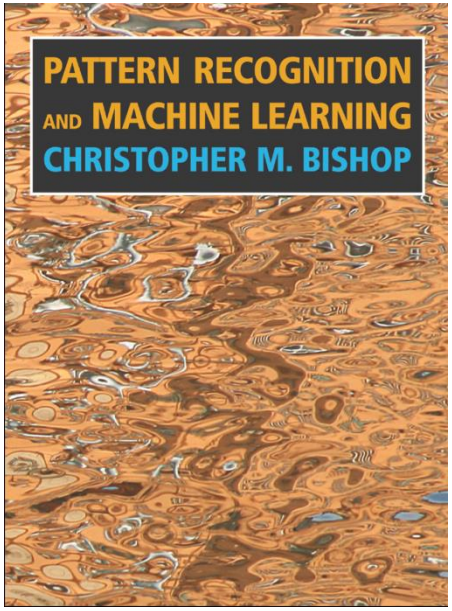
机器学习

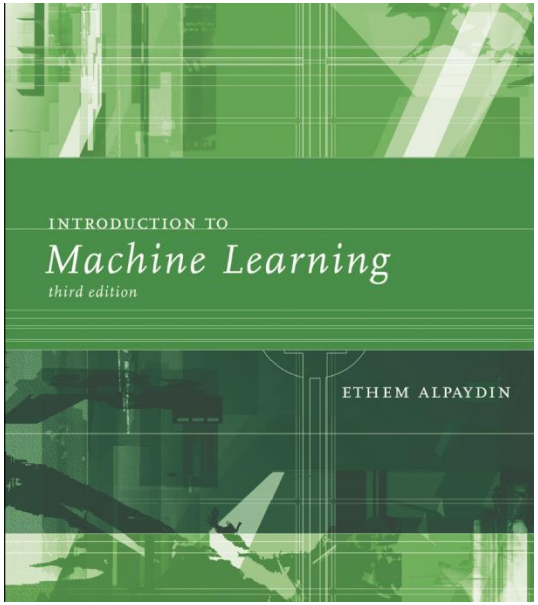
于元隆、朱丹红

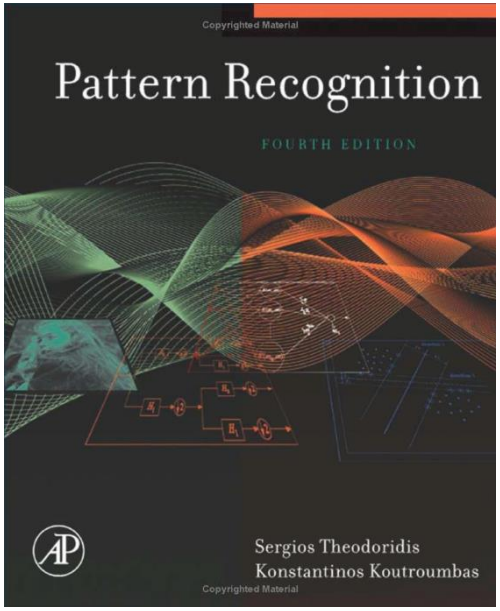
福州大学计算机与大数据学院
Email: yu.yuanlong@fzu.edu.cn

参考书目

- Pattern Recognition and Machine Learning, *Christopher M. Bishop*, Springer
- Introduction to Machine Learning (3rd version) *Ethern Alpaydin*, MIT Press
- Pattern Recognition (4th version), *Sergios Theodoridis* and *Konstantinos Koutroumbas*.







参考资料

- 重要的国际期刊：
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
  - Journal of Machine Learning Research
  - Machine Learning
  - IEEE Computational Intelligence Magazine
  - Pattern Recognition
  - IEEE Transactions on Neural Networks and Learning Systems
  - IEEE Transactions on Cybernetics
  - IEEE Transactions on System and Man and Cybernetics: Systems
  - IEEE Transactions on Cognitive and Developmental Systems
  - IEEE Transactions on Robotics
  - IEEE Transactions on Image Processing
  - International Journal of Computer Vision
- 重要的国际会议：
  - ICML, NIPS,CVPR……

线上学习管理

- ✓ 中国大学MOOC，《模式识别与机器学习》，于元隆教授
- ✓ SPOC模式——我的学校云

- ✓ 课程实践——华为云+Harmony OS系统

- 基础理论+科研创新：
  - 期末笔试
  - 文献阅读+科研实训
  - 雨课堂、上机实践

小目标检测

异常检测

目标跟踪

虚拟现实

机器人

MindSpore框架
训练模型

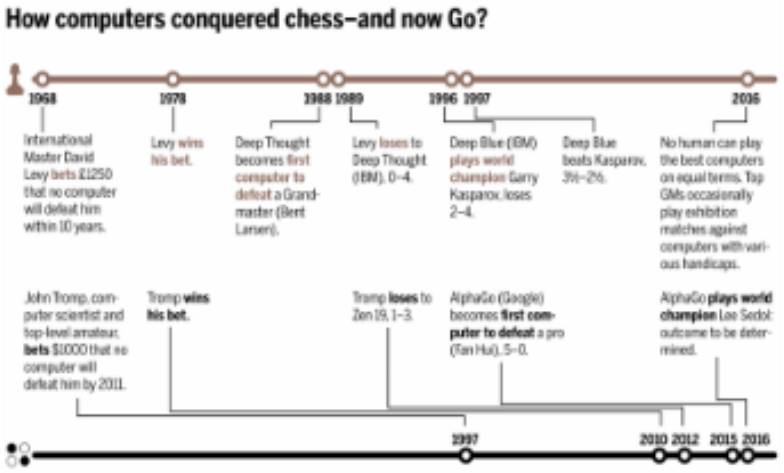
HAIWEI DevEco Studio工具
OpenHarmony软件
移植部署工具

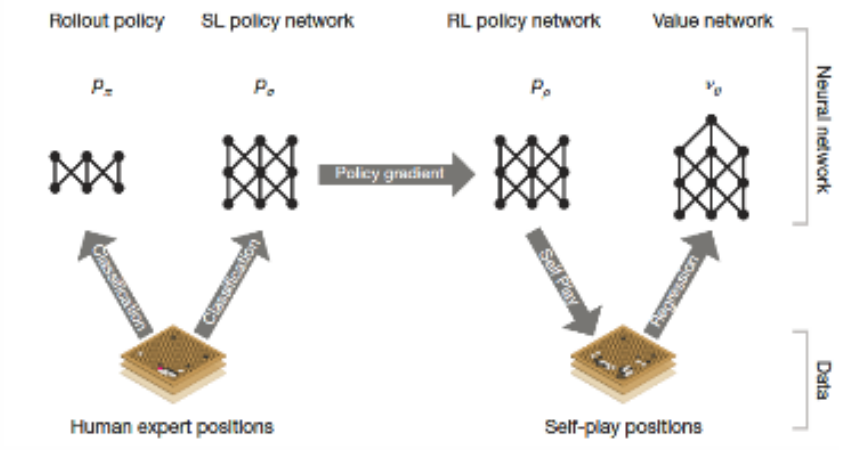
HarmonyOS体系架构与
AI开发接口

机器学习算法

昇腾全栈AI技术

第一代 AlphaGo





监督学习：16万人类选手棋局（约3000万棋谱）
围棋方格：19*19=361

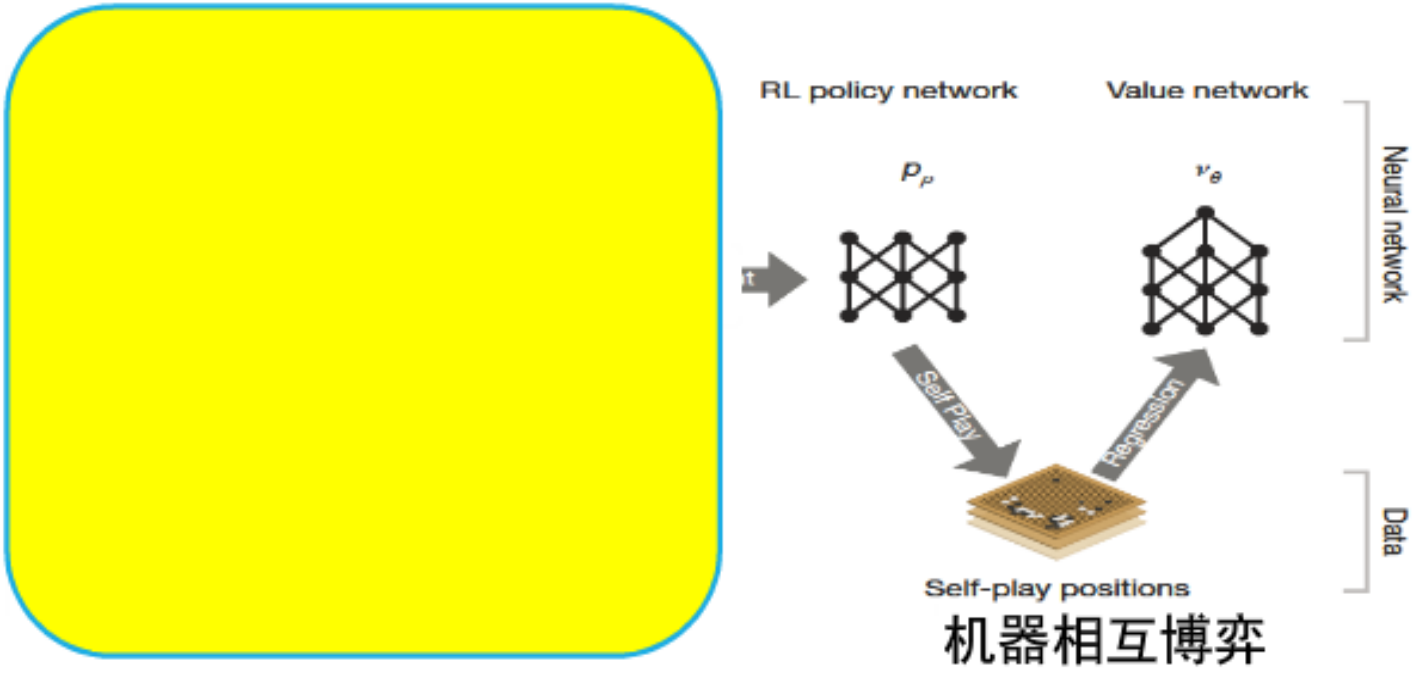
强化学习：机器对弈产生数以万计棋局

棋谱，九段数据
五段水平

九段水平
2016年战胜李世石

第二代AlphaGo：AlphaGo Master

AlphaGo vs AlphaGo: ***self-play games***

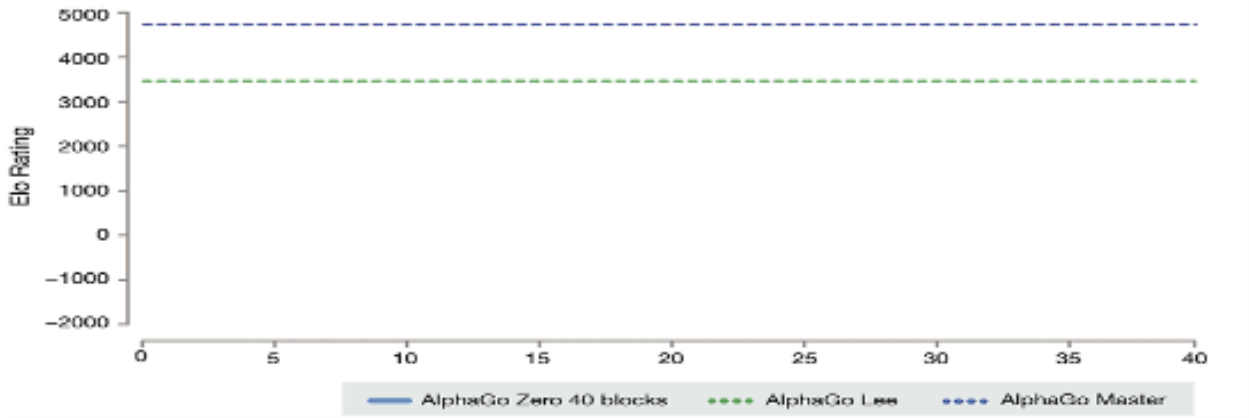


机器相互博弈

两个九段互下，基于谷歌公司的TPU的计算能力，战胜柯洁

第三代 ALphaGo Zero: starting *tabula rasa* (一张白纸绘蓝图)

based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules (learns 4,000 years of human knowledge in 40 days, “尧造围棋，丹朱善之”)

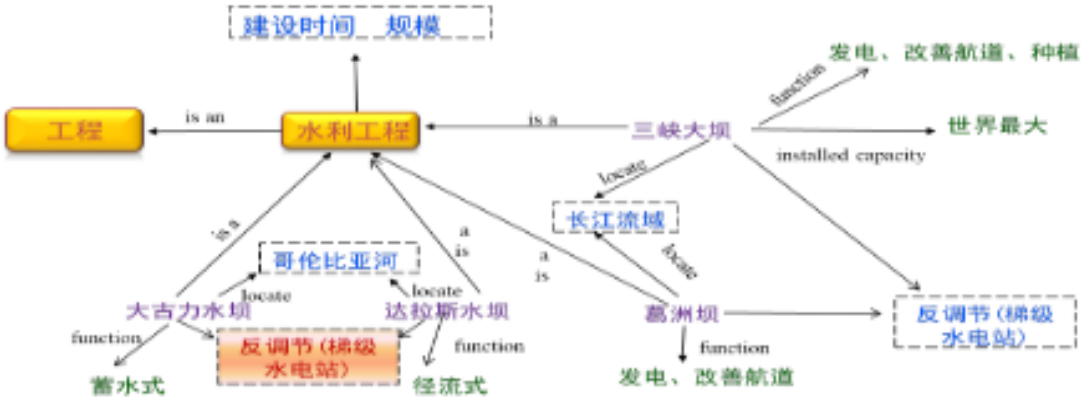


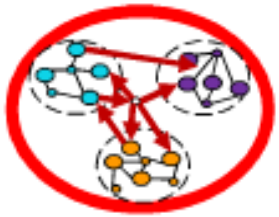
经过40天训练后，Zero总计运行约2900万次自我对弈，得以击败AlphaGo Master，比分为89比11

3天战胜1.0版（1.0打败李世石），21天战胜2.0（2.0打败柯洁）。
但Zero调用了谷歌公司所有的算力。

人工智能发展中的主流方法（1）

符号主义人工智能（Symbolic AI）为核心的逻辑推理





符号逻辑表示下的推理

用规则教

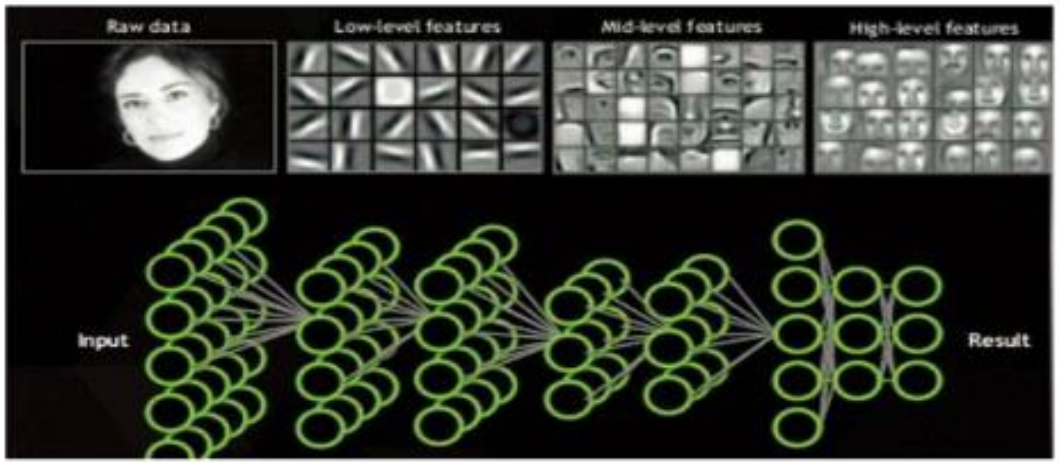
知识图谱

人工智能发展中的主流方法（2）

数据驱动（Data-Driven）为核心的机器学习

挖掘数据所蕴含的内在模式

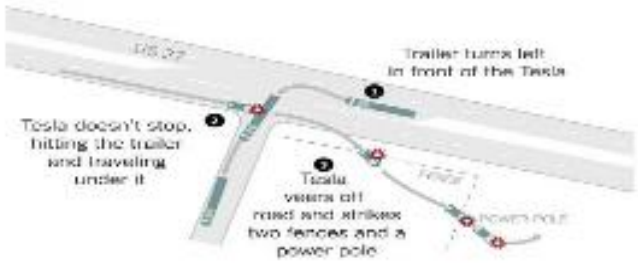
用大数据学（有监督）



机器挖掘得到的视觉模式

How Many Computers to Identify a Cat? 16,000

多少台机器机可识别一只猫(2012.6)



