略

损失函数（loss function）：预测值与真实标签值的差别，度量预测的好坏

1） 0-1损失函数 （0-1 loss function）

$$L(Y,f(X))=\begin{cases}1, Y\neq f(X)\\ 0, Y=f(X)\end{cases}$$

2）平方损失函数 （quadratic loss function）

$$L(Y,f(X))=(Y-f(X))^2$$

3）绝对损失函数（absolute loss function）

$$L(Y,f(X))=|Y-f(X)|$$

4）对数似然损失函数（log- likelihood loss function）

$$L\big(Y,P(Y|X)\big)=-\log P(Y|X)$$

统计机器学习三要素：策略

经验风险最小化

$$R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i)) \qquad \min_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$$

结构风险最小化

$$R_{srm}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$$

$$\min_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}L\big(y_i,f(x_i)\big)+\lambda J(f)$$

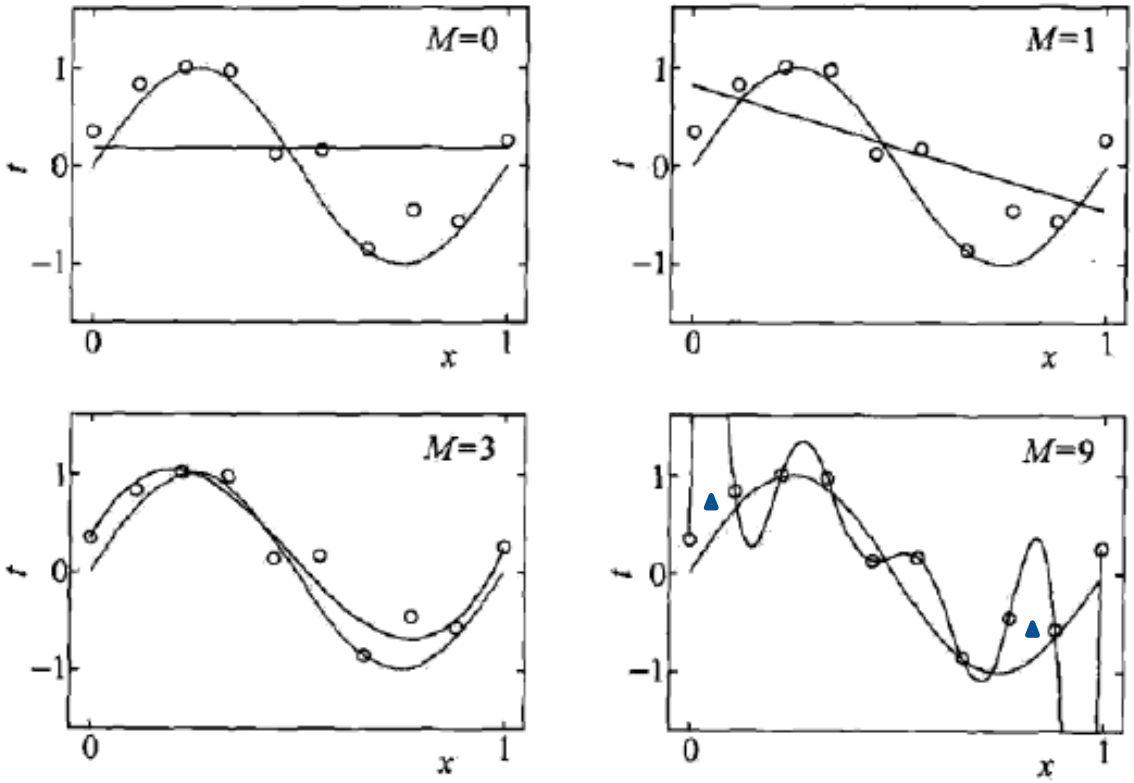
统计机器学习三要素：算法

- 指学习模型的具体计算方法，用什么计算方法求解最优模型
- 统计学习问题归结为最优化问题
- 显式的解析解，数值计算方法求解
- 随机梯度下降法 – 最常用的方法

模型评估与模型选择

- ✓ 训练误差（training error）与测试误差（test error）：平均损失。
- ✓ 泛化能力（generalization ability）：对未知数据的预测能力。
- ✓ 过拟合（over-fitting）：只追求提高对训练数据的预测能力，所得模型复杂度比真模型高。
- ✓ 模型选择目的：避免过拟合，提高模型的预测能力。