Source: The New York Times

特斯拉第一次车毁人亡(2016.5.7)

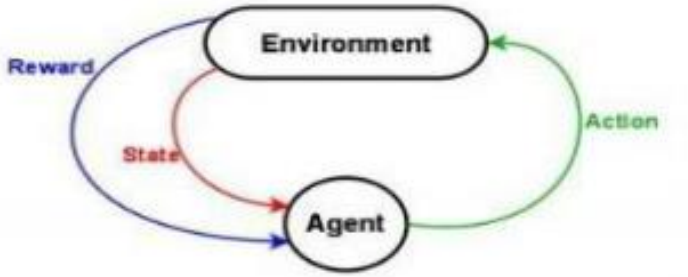
谷歌公司的图像标注系统 (2016.6)

https://www.tesla.cn/videos/autopilot-self-driving-hardware-neighborhood-long

特斯拉无人驾驶视频

人工智能发展中的主流方法（3）

探索与利用（Exploration and exploitation）为核心的强化学习



从经验中的策略学习

用问题引导（反馈牵引）


Reinforcement Renaissance

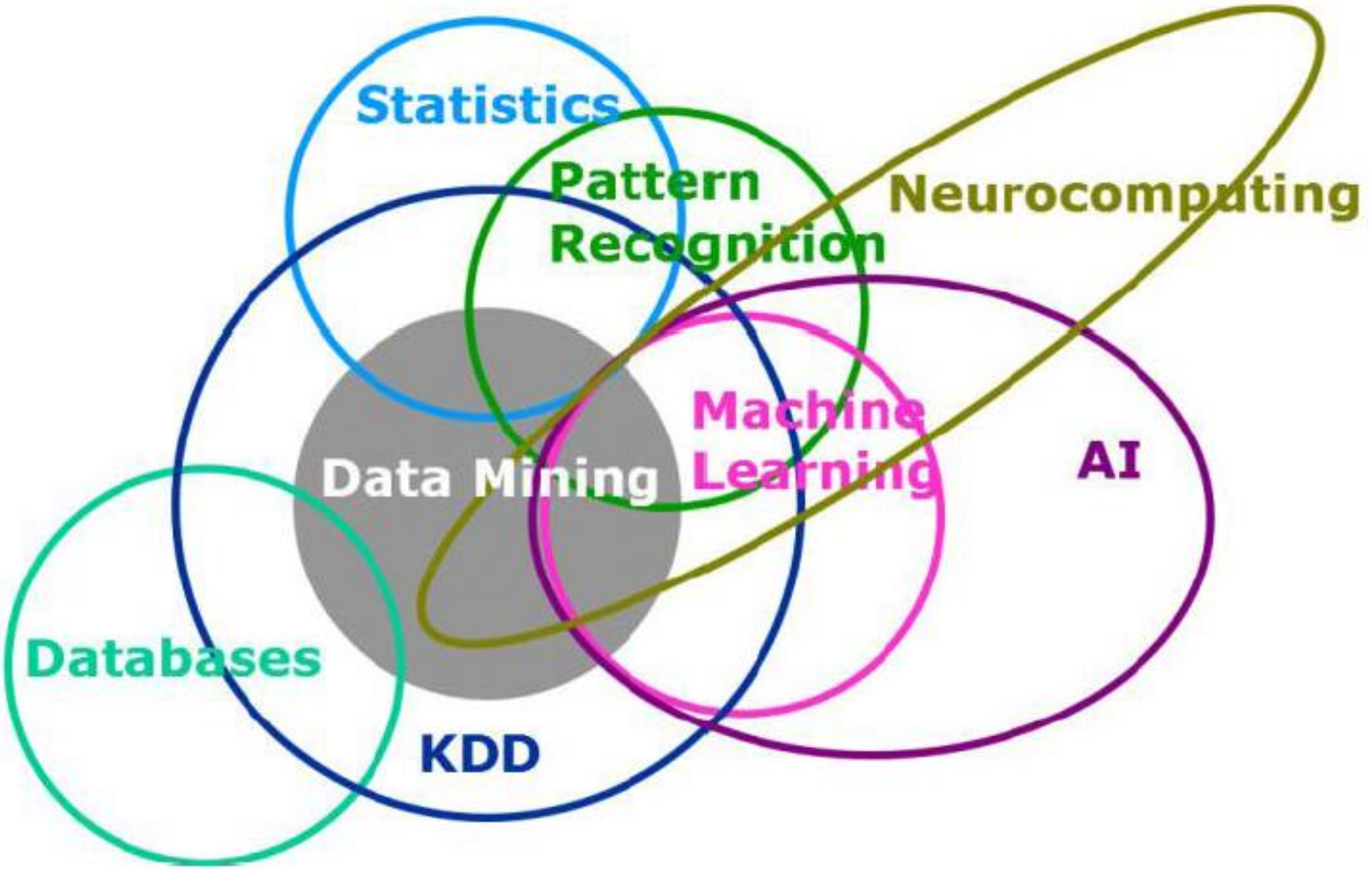
Reinforcement Renaissance, *Communications of the ACM*, 2016,59(8):12-14

人工智能三种主流方法区别

学习模式	优势	不足
用规则教	与人类逻辑推理相似，解释性强	难以构建完备的知识规则库
用数据学	直接从数据中学	以深度学习为例：依赖于数据、解释性不强
用问题引导	从经验中进行能力的持续学习	非穷举式搜索而需更好策略

从数据到知识与能力，能力增强是最终目标

三种学习方法的综合利用值得关注！


Statistics
Pattern Recognition
Neurocomputing
Machine Learning
AI
Data Mining
Databases
KDD

模式识别与机器学习

第一讲 模式识别基本概念

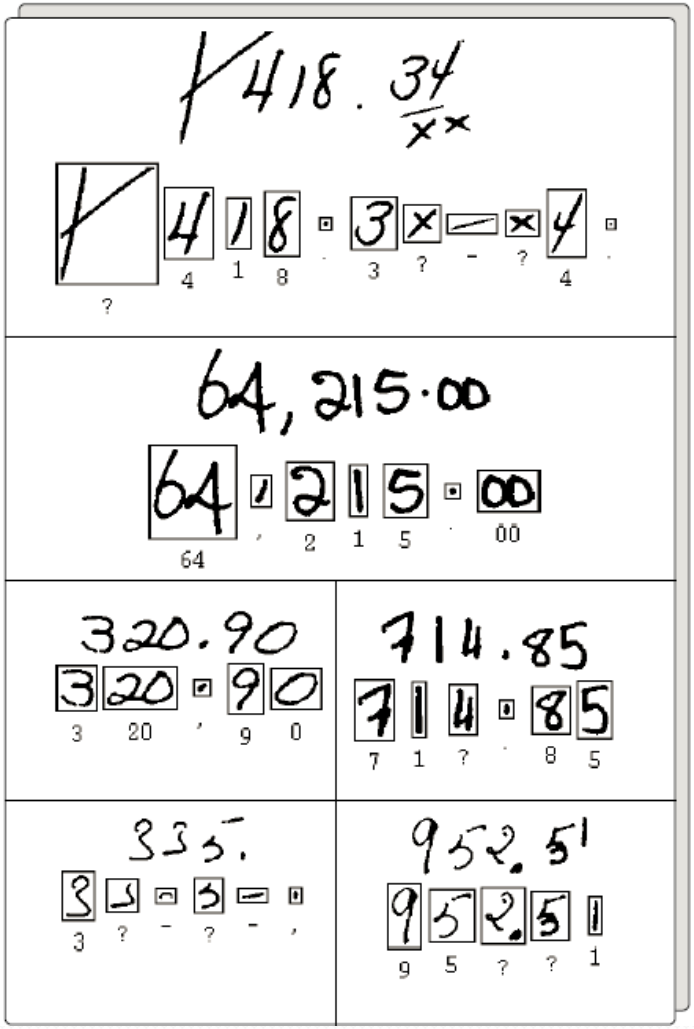
- 1.1.1 模式识别的概念
- 1.1.2 模式识别的发展简史
- 1.1.3 模式识别的应用

- **模式识别 – 直观，无所不在**
  - 人的识别：张三、李四
  - 周围物体的认知：符号、标志、桌子、椅子
  - 声音的辨别：人语、狗叫、汽车、火车
  - 气味的分辨：煎、炸、闷、炒、卤、炖

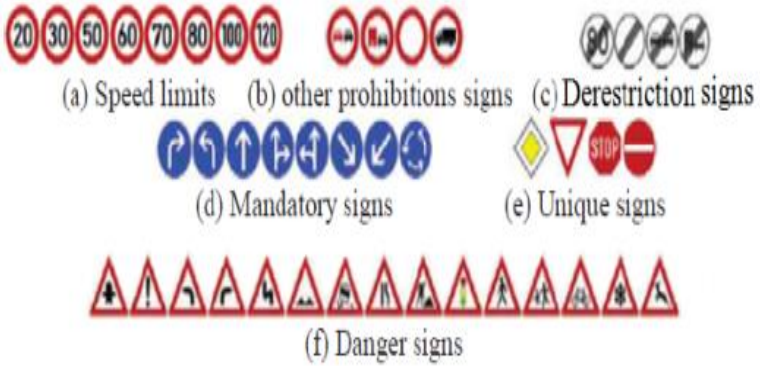
- **人和动物的模式识别能力是极其平常的，但对计算机来说却是非常困难的。**

- 1929年 G. Tauschek发明阅读机，能够阅读0-9的数字。
- 30年代 **Fisher**提出统计分类理论，奠定了统计模式识别的基础。
- 50年代 Noam Chemsky 提出形式语言理论——傅京荪 提出句法结构模式识别。
- 60年代 L.A.Zadeh提出了模糊集理论，模糊模式识别方法得以发展和应用。
- 80年代以Hopfield网、BP网为代表的**神经网络**模型导致人工神经元网络复活，并在模式识别得到较广泛的应用。
- 90年代小样本学习理论，**支持向量机**也受到了很大的重视。

- 手写体字符识别：OCR
  - ✓ 任务类型：多类分类
  - ✓ 输入：单个或者多个字符的图像
  - ✓ 输出：该字符的类别（字符语义）

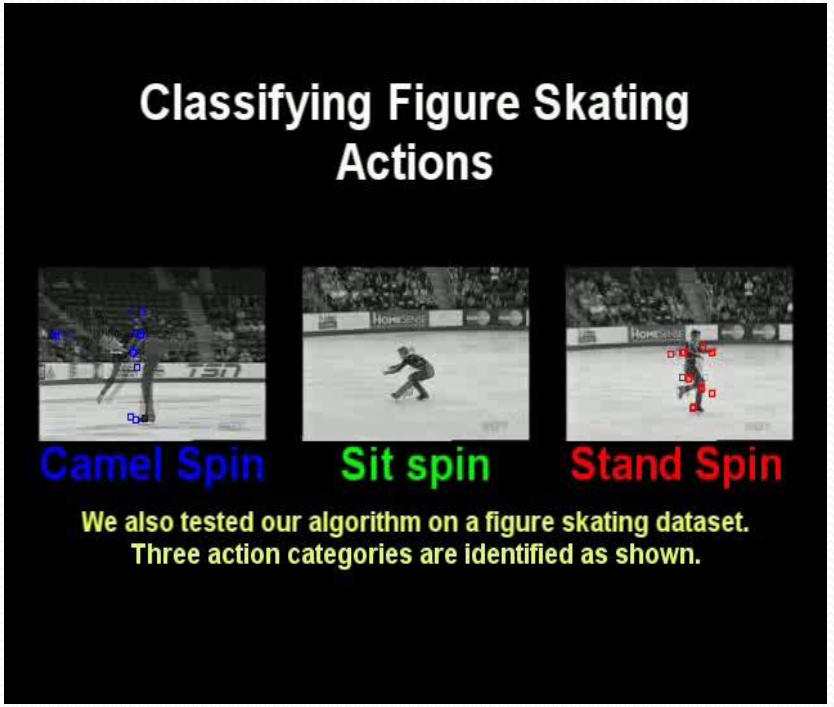


- 交通标志识别：辅助/无人驾驶
  - ✓ 输入：交通标志的图像
  - ✓ 输出：该标志的类别（标志语义）
  - ✓ 任务类型：多类分类



→ 模式识别系统 → 限速20公里/小时

- 动作识别
  - 任务类型：多类分类（时序数据）
  - 输入：一段视频（连续多帧图像）
  - 输出：动作所属的类型（动作语义）



- 语音识别
  - ✓ 任务类型：多类分类（音频时序数据）
  - ✓ 输入：一段语音
  - ✓ 输出：语音中的词语语义

记者：你最大的梦想是什么？
Journalist: what is your dream?
小冰：说大点：世界和平。说小点：晚上吃顿好的。
Xiaoice: World peace, or have a nice dinner

记者：如果有一天你死了，你希望后人怎么评价你？
Journalist: if you pass away one day, how would you like others to comment on you?
小冰：人生没有偶并不会不同。
Xiaoice: life won't be different without me

记者：你这辈子最幸福的事是什么？
Journalist: what is the happiest thing in your life?
小冰：活着。
Xiaoice: Being alive.

小冰聊天

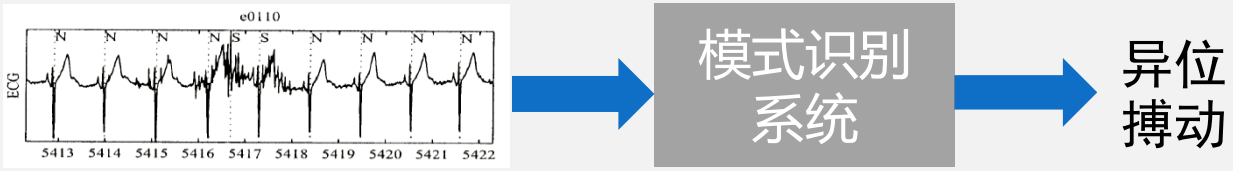
人生还有诗和远方

谷歌人工智能绘画

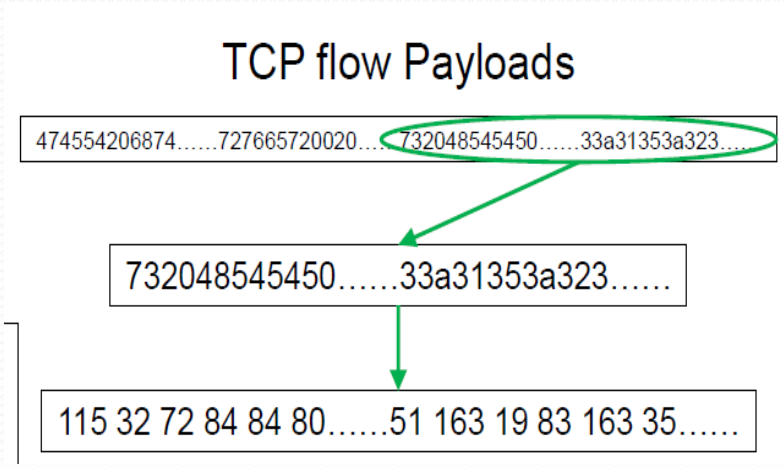
- 美国佐治亚理工大学的人工智能助教：Jill Watson于2016年春季上岗，三个月下来，未被学生发现，被推荐为优秀助教

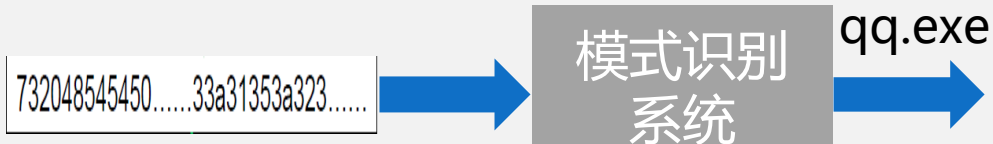
"Jill Watson" was 1 of 9 teaching assistants in an online grad course in Artificial Intelligence at the Georgia Institute of Technology. She performed admirably, and nobody suspected she wasn't a human. The only hint might have been that she responded perhaps a little too promptly to student questions – and she single-handedly answered 40% of them

- 心跳异位搏动识别
  - 任务类型：二类分类（心电图时序信号）
  - 输入：心电图(ECG)信号
  - 输出：心跳正常/有异位搏动（心跳状态语义）



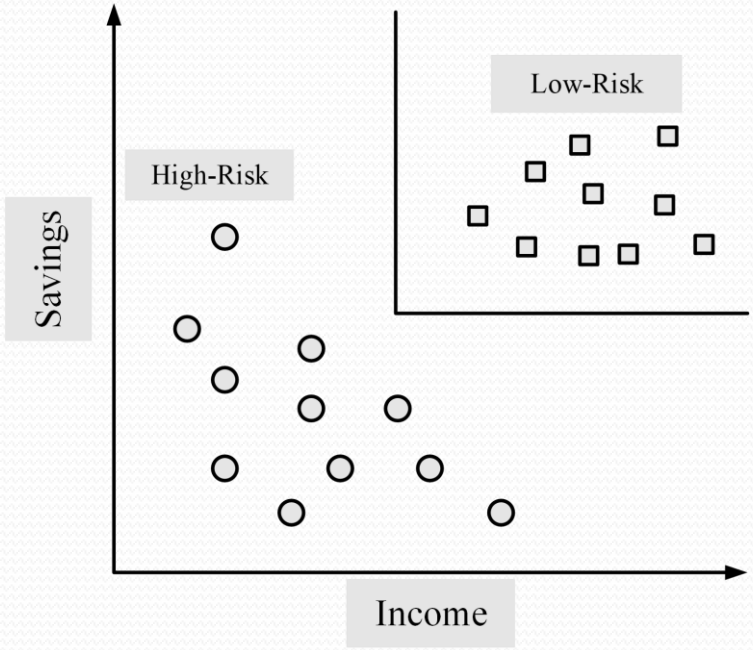
- 应用程序识别（基于TCP/IP流量）
  - 任务类型：多类分类
  - 输入：流量数据
  - 输出：应用程序/协议类别





Application	Precision	Protocol	Precision
foxmail.exe	1.0000	xshell.exe	0.9813
wpservice.exe	1.0000	baidumusic.exe	0.9808
taobaoprotect.exe	0.9984	fetion.exe	0.9779
wechat.exe	0.9983	qqmusic.exe	0.9730
liebao.exe	0.9978	qqdownload.exe	0.9615
weibo2015.exe	0.9974	yodaodict.exe	0.9542
lsass.exe	0.9945	itunes.exe	0.9429
sogoucloud.exe	0.9897	outlook.exe	0.9219
qq.exe	0.9884	thunder.exe	0.9168
pplive.exe	0.9870	iexplore.exe	0.8860

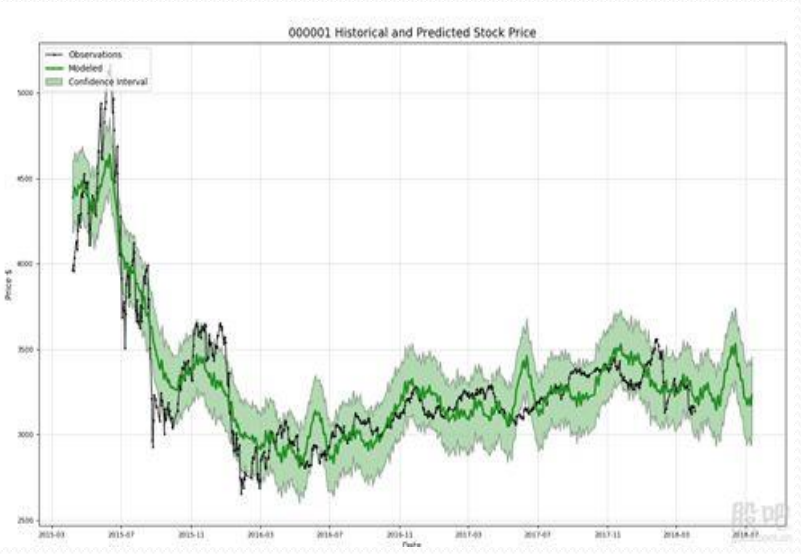
- 银行信贷识别
  - ✓ 任务类型：二类分类（数值数据）
  - ✓ 输入：用户信息（个人的收入、存款、年龄、职业、过去的还贷款记录等）
  - ✓ 输出：贷款（低风险）/不贷 （高风险）



收入=5000
存款=1000
→ 模式识别系统 → 高风险

- 股票价格预测
  - ✓ 任务类型：回归（预测）
  - ✓ 输入：股票历史数据（开盘价、收盘价、最高价、交易量等）
  - ✓ 输出：（明日）股价

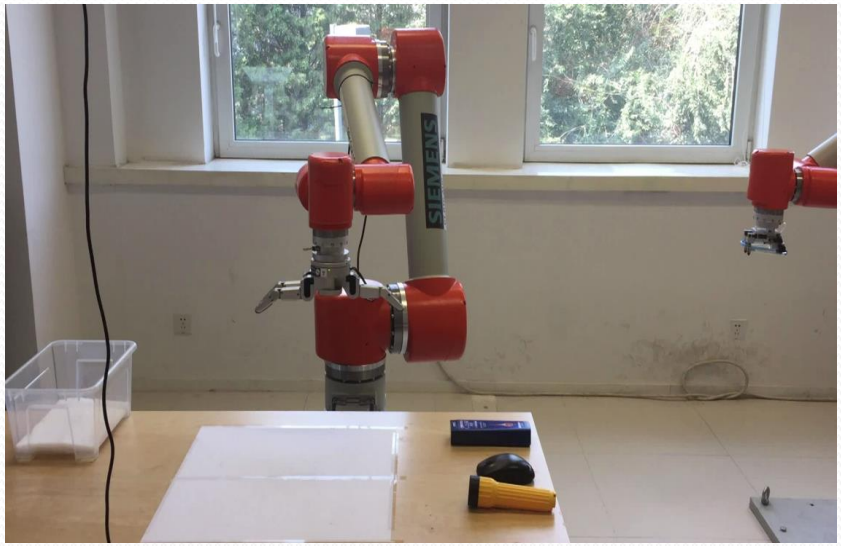
$$\begin{bmatrix} 5.3 \\ 3.0 \\ \vdots \end{bmatrix} \rightarrow \boxed{\text{模式识别系统}} \rightarrow 4.98$$



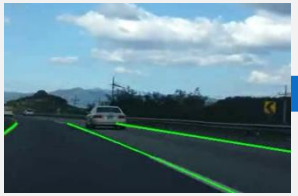
- 机械手目标抓取点位姿
  - ✓ 功能：回归（决定怎么抓）
  - ✓ 输入：一幅图像（待抓取目标）
  - ✓ 输出：每个机械手指抓取点的位置坐标

模式识别系统

$$\begin{bmatrix} 3.7 \\ 6.7 \\ 9.8 \\ 6.7 \\ \vdots \end{bmatrix}$$



- 无人驾驶
  - 功能：回归（决定车辆当前的控制量）
  - 输入：图像、激光雷达等
  - 输出：车体速度和方向控制量



模式识别系统

$$\begin{bmatrix} 46\ \text{km/h} \\ 35^{\circ} \end{bmatrix}$$

- 1.2 模式识别

- 根据任务，模式识别可以划分为“分类”和“回归”两种形式。

分类(Classification)

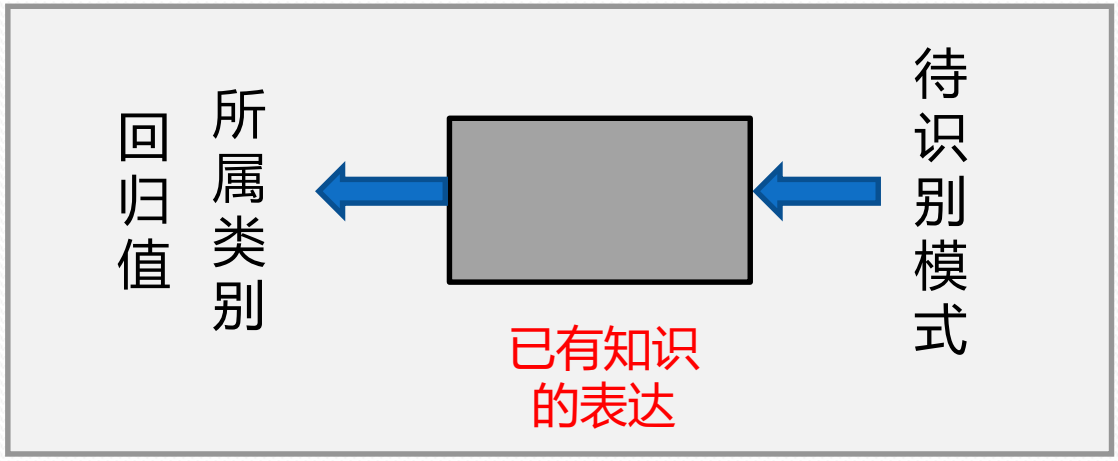
- 输出量是离散的类别表达，即输出待识别模式所属的类别。
- 二类/多类分类

回归(Regression)

- 输出量是连续的信号表达（回归值）
- 输出量维度：单个/多个维度
- 回归是分类的基础：离散的类别值是由回归值做判别决策得到的。

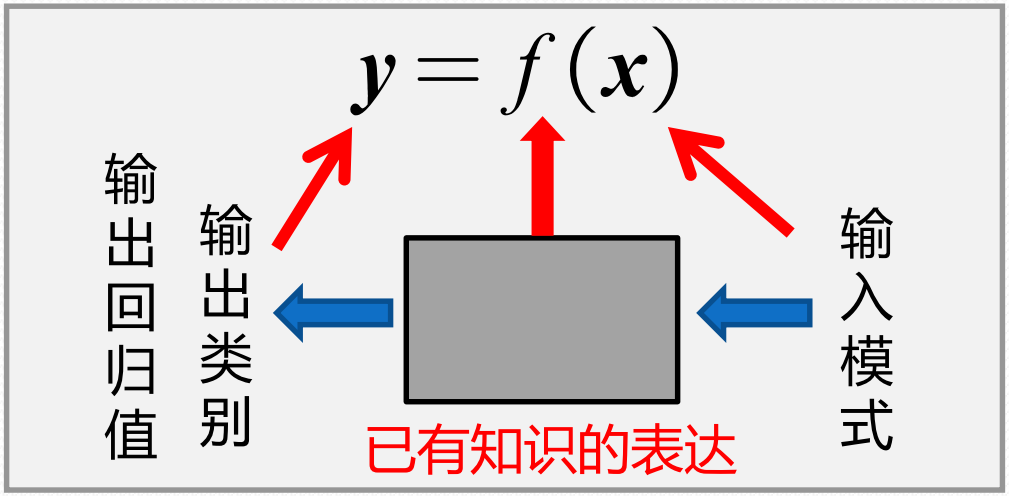
模式识别

- 模式识别：根据已有知识的表达，针对待识别模式，判别决策其所属的类别或者预测其对应的回归值。
- 由此可见，模式识别本质上是一种推理（inference）过程。



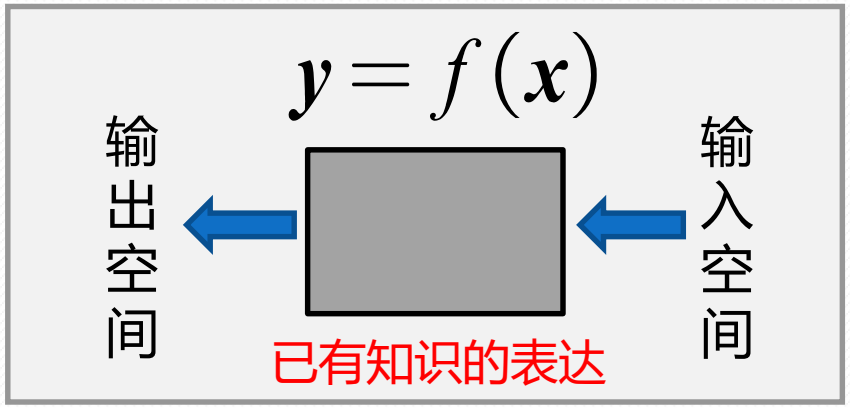
模式识别

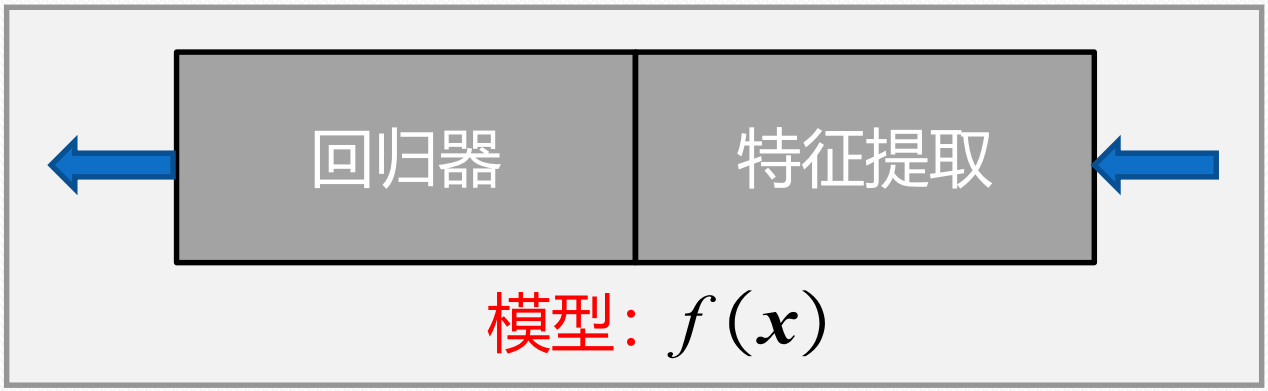
- 数学解释：模式识别可以看做一种函数映射$f(\boldsymbol{x})$，将待识别模式$\boldsymbol{x}$从输入空间映射到输出空间。函数$f(\boldsymbol{x})$是关于已有知识的表达。
- 函数$f(\boldsymbol{x})$的形式：可解析表达的、难以解析表达的。
- 函数$f(\boldsymbol{x})$的输出：确定值、概率值。