过拟合与模型选择举例

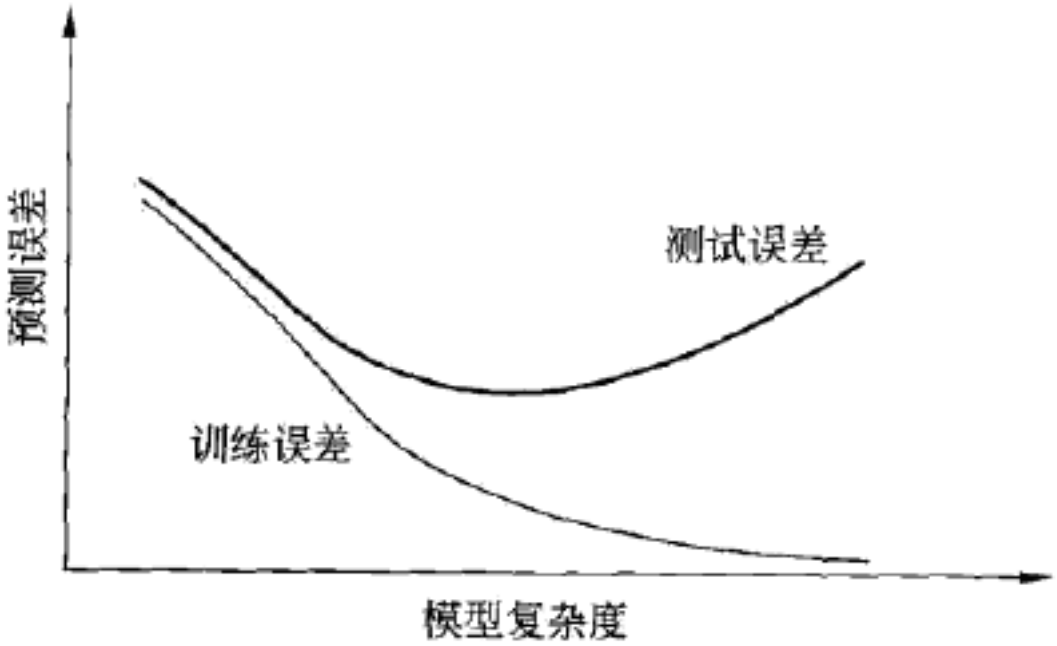


$$f_M(x,w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}(f(x_i,w) - y_i)^2$$

模型评估与模型选择

训练误差和测试误差与模型复杂度关系



正则化 regularization

结构风险最小化： $\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(x_i)\right) + \lambda J(f)$

参数向量的L2范数:

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i;w) - y_i)^2 + \frac{\lambda}{2}\|w\|^2$$

参数向量的L1范数:

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i;w) - y_i)^2 + \lambda\|w\|_1$$

范数?