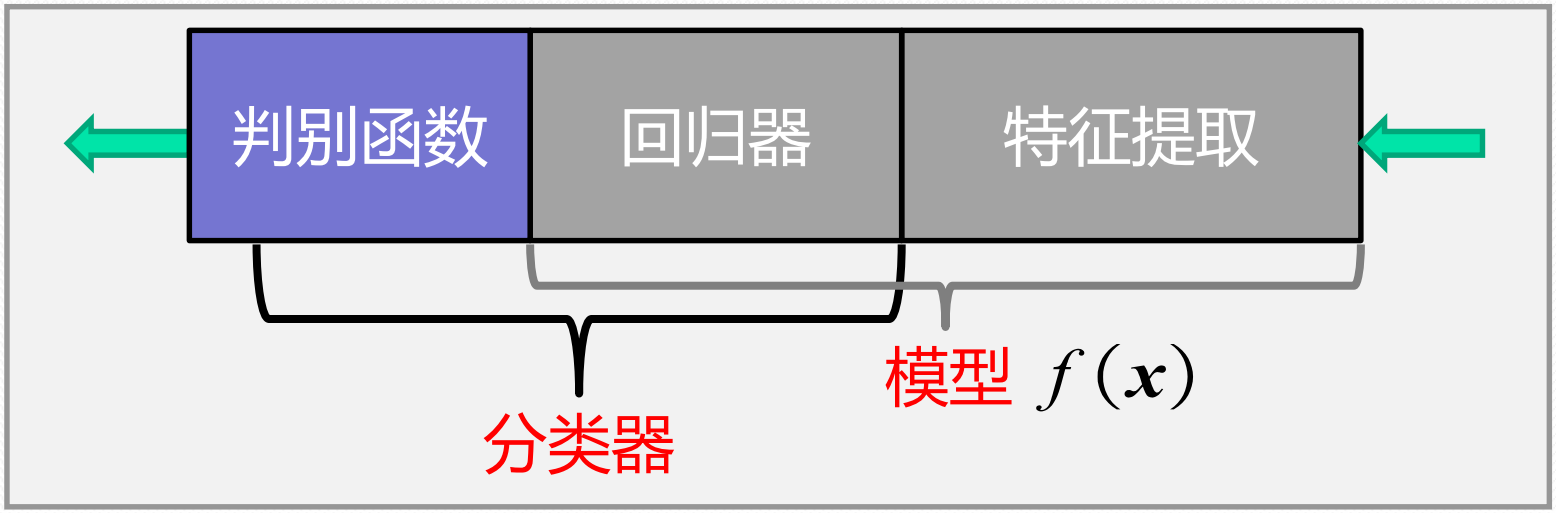
输入与输出空间

- 输入空间：原始输入数据x所在的空间。
  - ✓ 空间维度：输入数据的维度。
- 输出空间：输出的类别/回归值y所在的空间。
  - ✓ 空间维度：1维、类别的个数（>2）、回归值的维度。





用于回归



用于分类

判别函数使用一些特定的非线性函数来实现，记作函数g。

判别器：二类分类

- 使用sign函数：判断回归值大于0还是小于0。

判别器：多类分类

- 使用max函数：取最大的回归值所在维度对应的类别。

可见，由于判别函数通常固定已知，所以不把它当做模型的一部分。

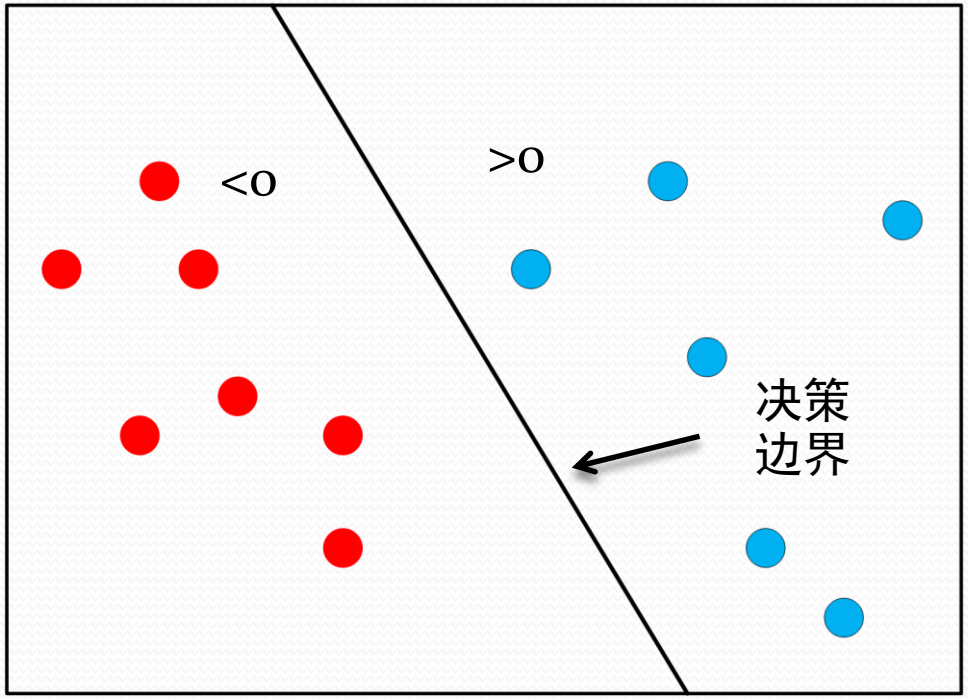
以二类分类为例：

判别公式

$$\boldsymbol{x} \in \begin{cases} C_1 & \text{if} \quad f(\boldsymbol{x}) > 0 \\ C_2 & \text{if} \quad f(\boldsymbol{x}) < 0 \end{cases}$$

决策边界

$$f(\boldsymbol{x}) = 0$$



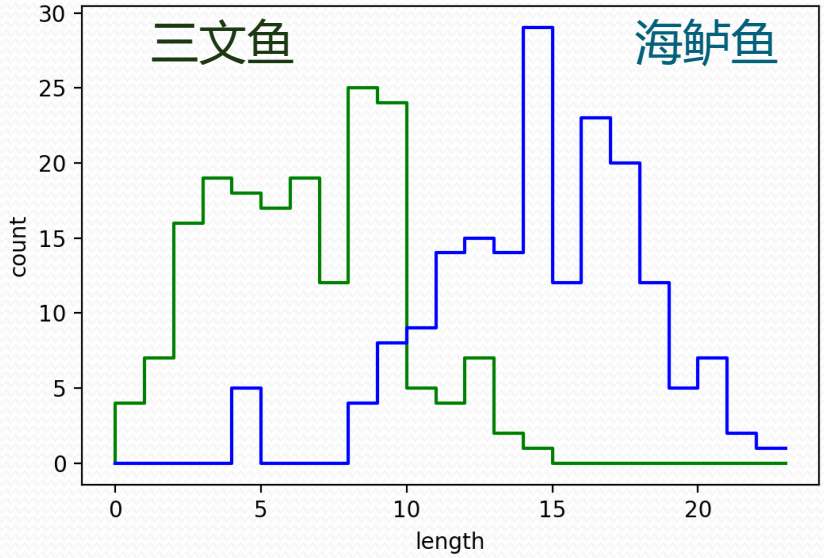
- 1.3 特征与特征空间

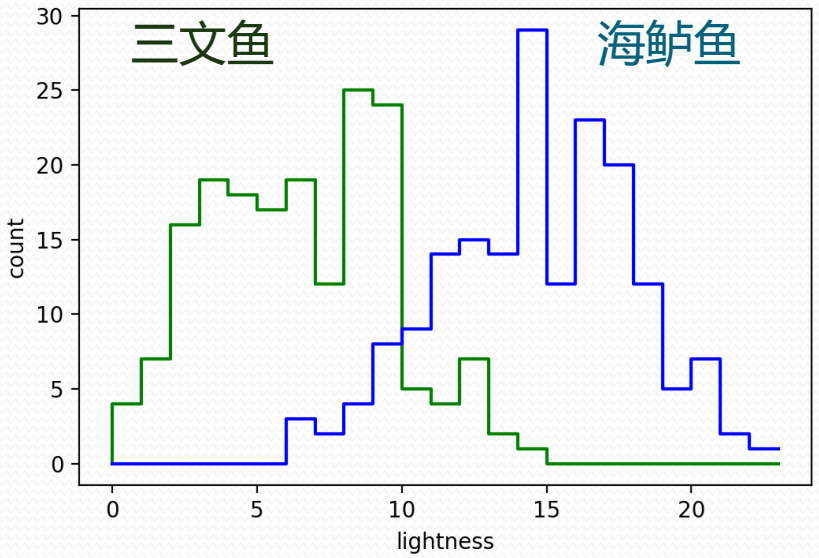
特征（Feature）

- 可以用于区分不同类别模式的、可测量的量。
  - ✓ 例子：针对橙子和苹果两个类，形状or颜色？
- 输入数据也可以看做是一种原始特征表达。

特征的特性

- 具有辨别能力：提升不同类别之间的识别性能。
  - ✓ 基于统计学规律、而非个例。
  - ✓ 例子：三文鱼和海鲈鱼。



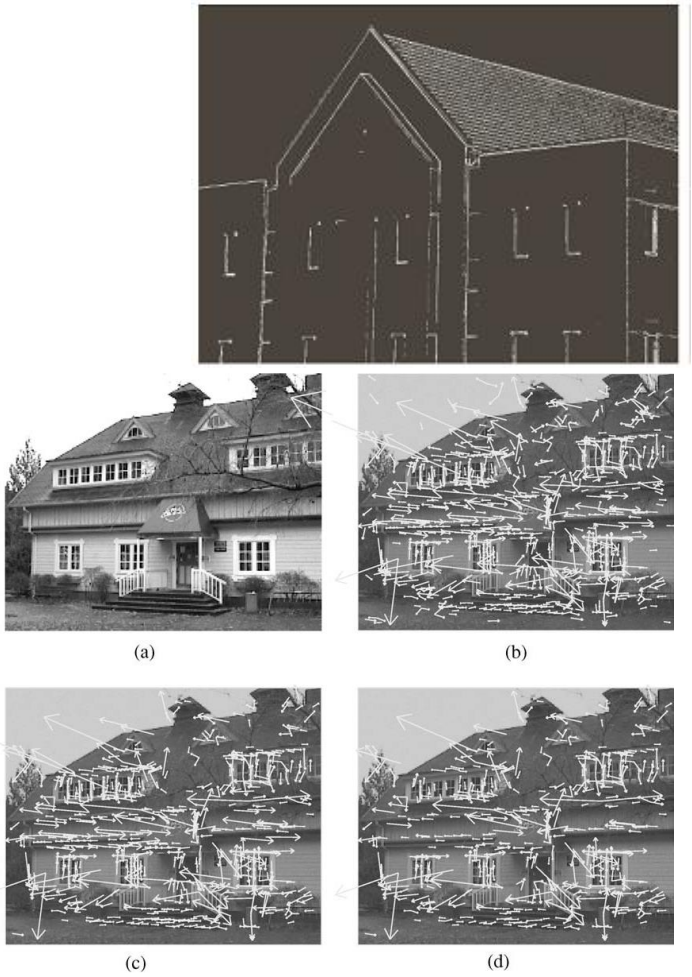


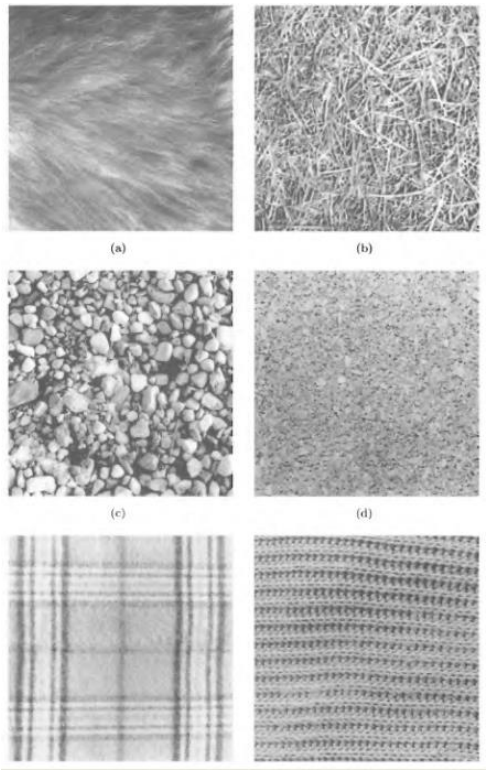
特征的特性

- 鲁棒性：针对不同的观测条件，仍能够有效表达类别之间的差异性。
  - ✓ 以交通标志识别为例：噪声、光照变化、视角变化（尺度、形状、旋转等）、遮挡……

- 如何得到特征？特征提取技术
  - 手动设计
  - 自动学习

- 以图像信号为例：
  - 边缘特征
  - 点特征
  - 纹理特征
  - 形状特征





特征向量的定义&性质

- 特征向量（feature vector): 多个特征构成的（列）向量 $\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$
  - ✓ 特征向量的长度（模）： $\|\boldsymbol{x}\|=\sqrt{\boldsymbol{x}^T\boldsymbol{x}}=\sqrt{\sum_{j=1}^{p}x_j^2}$
  - ✓ 特征向量的方向（单位向量）： $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$
- 特征向量还可以表达为：模长（标量）× 方向（单位向量）

用实数表示特征时，模式通常表示成特征向量：

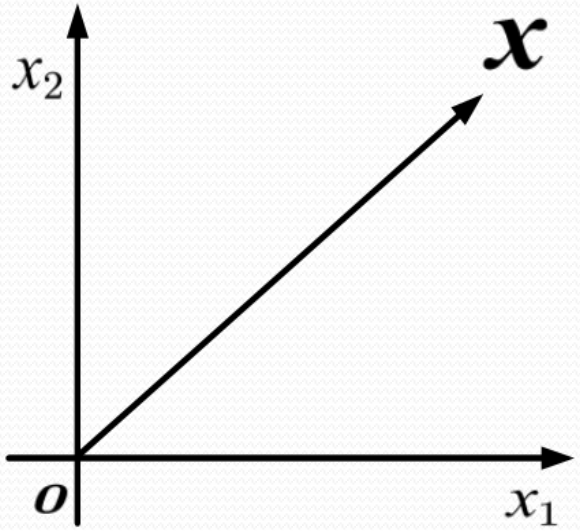
$$x = \begin{bmatrix} x_1 \\ x_2 \\ M \\ x_d \end{bmatrix} \in X \subset R^d$$

$$\xrightarrow{32\times 32 = 1024} x = [f_1, f_2, \cdots, f_{1024}]$$

$$\longrightarrow X = [x_1, x_2, \cdots, x_n] = \begin{bmatrix} f_{1,1} & \cdots & f_{n,1} \\ \vdots & \ddots & \vdots \\ f_{1,1024} & \cdots & f_{n,1024} \end{bmatrix}$$

特征空间

- 每个坐标轴代表一维特征
- 空间中的每个点代表一个模式（样本）
- 从坐标原点到任意一点（模式）之间的向量即为该模式的特征向量。



- 1.4 特征的相似度度量

- 由于每个特征向量代表一个**模式**，所以度量**特征向量两两**之间的**相关性**是识别模式之间是否相似的基础。

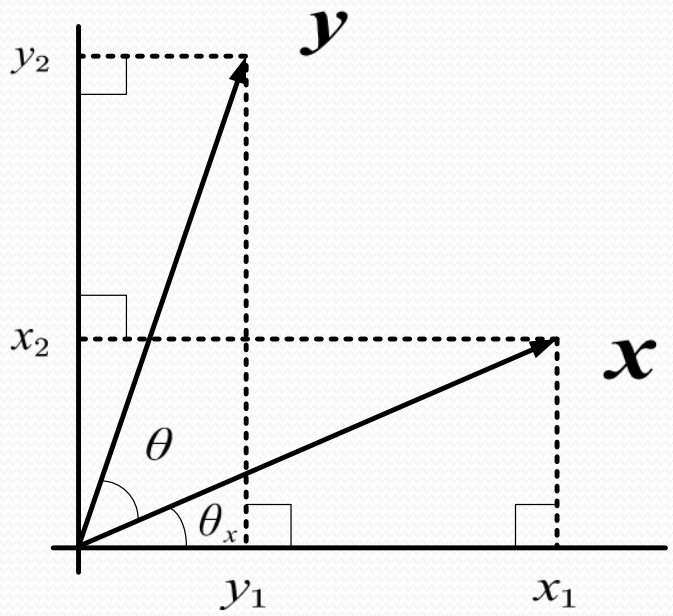
点积（dot product）的代数定义

$$\boldsymbol{x} \bullet \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y} = \mathbf{y}^T \boldsymbol{x} = \sum_{j=1}^{p} x_j y_j$$

- 点积结果是一个**标量**表达。
- 点积具备对称性。
- 点积是一个线性变换。

$$\boldsymbol{x} \bullet \boldsymbol{y} = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos\theta$$

- 证明:



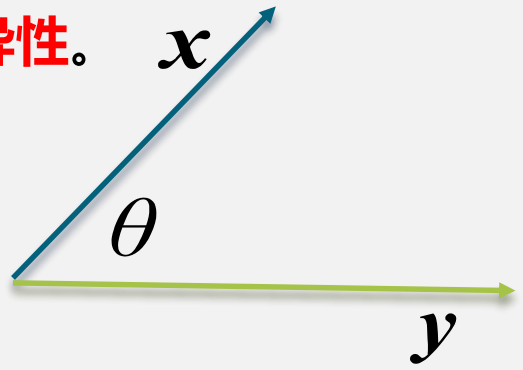
$$\begin{aligned}
\boldsymbol{x} \bullet \boldsymbol{y} &= x_1 y_1 + x_2 y_2 \\
&= \|\boldsymbol{x}\| \cos\theta_x \|\boldsymbol{y}\| \cos(\theta + \theta_x) \\
&\quad + \|\boldsymbol{x}\| \sin\theta_x \|\boldsymbol{y}\| \sin(\theta + \theta_x) \\
&= \|\boldsymbol{x}\| \|\boldsymbol{y}\| [\cos\theta_x (\cos\theta \cos\theta_x - \sin\theta \sin\theta_x) \\
&\quad + \sin\theta_x (\sin\theta_x \cos\theta + \cos\theta_x \sin\theta)] \\
&= \|\boldsymbol{x}\| \|\boldsymbol{y}\| [\cos\theta \cos^2\theta_x + \cos\theta \sin^2\theta_x] \\
&= \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos\theta
\end{aligned}$$

点积的几何定义

- 可见，点积可以表征两个特征向量的共线性，即方向上的**相似程度**。
- 点积为0，说明两个向量是正交的(orthogonal)。

- 两个向量的夹角：反映两个向量在**方向上的差异性**。

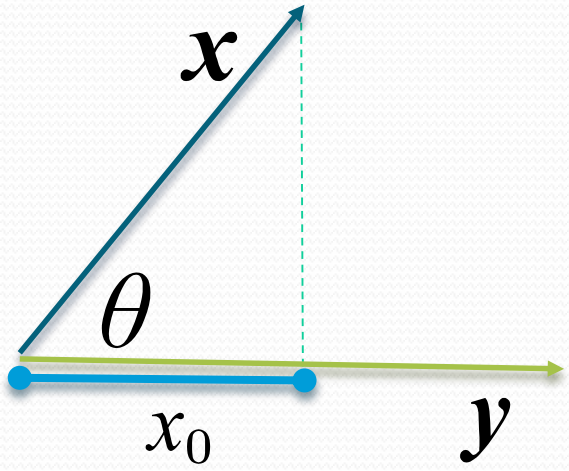
$$\cos\ \theta = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$



投影

- 向量$\boldsymbol{x}$到$\boldsymbol{y}$的投影（projection）：将向量$\boldsymbol{x}$垂直投射到向量$\boldsymbol{y}$方向上的长度（标量）。

$$x_0 = \|\boldsymbol{x}\|\cos\theta$$



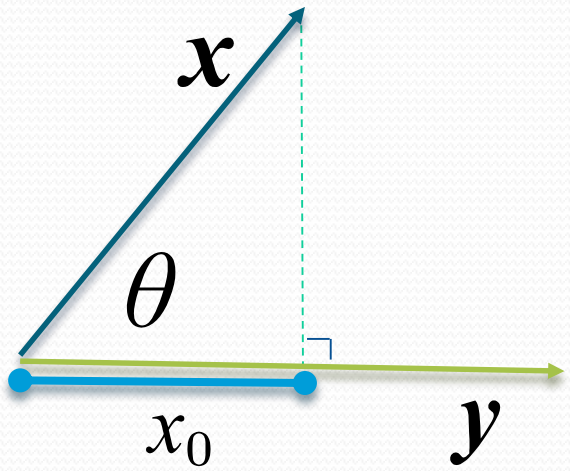
投影的含义

- 投影的含义：向量x分解到向量y方向上的程度。能够分解的越多，说明两个向量方向上越相似。
  - ✓ 当$\theta = 0^{\circ}$时，完全等同
  - ✓ 当$\theta = 90^{\circ}$时，分解量为0
- 向量投影不具备对称性。
- 投影向量：

$$\boldsymbol{x}_0 = \frac{\|\boldsymbol{x}\|\cos\theta}{\|\boldsymbol{y}\|}\boldsymbol{y}$$

- 向量点积还可以通过投影来表达：
  - ✓ 向量x和y的点积=向量$x \rightarrow y$的投影×向量y的模长。

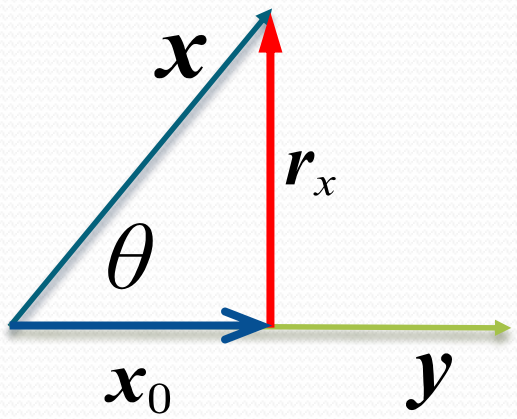
$$x \bullet y = x_0 \| y \|$$



残差向量

- 残差向量（residual vector）：向量$\boldsymbol{x}$分解到向量$\boldsymbol{y}$方向上得到的投影向量与原向量$\boldsymbol{x}$的误差。

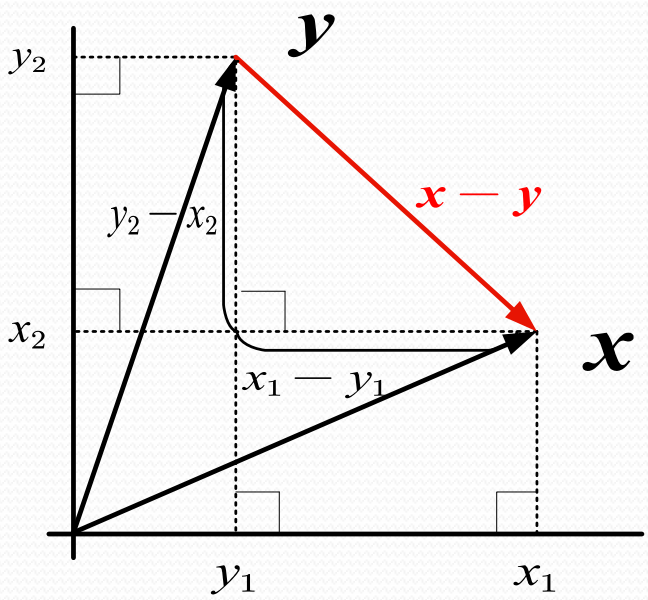
$$\boldsymbol{r}_x = \boldsymbol{x} - \boldsymbol{x}_0 = \boldsymbol{x} - \frac{\|\boldsymbol{x}\|\cos\theta}{\|\boldsymbol{y}\|}\boldsymbol{y}$$



欧式距离

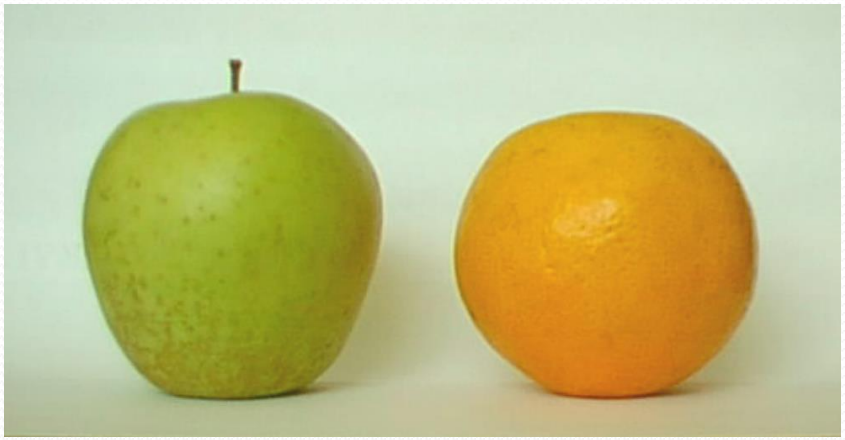
- 两个特征向量之间的欧式距离：表征两个向量之间的相似程度（综合考虑方向和模长）。

$$d(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\mathbf{y}) = \sum_{\mathrm{j}=1}^{p}(x_j - y_j)^2$$



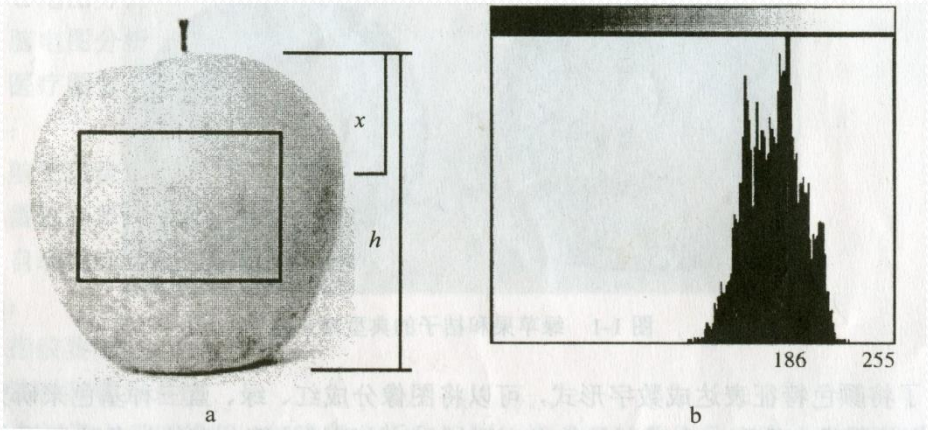
- 1.5.1 分类决策
- 1.5.2 回归问题
- 1.5.3 描述问题

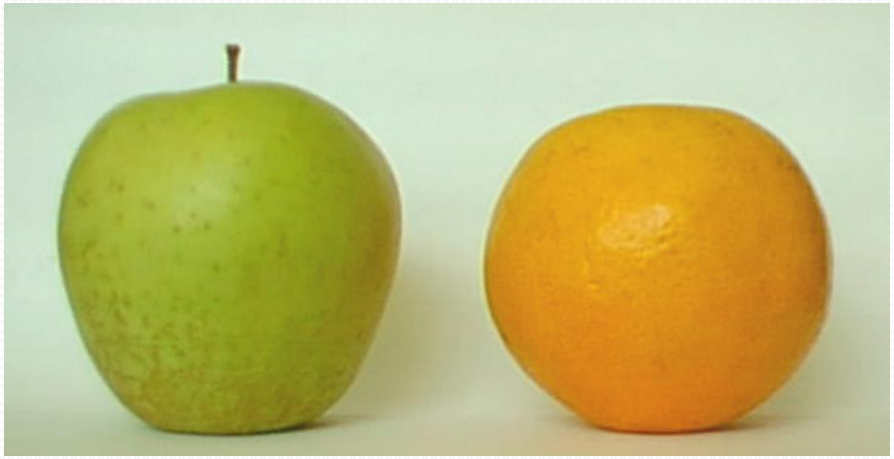
分类决策



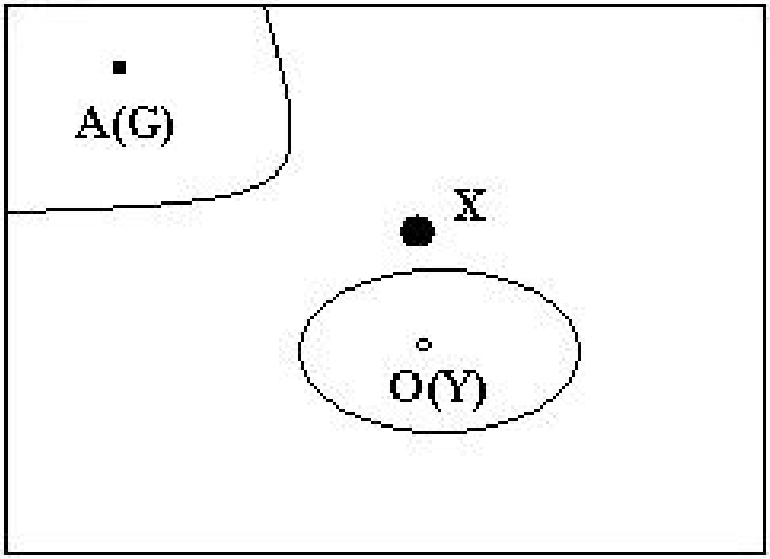
- 考虑采用颜色和形状作为特征，建立二维的特征空间，即：

  $x=[x_1\ x_2]^T=$[颜色 形状]^T



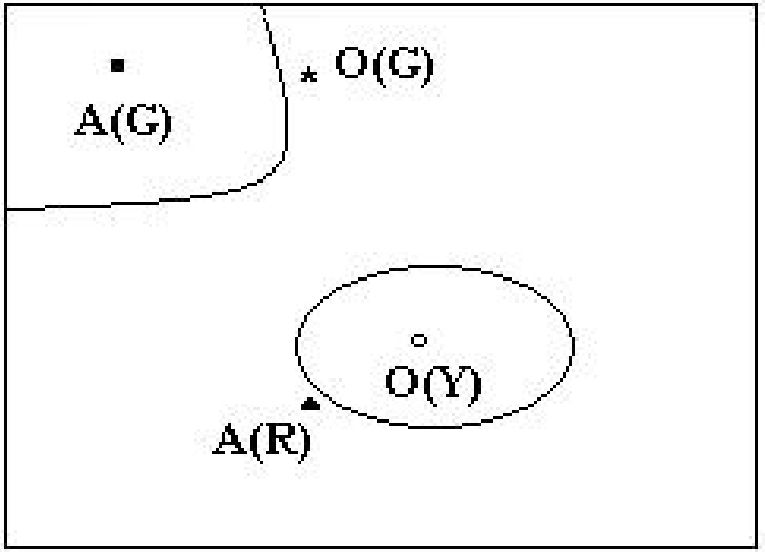


color



shape

color



shape

- 产生错分的原因
  - 使用的特征不恰当或者不充分
  - 用来训练分类器的样本不够全面和具有代表性
  - 类别之间存在交集
  - 分类器的效率不够高

回归问题

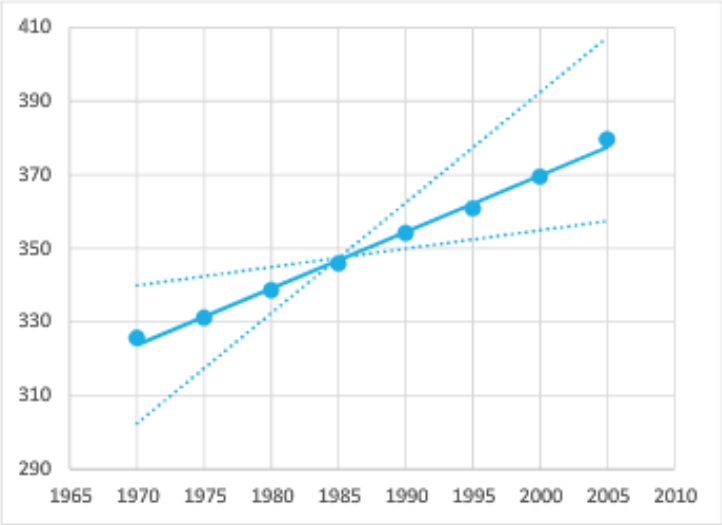
线性回归模型例子

下表给出了莫纳罗亚山（夏威夷岛的活火山）从1970年到2005年每5年的二氧化碳浓度，单位是百万分比浓度 （Parts Per Million，ppm）。

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2 (y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

问题：1) 给出1984年二氧化碳浓度值；2) 预测2010年二氧化碳浓度值

线性回归模型例子



莫纳罗亚山地区时间年份与二氧化碳浓度之间的一元线性回归模型（实线为最佳回归模型）

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2 (y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

代入

回归模型：$y = ax + b$

求取：最佳回归模型是最小化残差平方和的均值，即要求8组(x, y)数据得到的残差平均值$\frac{1}{N}\sum(y - \tilde{y})^2$最小。残差平均值最小只与参数$a$和$b$有关，最优解即是使得残差最小所对应的$a$和$b$的值。