正则化

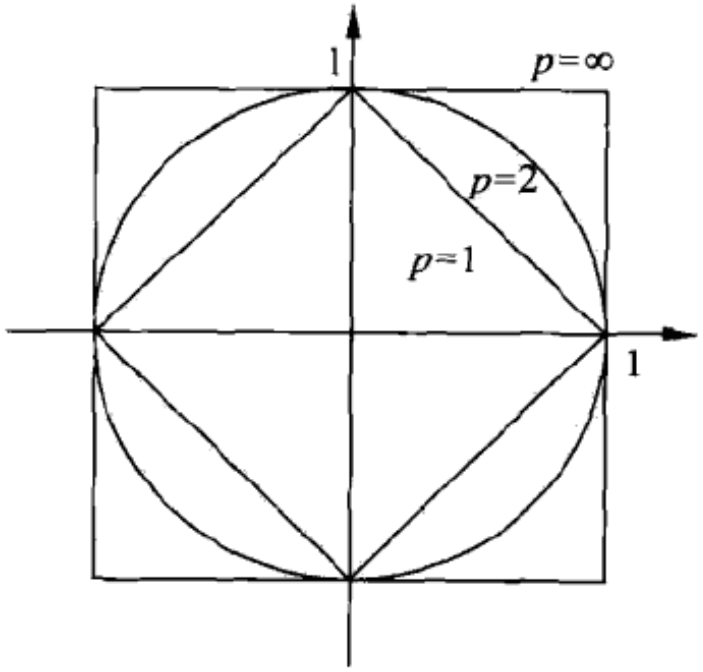
距离度量：欧式距离，Lp距离等

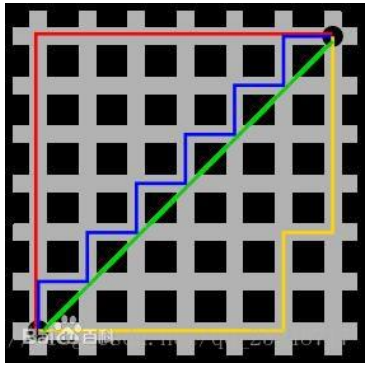
$$L_p(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}}$$

$$L_2(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^2\right)^{\frac{1}{2}}$$

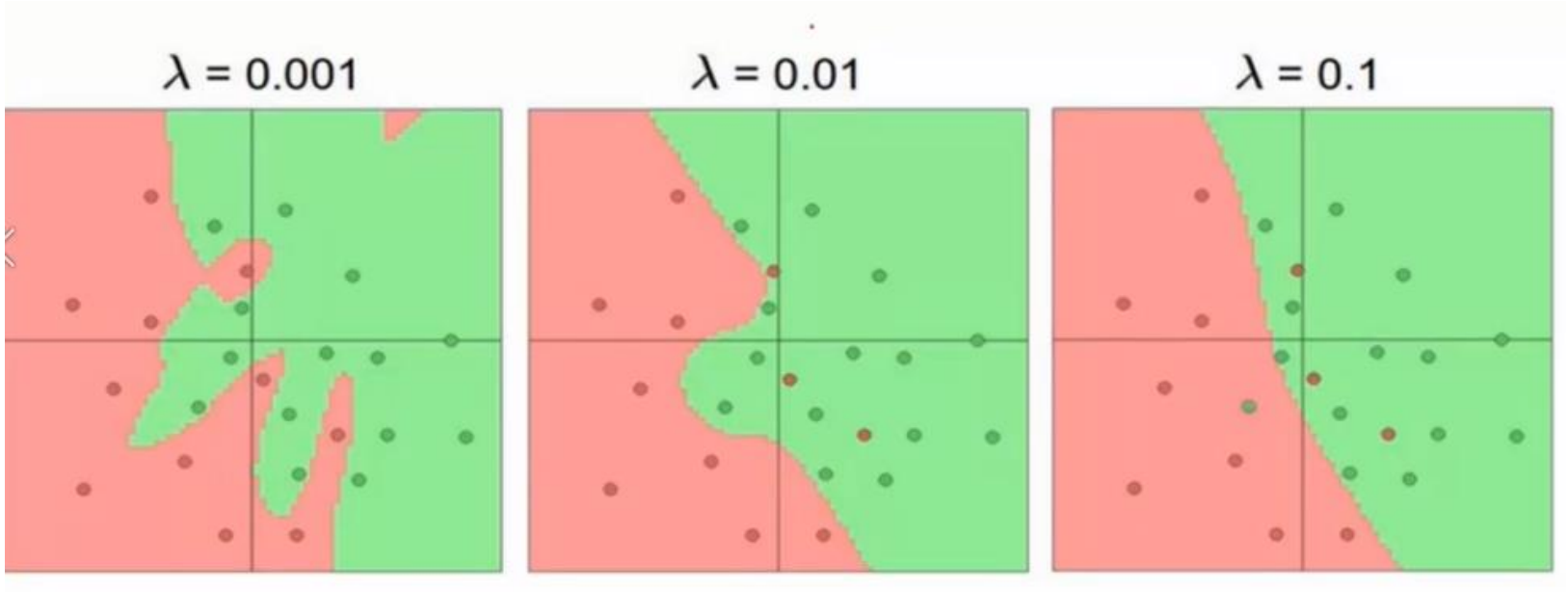
$$L_1(x_i, x_j) = \sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|$$

$$L_\infty(x_i, x_j) = \max_l \left|x_i^{(l)} - x_j^{(l)}\right|$$





正则化



正则化

奥卡姆剃刀（Occam ‘s razor）原理：

“简单有效原理”

“如无必要，勿增实体”

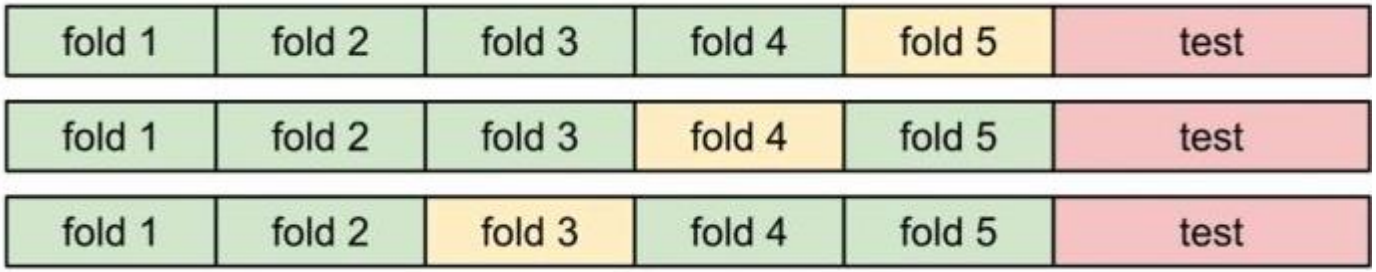
在所有可能选择的模型中，我们应该选择能够很好地解释已知数据并且十分简单的模型。

交叉验证

交叉验证 cross validation

- ✓ 训练集 training set ： 用来训练模型
- ✓ 验证集 validation set：用于模型选择，确定超参
- ✓ 测试集 test set：用于对学习方法的评估