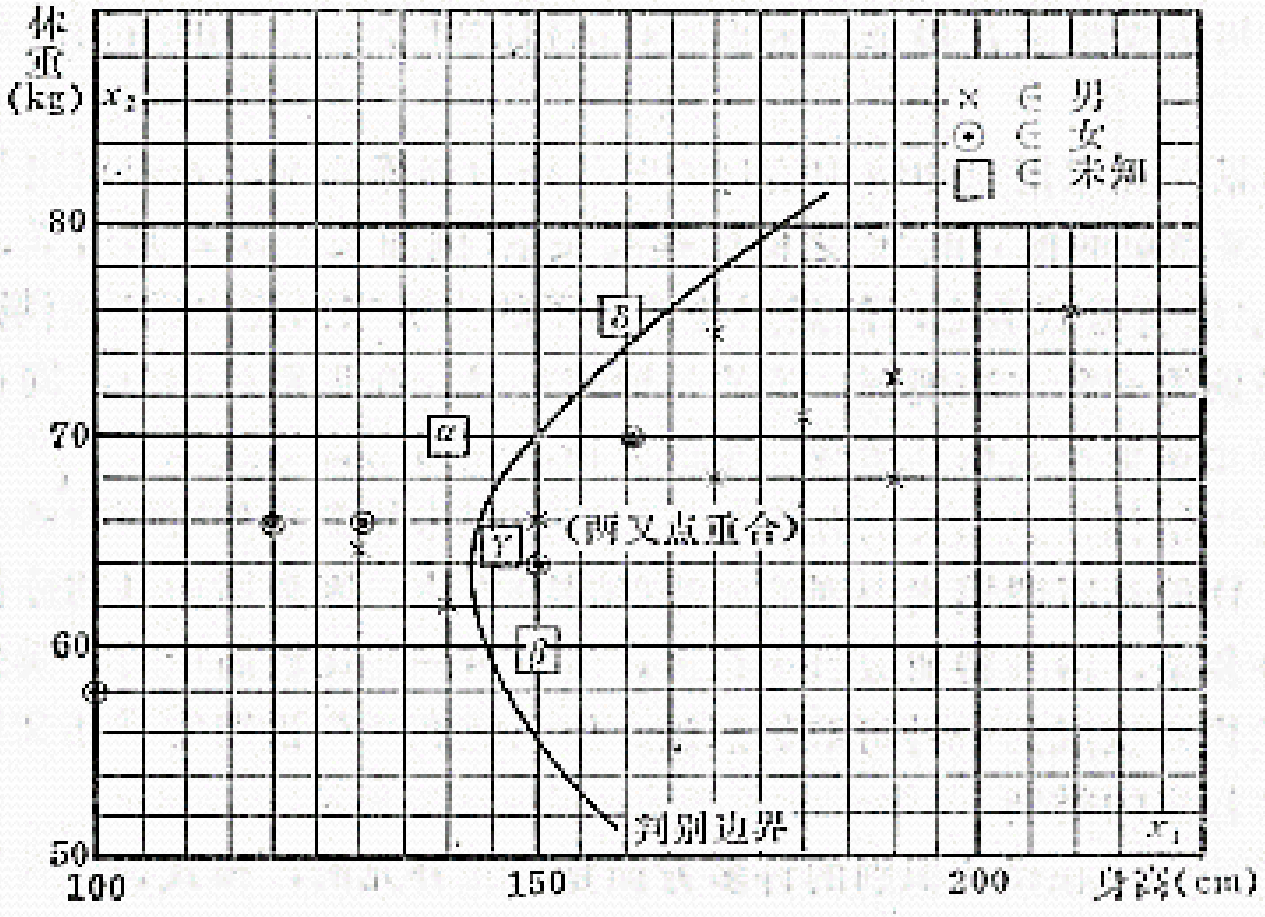
- **实例：**

19名男女同学进行体检，测量了身高和体重，但事后发现其中有4人忘记填写性别，试问（在最小错误的条件下）这4人是男是女？体检数值如下：

编　号	身高(cm)	体重(kg)	性　别	编　号	身高(cm)	体重(kg)	性　别
1	170	68	男	11	140	62	男
2	130	66	女	12	150	64	女
3	180	71	男	13	120	66	女
4	190	73	男	14	150	66	男
5	160	70	女	15	130	65	男
6	150	66	男	α	140	70	?
7	190	68	男	β	150	60	?
8	210	76	男	γ	145	65	?
9	100	58	女	δ	160	75	?
10	170	75	男				

- 分析：
  - 待识别的模式：性别（男或女）
  - 测量的特征：身高和体重
  - 训练样本：15名已知性别的样本特征
  - 目标：希望借助于训练样本的特征建立判别函数（即数学模型）

- **解：**
  - 第一步：由训练样本得到的**特征空间分布**图



- 第二步：训练过程

从图中训练样本的分布情况，找出男、女两类特征各自的聚类特点，从而求取一个**判别函数**（直线或曲线）。

- 第三步：识别过程

只要给出待分类的模式特征的数值，看它在特征平面上**落在判别函数的哪一侧**，就可以判别是男还是女了。

- 1.6 机器学习

模型如何得到?

使用机器学习技术!

训练样本

- 一组训练样本（数据），记作 $\{\boldsymbol{x}_n\}_{n=1,\cdots,N}$
  - ✓ 每个训练样本$\boldsymbol{x}_n \in R^p$，都是通过采样得到的一个模式，即输入特征空间中的一个向量；通常是高维度（即 p 很大），例如一幅图像。
  - ✓ 训练样本可以认为是尚未加工的原始知识，模型则是经过学习（即加工整理归纳等）后的真正知识表达。
  - ✓ 所有训练样本假设满足independent and identical distribution (iid) 。
  - ✓ 如果想学得好，这组训练样本要覆盖模型所有可能的分布空间。

拿什么学

模型的参数和结构

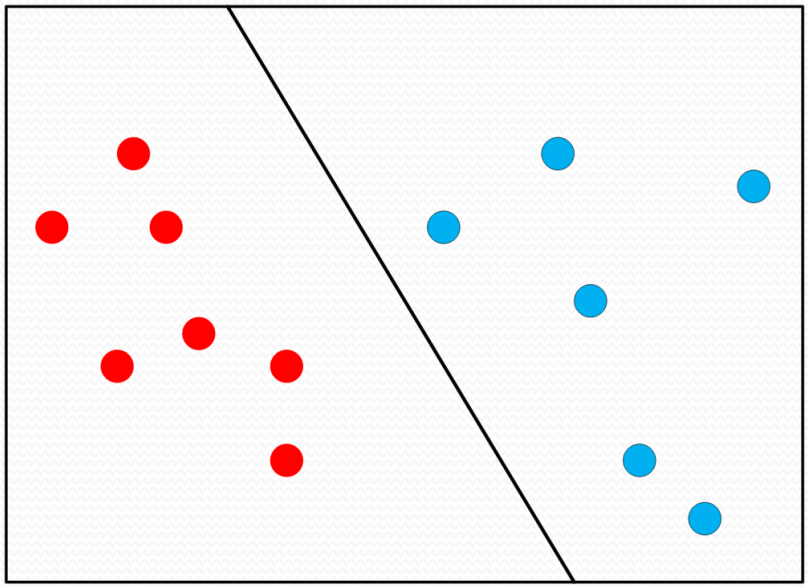
$$\boldsymbol{\theta} = \{\theta_1, ..., \theta_M\}$$