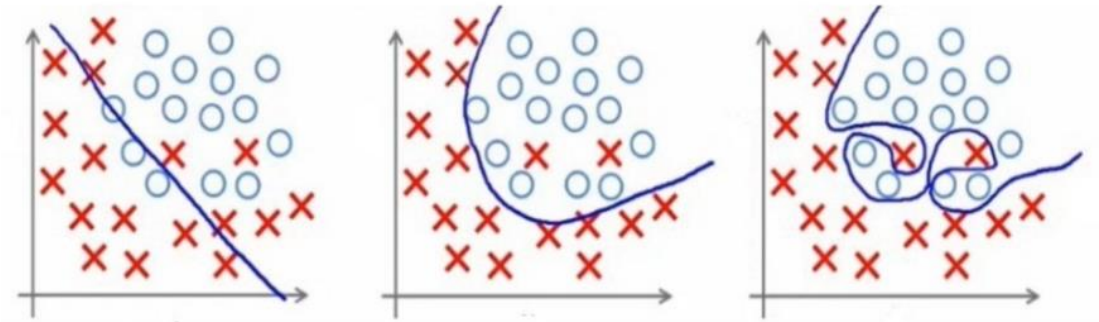
S折交叉验证：



分类问题和回归问题

- 分类 classification：输出变量取有限个离散值
- 回归 regression：输出变量取连续值

  回归问题的学习其实是函数拟合

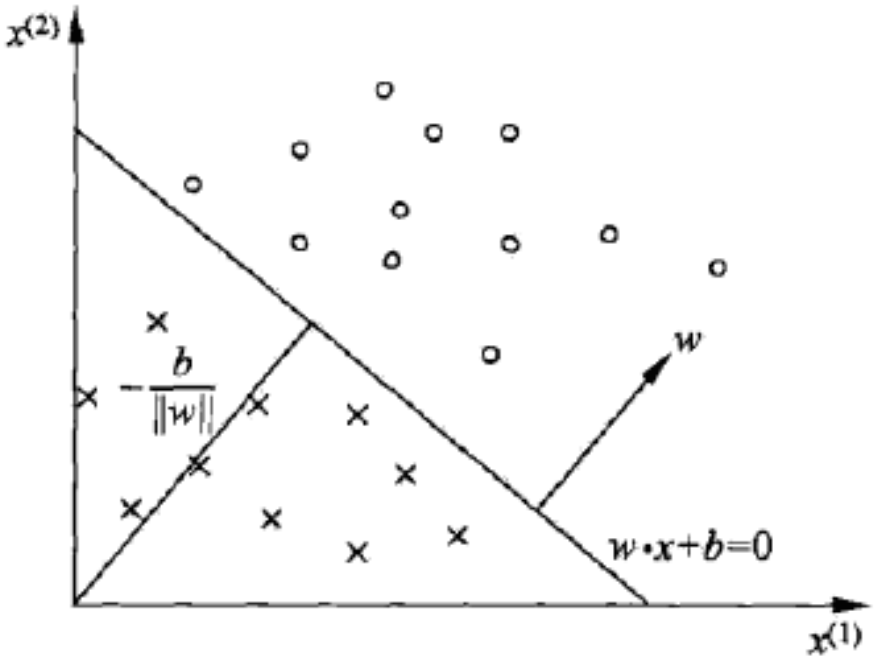


损失函数的形式不同： 分类loss是0-1损失或交叉熵损失，回归loss是平方损失或绝对损失函数。

感知机(PERCEPTRON)模型

- 二分类，线性分类 $f(x) = sign(w \cdot x + b)$ $sign(x) = \begin{cases} +1, x \geq 0 \\ -1, x < 0 \end{cases}$
- 输入是特征向量，输出为类别 +1和-1
- 分离超平面 $w \cdot x + b = 0$ $w \in \boldsymbol{R}^n$

$$w \cdot x = \left(w^{(1)}, w^{(2)}, \cdots, w^{(m)}\right) \cdot \left(x^{(1)}, x^{(2)}, \cdots, x^{(m)}\right)$$
$$= w^{(1)}\, x^{(1)} + w^{(2)}\, x^{(2)} + \cdots + w^{(m)}\, x^{(m)}$$



感知机学习策略

$$d=\frac{|Ax_0+By_0+C|}{\sqrt{A^2+B^2}}$$

- 模型参数： w， b
- 定义损失函数：
  - ✓ 误分类点的个数?
  - ✓ 误分类点到超平面的总距离

$$\begin{cases}\frac{1}{\|w\|}(w\cdot x_i+b), & w\cdot x_i+b\geq 0\\ -\frac{1}{\|w\|}(w\cdot x_i+b), & w\cdot x_i+b<0\end{cases}$$

任一点到超平面的距离： $\frac{1}{\|w\|}|w\cdot x_0+b|$

误分类的点到超平面的距离： $-\frac{1}{\|w\|}y_i(w\cdot x_i+b)$

所有误分类点到超平面的距离： $-\frac{1}{\|w\|}\sum_{x_i\in M}y_i(w\cdot x_i+b)$

损失函数： $L(w,b)=-\sum_{x_i\in M}y_i(w\cdot x_i+b)$

随机梯度下降法 stochastic gradient descent

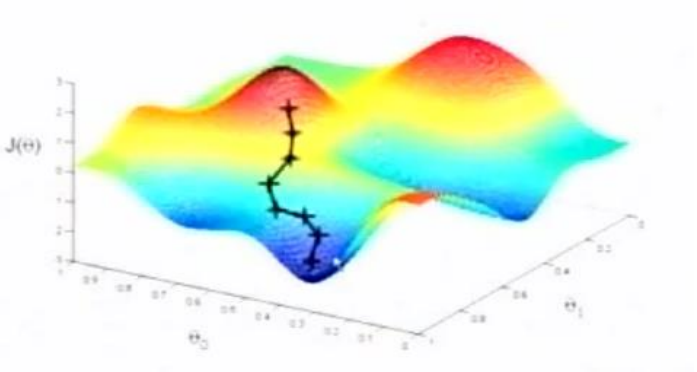
损失函数：$L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$

梯度：

$$\nabla L(w,b) = (\frac{\partial L}{\partial w}, \frac{\partial L}{\partial b})$$

$$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$$

$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$



对参数进行更新：$w \leftarrow w + \eta y_i x_i$

$$b \leftarrow b + \eta y_i$$

感知机学习算法步骤

1) 选取初值 w_0, b_0

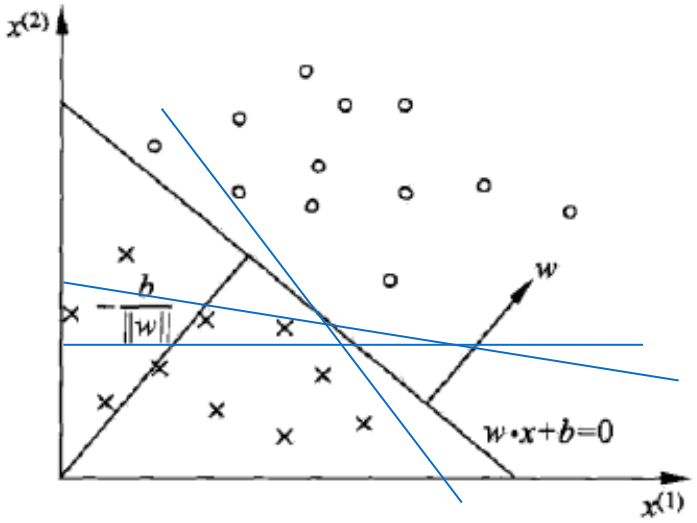
2) 在训练集中选取数据 (x_i, y_i)