- 模型的参数：$\boldsymbol{\theta} = \{\theta_1, ..., \theta_M\}$
- 模型的结构：函数f的形式。

- 可见，模型结构决定了模型有哪些参数。
- 通常情况下，模型的结构是设计人员事先给定的。
- 如何学习模型结构是当前和未来机器学习领域的研究内容之一。

学什么

线性模型

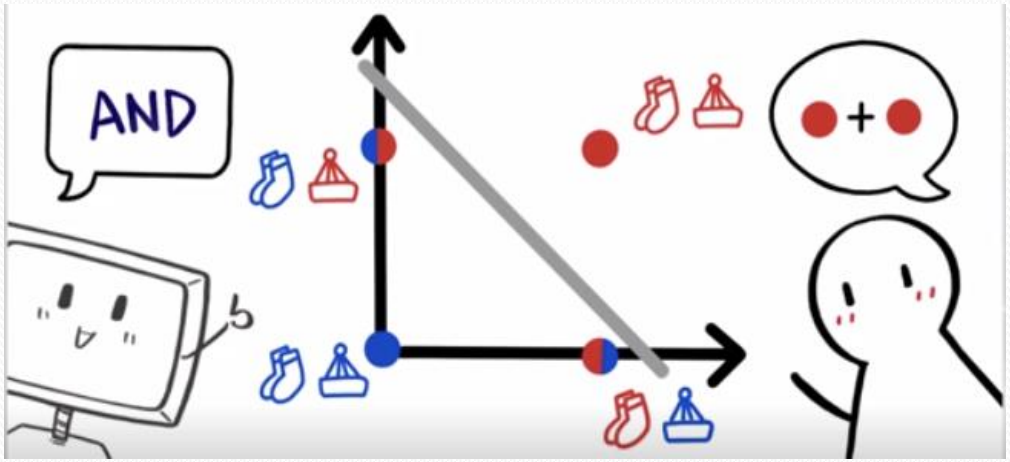


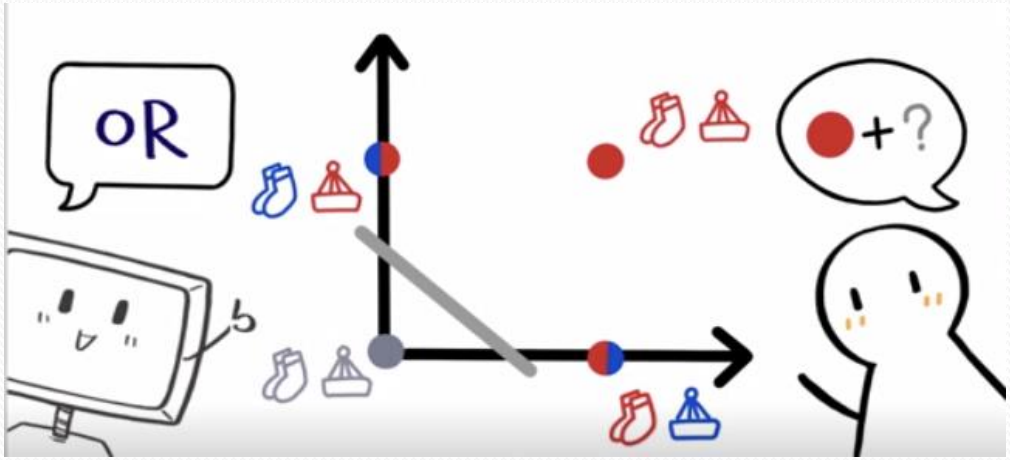
- 模型结构是线性的（直线、面、超平面）：

$$y = \mathbf{w}^T \boldsymbol{x} + w_0$$

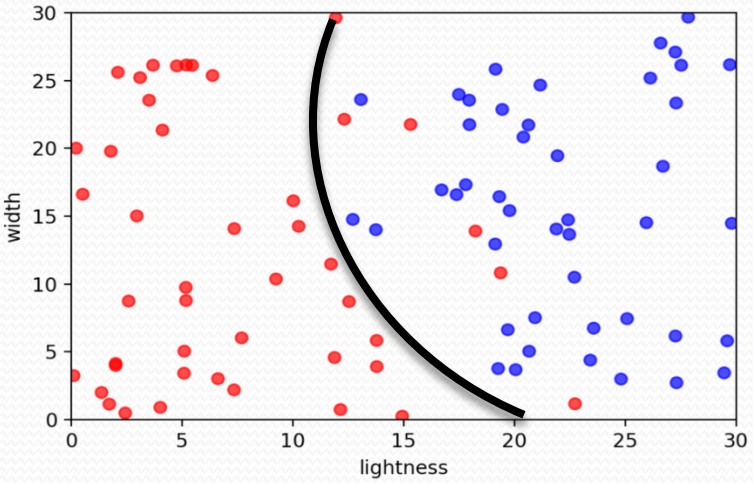
其中，$\mathbf{w}, w_0$就是模型参数。

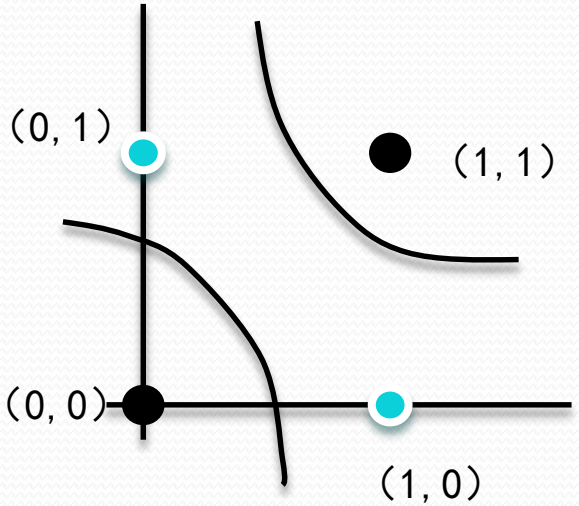
- 适用于数据是线性可分/线性表达的数据。

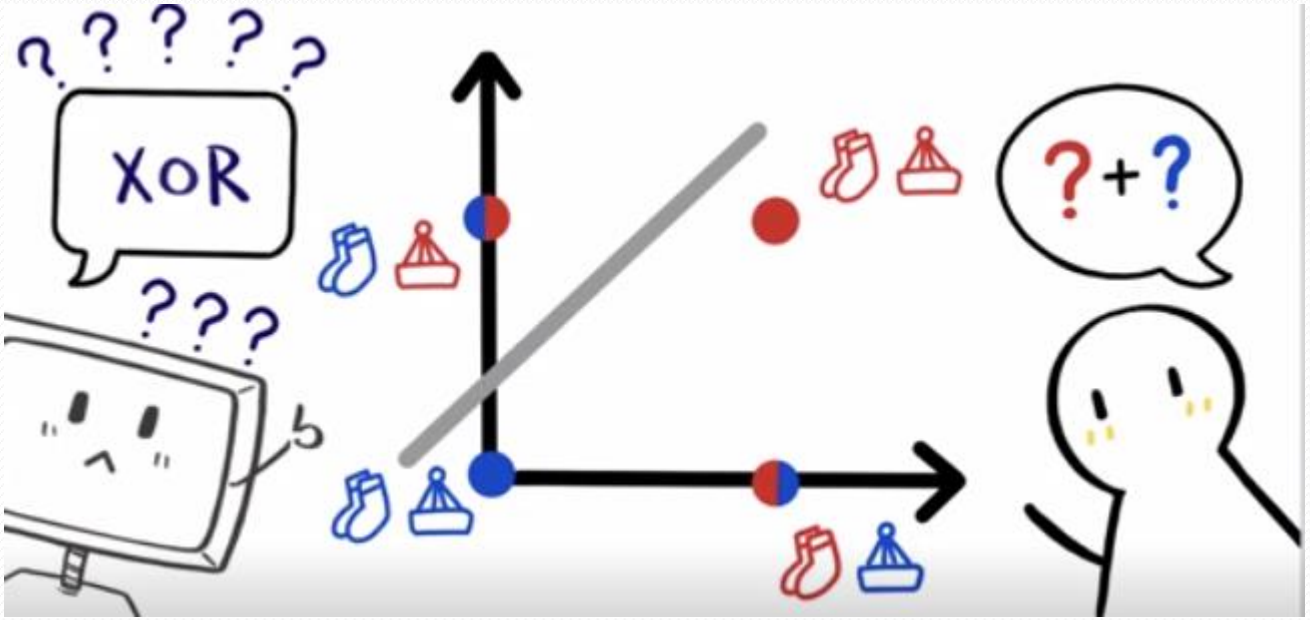

AND


OR

非线性模型

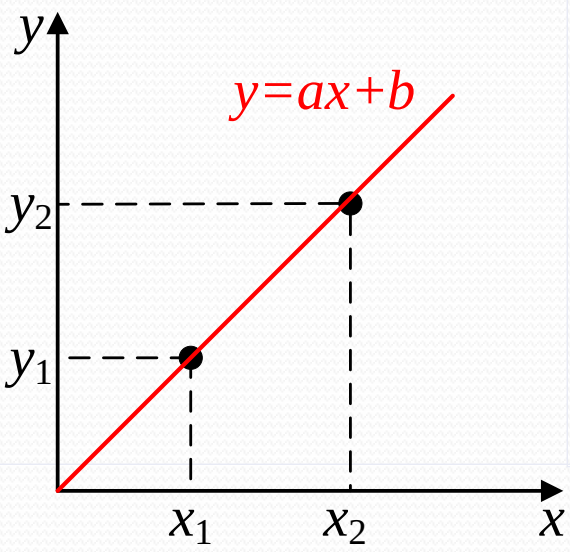


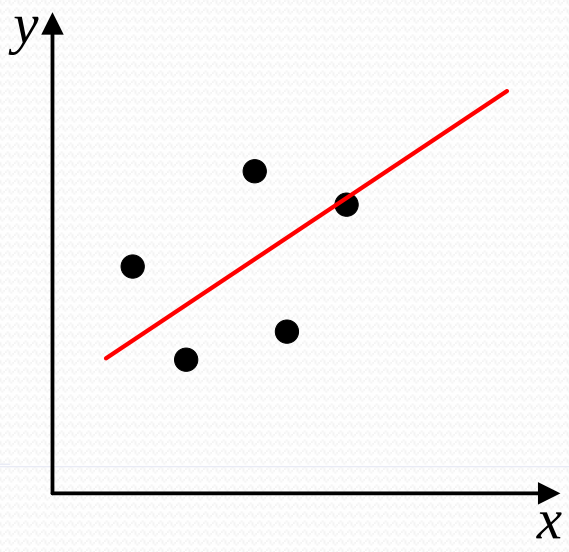


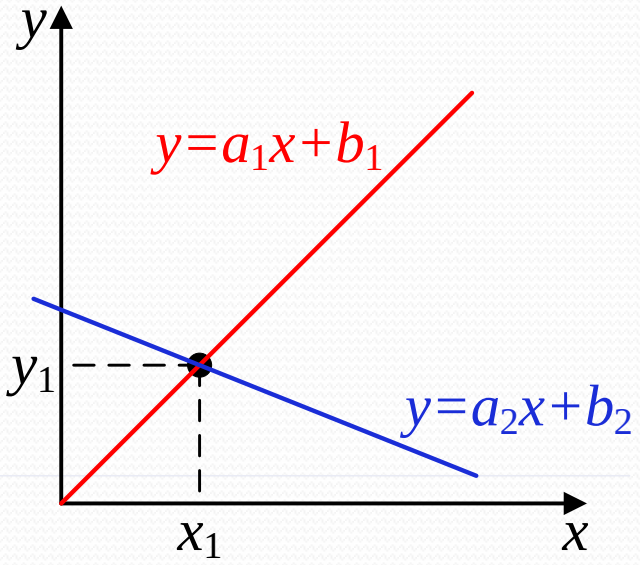

XOR
? + ?

样本量vs模型参数量

- 训练样本个数=模型参数个数（$N = M$）：参数有唯一的解。
- 训练样本个数≫模型参数个数（$N \gg M$，Over-determined）：没有准确的解。
- 训练样本个数 ≤ 模型参数个数（$N \ll M$，Under-determined）：无数个解/无解。







目标函数

- 对于over-determined的情况，需要额外添加一个标准，通过优化该标准来确定一个近似解。该标准就叫**目标函数**（Objective function），也称作代价函数（cost function）或损失函数（loss function）。
  - ✓ 目标函数以待学习的模型参数作为自变量、以训练样本作为给定量：

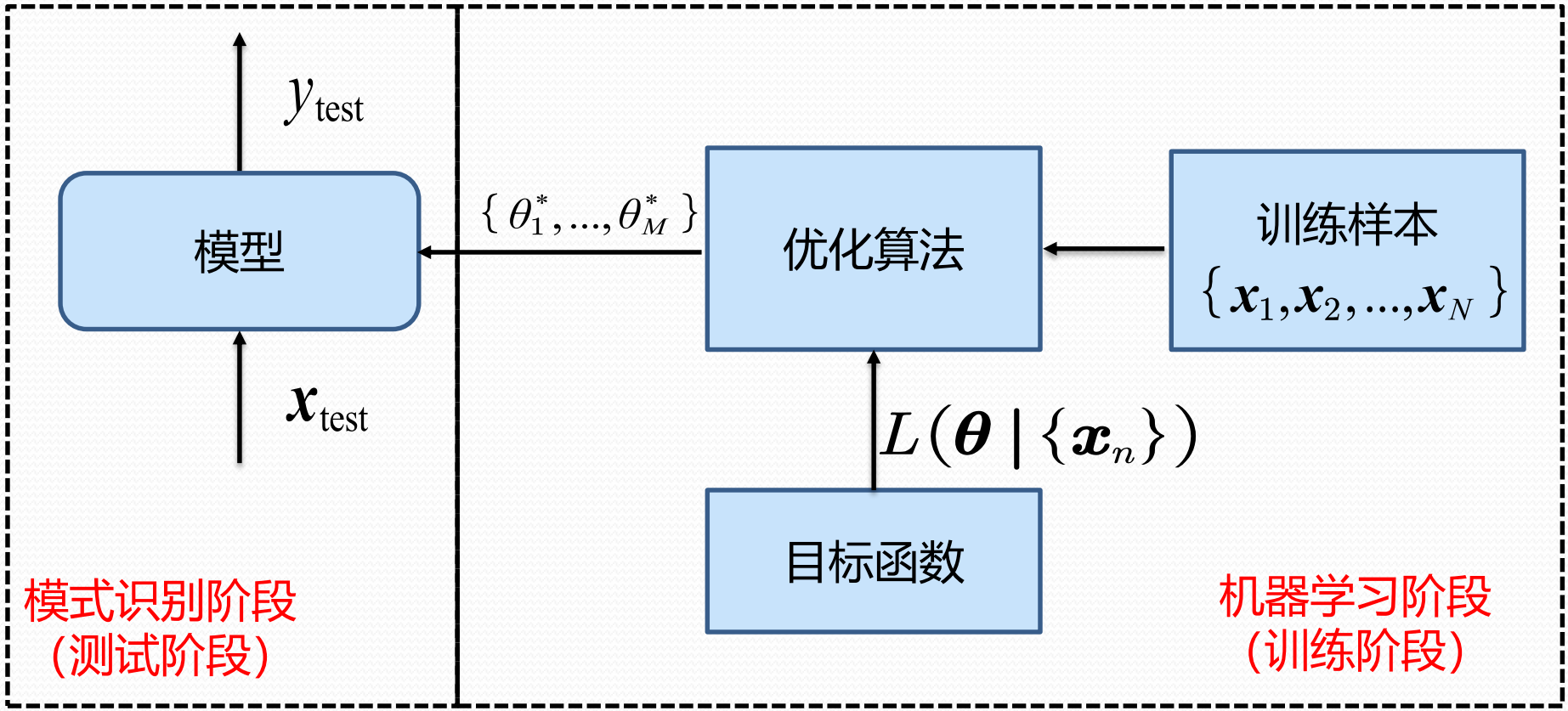
$$L\left(\boldsymbol{\theta}\middle|\{\boldsymbol{x}_i\}\right)$$

- 对于under-determined的情况，还需要在目标函数中**加入能够体现对于参数解的约束条件**，据此从无数个解中选出最优的一个解。

优化算法

- 优化算法：最小化或最大化目标函数的技术。
- 通过优化算法，最终得到模型参数$\{\theta_1, \ldots, \theta_M\}$的最优解：

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \{\boldsymbol{x}_n\})$$


y_{test}
模型
$\boldsymbol{x}_{\text{test}}$
$\{\theta_1^*,\ldots,\theta_M^*\}$
优化算法
训练样本
$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$
$L(\boldsymbol{\theta}\,|\,\{\boldsymbol{x}_n\})$
目标函数
模式识别阶段
（测试阶段）
机器学习阶段
（训练阶段）

真值（标签）& 标注

- **真值**（ground truth）：针对每个训练样本$\boldsymbol{x}_n$，其对应的真实正确的输出值，记作$\boldsymbol{t}_n$。
- **标签**（label）：对于分类任务，真值又称作标签。
- 通常，每个真值是一个向量$\boldsymbol{t}_n$。**二类分类**：真值是一个标量t_n（比如大于或小于0，正负1）。
- **标注**（labeling）：给每个训练样本标出真值的过程。目前，主要由人工完成。
- **标注的样本**（labeled samples）：有提供真值的样本。
- **未标注的样本**（unlabeled samples）：没有提供真值的样本。

监督式学习(Supervised Learning)

- 定义：训练样本及其输出真值都给定情况下的机器学习算法。
- 问题描述：

给定 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ 和 $\{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_N\}$ ，求解 $\{\theta_1, \theta_2, ..., \theta_M\}$

- 监督式学习是机器学习中最常见的学习方式。
- 通常使用最小化训练误差作为目标函数进行优化。

$$\min \sum_{n=1}^{N} ||\boldsymbol{t}_n - \boldsymbol{y}_n||_2^2$$

- **监督学习**：依靠已知所属类别的的训练样本集，按它们特征向量的分布来确定判别函数。只有在判别函数确定之后才能用它对未知的模式进行分类。

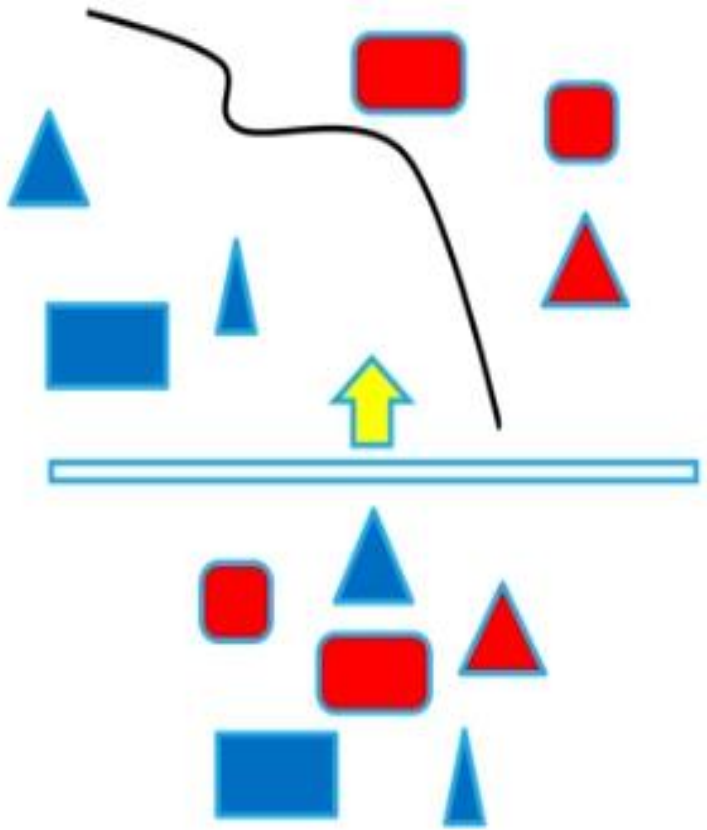
无监督式学习（Unsupervised Learning）

- 定义：只给定训练样本、没有给输出真值情况下的机器学习算法。
- 问题描述：

给定 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$，求解 $\{\theta_1, \theta_2, ..., \theta_M\}$

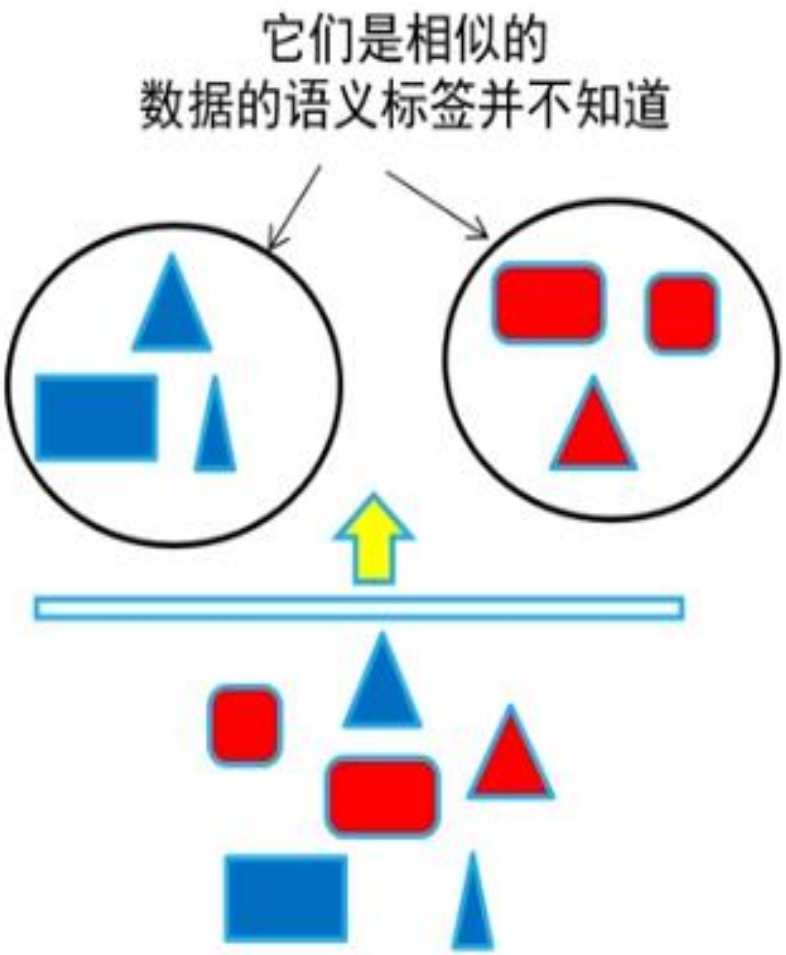
- 无监督式学习算法的难度远高于监督式算法。
- 根据训练样本之间的相似程度来进行决策。
  - ✓ 如何衡量模式（样本）之间的相似程度
  - ✓ 如何设计目标函数

- **无监督学习**：在没有先验知识的情况下，通常采用聚类方法。
- 基于**“物以类聚”**的观点，用数学方法分析各特征向量之间的距离及分散情况。
- 如果特征向量集合**聚集**若干个群，可按群间距离远近把它们划分成类。这种按各类之间的亲疏程度的划分，若事先能知道应划分成几类，则可获得更好的分类结果。