3) 如果 $y_i(w \cdot x_i + b) \leq 0$

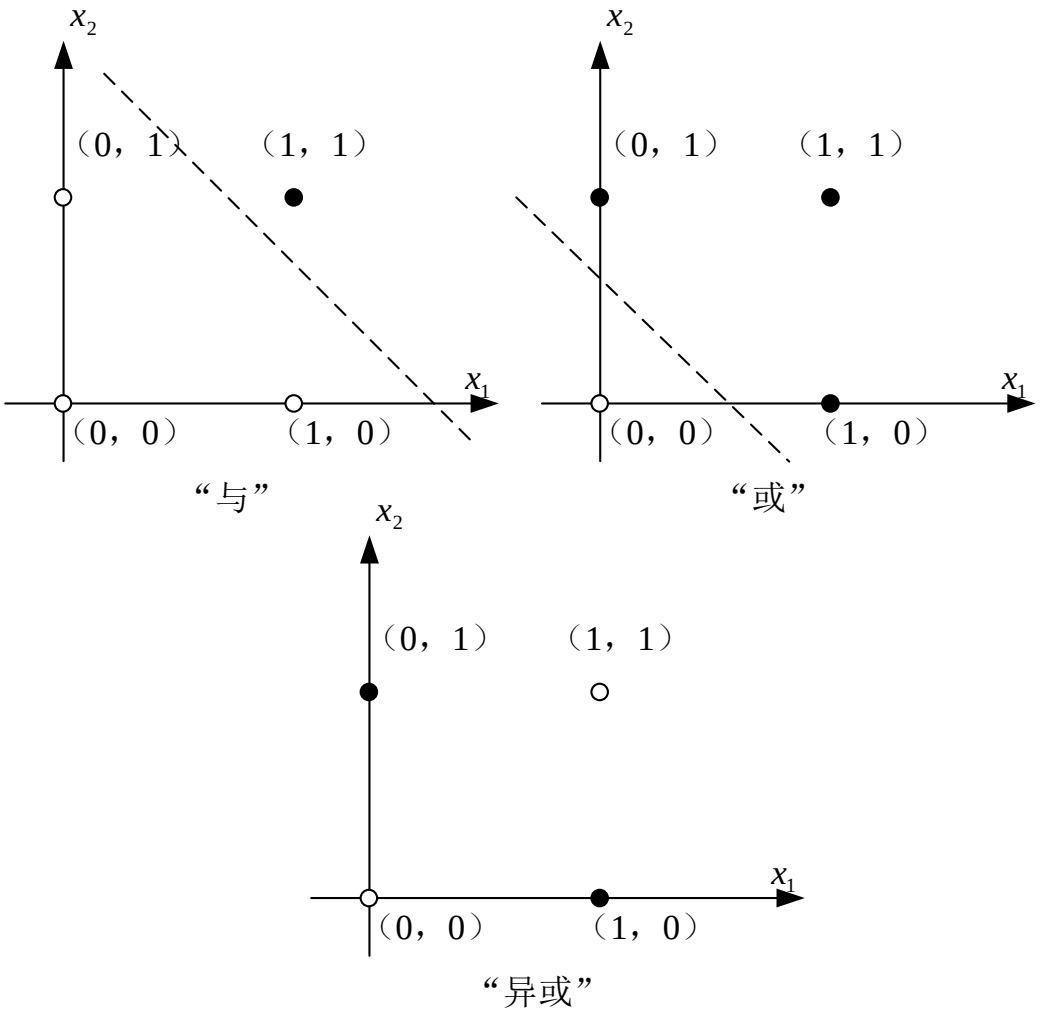
$$w \leftarrow w + \eta y_i x_i$$

$$b \leftarrow b + \eta y_i$$

4) 转至2), 直至训练集中没有误分类点



异或难题



异或难题

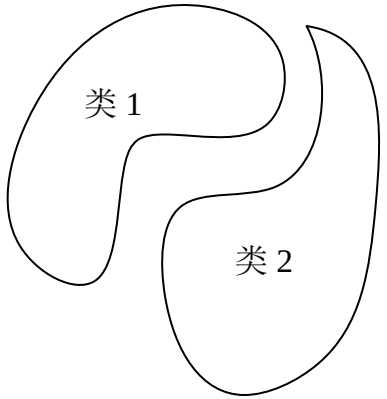
x_1	x_2	“与”	$Y = w_1 \cdot x_1 + w_2 \cdot x_2 - b = 0$	“条件”
0	0	0	$Y = w_1 \cdot 0 + w_2 \cdot 0 - b < 0$	$b > 0$
0	1	0	$Y = w_1 \cdot 0 + w_2 \cdot 1 - b < 0$	$b > w_2$
1	0	0	$Y = w_1 \cdot 1 + w_2 \cdot 0 - b < 0$	$b > w_1$
1	1	1	$Y = w_1 \cdot 1 + w_2 \cdot 1 - b \geq 0$	$b \leq w_1 + w_2$

x_1	x_2	“或”	$Y = w_1 \cdot x_1 + w_2 \cdot x_2 - b = 0$	“条件”
0	0	0	$Y = w_1 \cdot 0 + w_2 \cdot 0 - b < 0$	$b > 0$
0	1	1	$Y = w_1 \cdot 0 + w_2 \cdot 1 - b \geq 0$	$b \leq w_2$
1	0	1	$Y = w_1 \cdot 1 + w_2 \cdot 0 - b \geq 0$	$b \leq w_1$
1	1	1	$Y = w_1 \cdot 1 + w_2 \cdot 1 - b \geq 0$	$b \leq w_1 + w_2$

异或难题

x_1	x_2	“异或”	$Y = w_1 \cdot x_1 + w_2 \cdot x_2 - b = 0$	“条件”
0	0	0	$Y = w_1 \cdot 0 + w_2 \cdot 0 - b < 0$	$b > 0$
0	1	1	$Y = w_1 \cdot 0 + w_2 \cdot 1 - b \geq 0$	$b \leq w_2$
1	0	1	$Y = w_1 \cdot 1 + w_2 \cdot 0 - b \geq 0$	$b \leq w_1$
1	1	0	$Y = w_1 \cdot 1 + w_2 \cdot 1 - b < 0$	$b > w_1 + w_2$

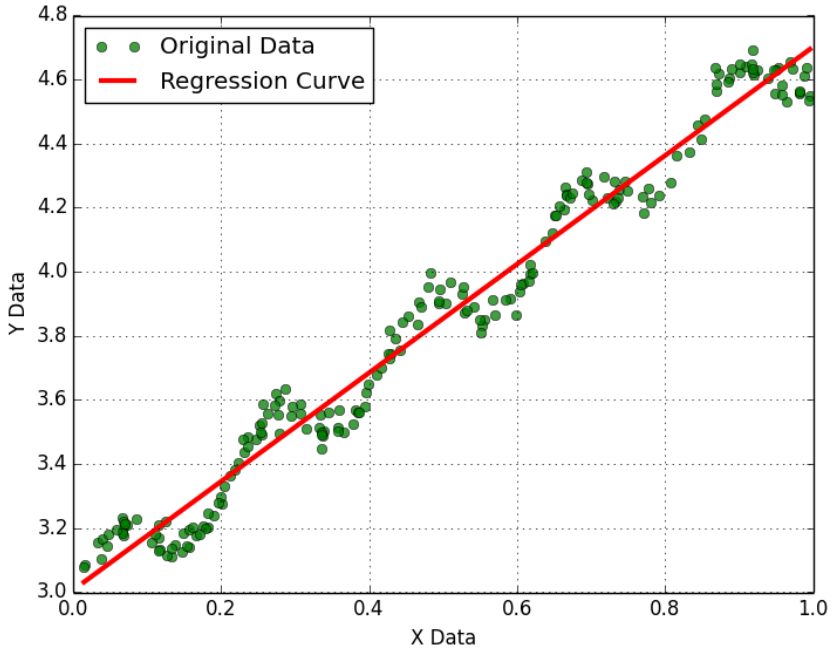
（单层）感知机不能解决线性不可分问题



线性回归 /LINEAR REGRESSION

线性回归是利用回归分析，确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。 “满足多元一次方程”

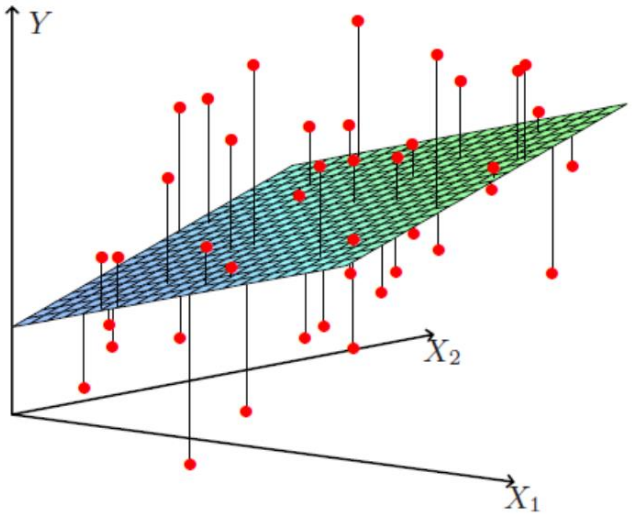
一元线性回归: 单变量线性回归， “拟合一条直线”



$$y = b + w_1 x$$

线性回归

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮



考虑两个变量，多元线性回归

$$y = b + w_1 x_1 + w_2 x_2$$

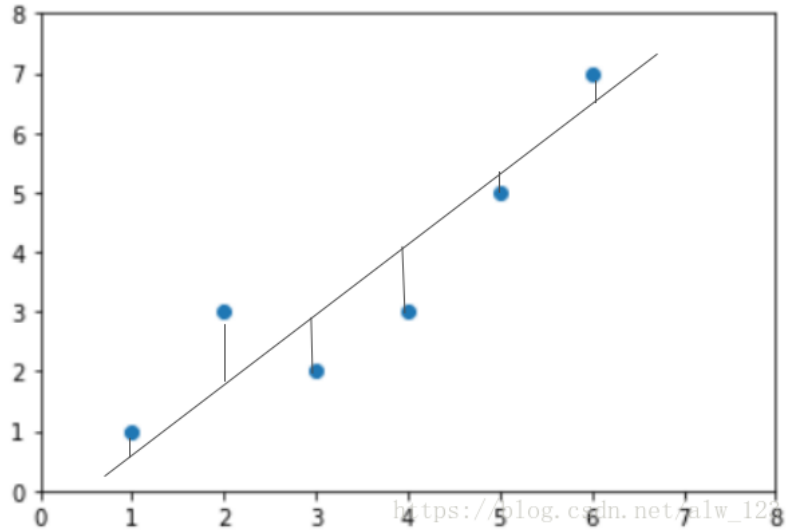
线性回归

损失函数，最小二乘法，梯度下降

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$L(w,b) = \frac{1}{n}\sum_{i=1}^{n}(wx_i + b - y_i)^2$$

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{n}(wx_i + b - y_i)^2$$



$$\frac{\partial L}{\partial w} = 2\left(w\sum_{i=1}^{n}x^2 - \sum_{i=1}^{n}x_i(y_i - b)\right)$$

$$\frac{\partial L}{\partial b} = 2\left(nb - \sum_{i=1}^{n}(y_i - wx_i)\right)$$

最优解

$$w = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i)$$