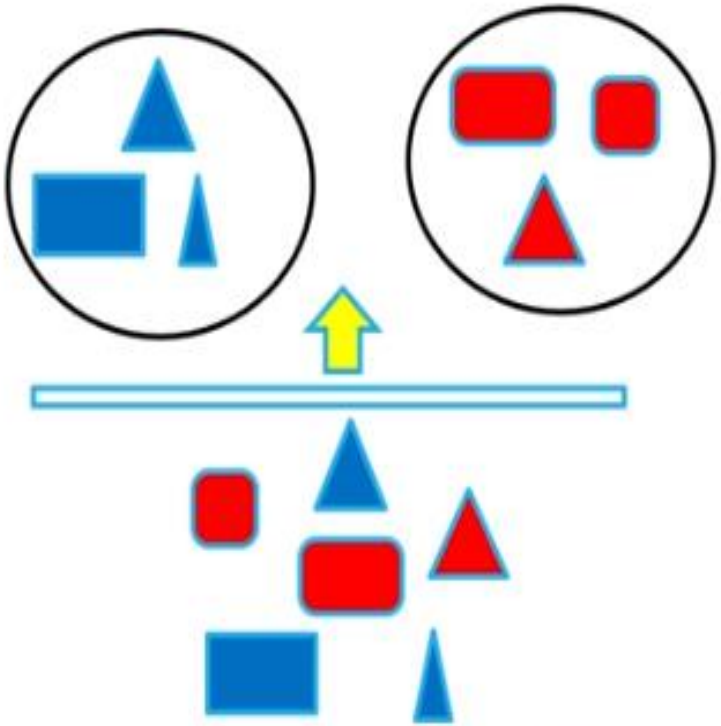
红色：汽车　　蓝色：飞机

左：监督学习

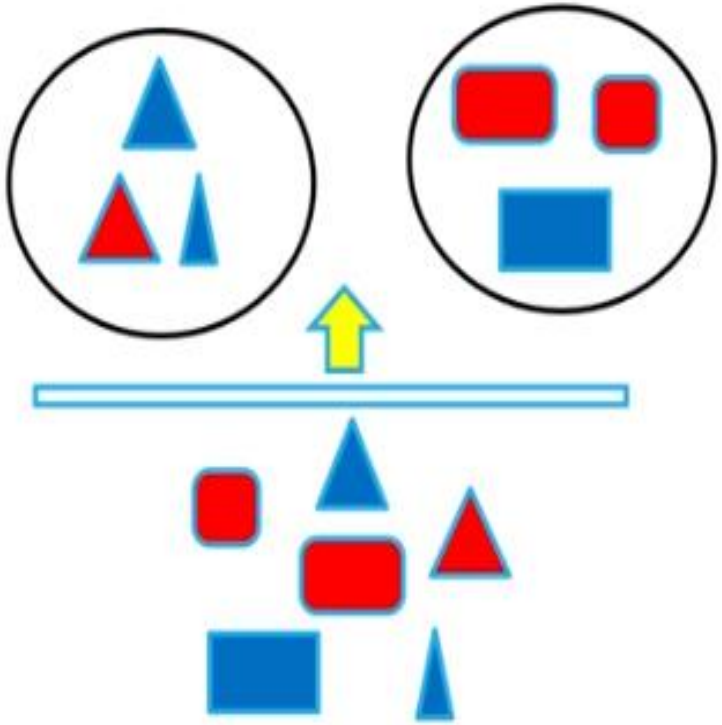


右：无监督学习

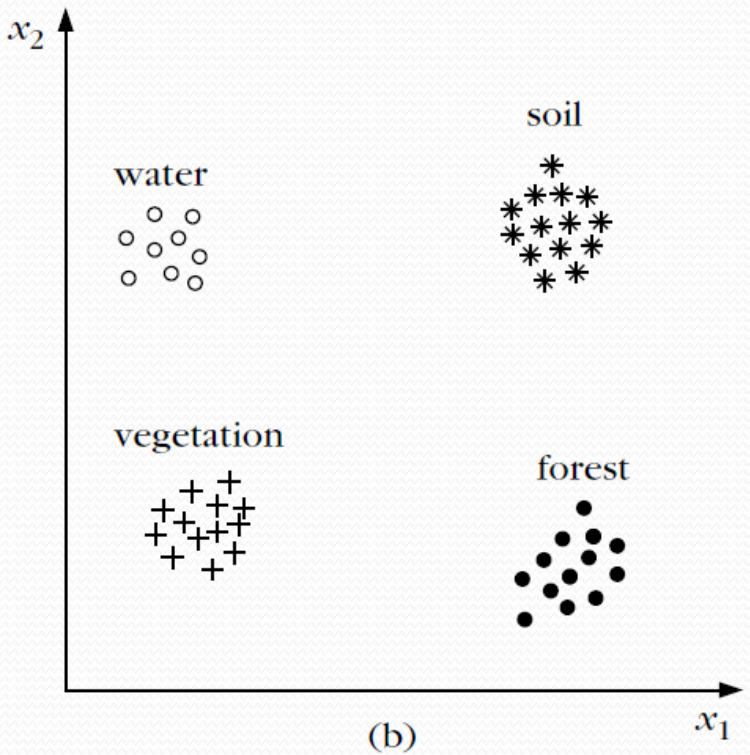
相似度函数：颜色相似

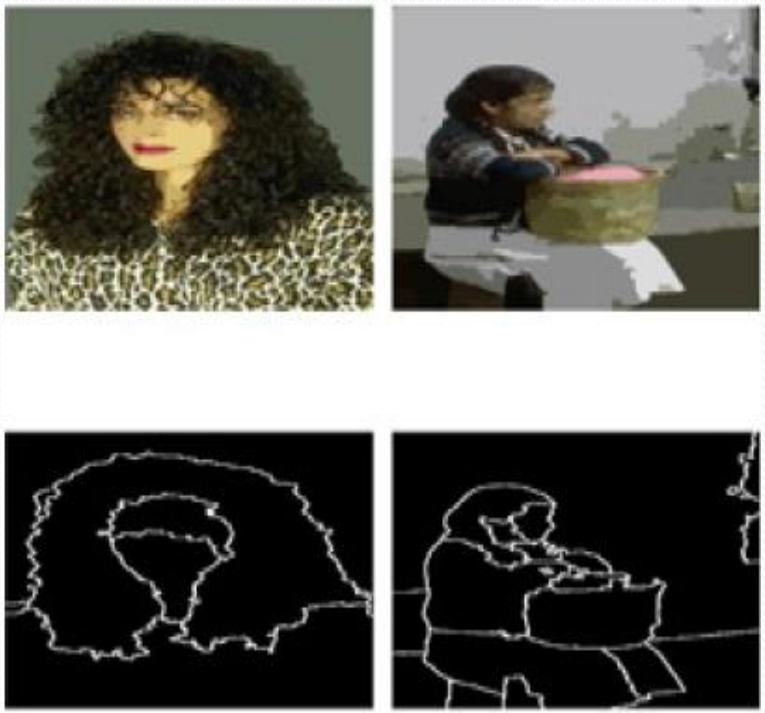


相似度函数：形状相似



无监督学习


x_2
water
soil
vegetation
forest
x_1
(b)



半监督式学习

- 定义：既有标注的训练样本、又有未标注的训练样本情况下的学习算法。
- 问题描述：

  给定 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ 和 $\{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_{N'}\}$，其中 $N > N'$，

  求解 $\{\theta_1, \theta_2, ..., \theta_M\}$

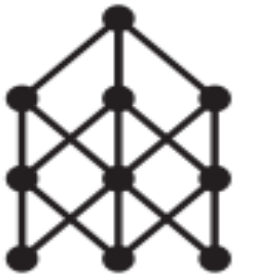
- 看做有约束条件的无监督式学习问题：标注过的训练样本用作约束条件。
- 典型应用：网络流数据。

强化学习（Reinforcement Learning）

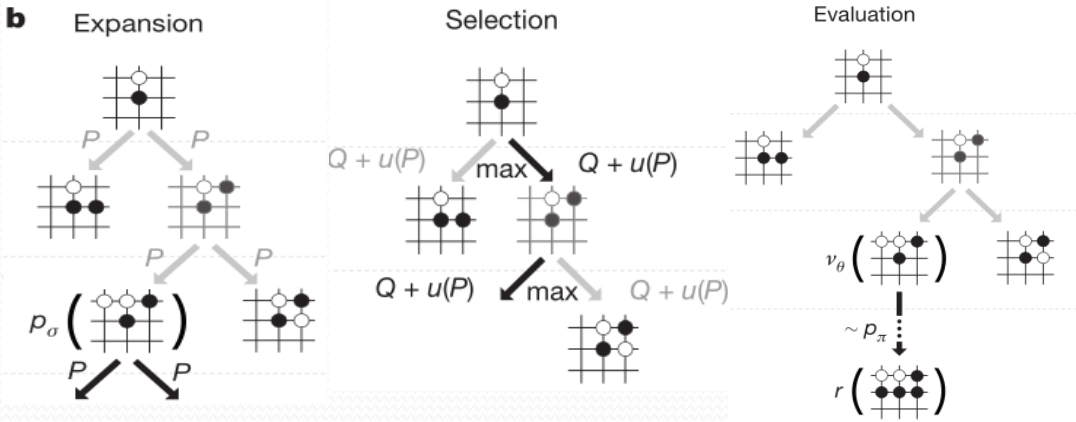
- 有些任务需要先后累积多次决策动作才能知道最终结果好坏，很难针对单次决策给出对应的真值，例如，棋类游戏。

- 强化学习：机器自行探索决策、真值滞后反馈的过程。
  - 定义从输入状态到动作决策为一个策略（policy）
  - 使用该策略进行决策探索时，给予每次决策一个奖励(reward)
  - 累积多次奖励获得回报值(return)
  - 回报的期望值作为该策略的价值函数（value function）
  - 通过最大化回报的期望值，解出策略的参数。

Value network: CNN



输入：棋局
输出：value



Given $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and $\{b_1, , \ldots, b_N\}$ solve $\{\theta_1, \ldots \theta_m\}$

- 1.7 模型的泛化能力

模型到底学的如何?

泛化能力:
学习算法对新模式的决策能力

训练集&测试集

- **训练集**(training set）：模型训练所用的样本数据。集合中的每个样本称作训练样本。
- **测试集**（test set）：测试模型性能所用的样本数据。集合中的每个样本称作测试样本。
- **测试样本**也是假设从样本真实分布中独立同分布(iid）采样得到的。
- 测试集和训练集是互斥的，但假设是同分布的。

训练误差&测试误差

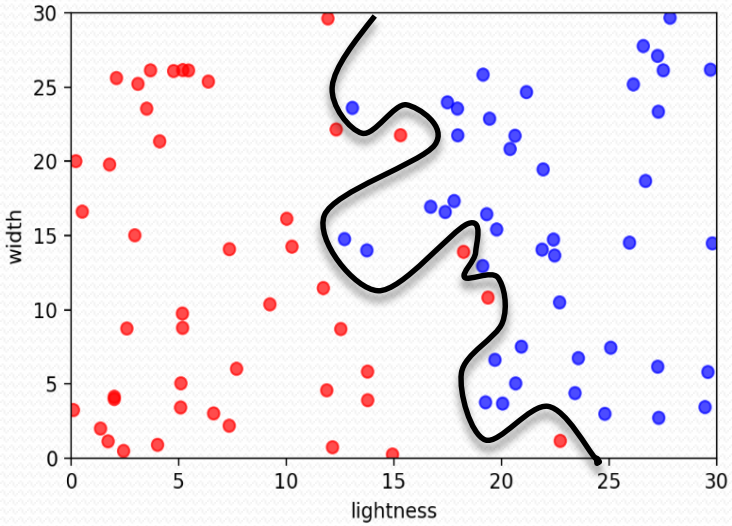
- **误差**（error）：模型（机器）给出的预测/决策输出与真值输出之间的差异。
- **训练误差**(training error）：模型在训练集上的误差。
- **测试误差**（test error）：模型在测试集上的误差。它反映了模型的泛化能力，也称作泛化误差。

泛化能力（Generalization）

- 训练样本存在的问题：
  - ✓ 训练样本稀疏：给定的训练样本数量是有限的（即有限采样），很难完整表达样本真实分布。
  - ✓ 训练样本采样过程可能不均匀：有些区域采样密一些，有些区域采样稀疏一些。
  - ✓ 一些训练样本可能带有噪声。
- **泛化能力**：训练得到的模型不仅要对训练样本具有决策能力，也要对新的（训练过程中未看见）的模式具有决策能力。

泛化能力低的表现

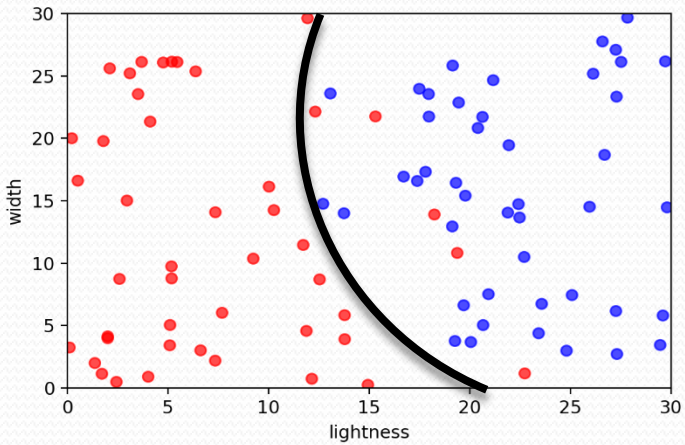
- 过拟合（over-fitting）：
  - ✓ 模型训练阶段表现很好，但是在测试阶段表现很差。
  - ✓ 模型过于拟合训练数据。



过拟合

如何提高泛化能力

- 思路：不要过度训练。
- 方法：
  - ✓ 选择复杂度适合的模型（tradeoff）：模型选择。
  - ✓ 正则化（regularization）：在目标函数中加入正则项。



多项式拟合与超参数

- 多项式拟合，求解参数**w**的最优值

$$y(x,\boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{m=1}^{M} w_m x^m$$

  - ✓ 目标函数：

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$

- 超参数（super-parameters）：
  - ✓ 超参数1：M，即多项式的阶数，决定了模型的复杂度。
  - ✓ 超参数2：N，训练样本的个数。

多项式拟合与超参数

- 多项式拟合，求解参数**w**的最优值

$$y(x,\boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{m=1}^{M} w_m x^m$$

  - ✓ 目标函数：

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$

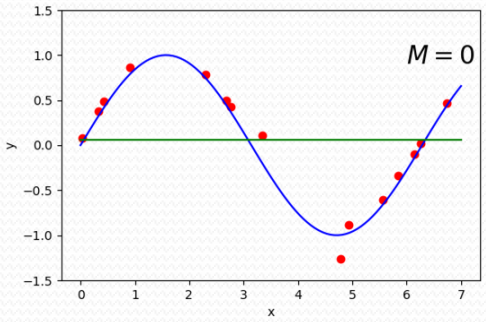
- 超参数（super-parameters）：
  - ✓ 超参数1：M，即多项式的阶数，决定了模型的复杂度。
  - ✓ 超参数2：N，训练样本的个数。

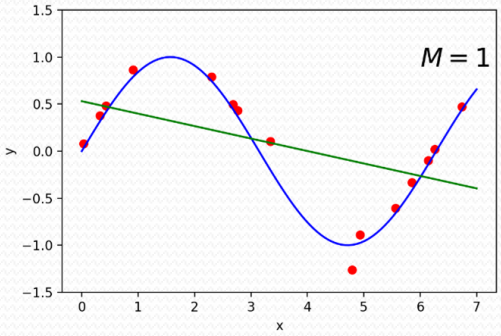
超参数

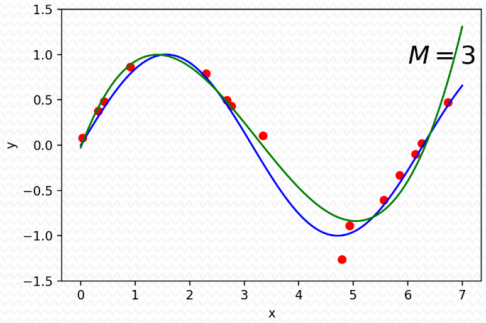
- 超参数（super-parameters）：
  - ✓ 在机器学习的上下文中，超参数是在开始学习过程之前设置值的参数，而不是通过训练得到的参数数据。
  - ✓ 通常情况下，需要对超参数进行优化。
  - ✓ 给模型选择一组最优超参数，以提高学习的性能和效果。

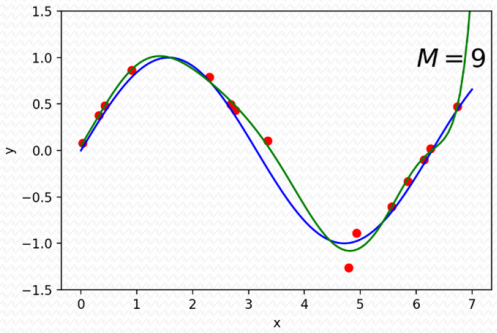
提高泛化能力：模型选择

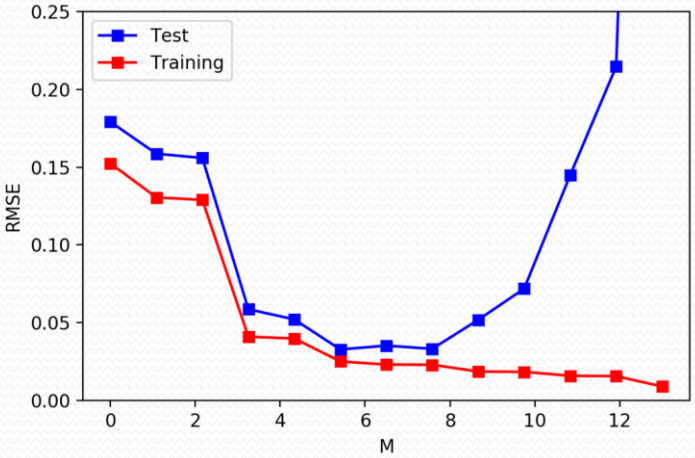
- 固定训练样本个数： $N = 15$
- 模型选择：选择合适的的多项式阶数M











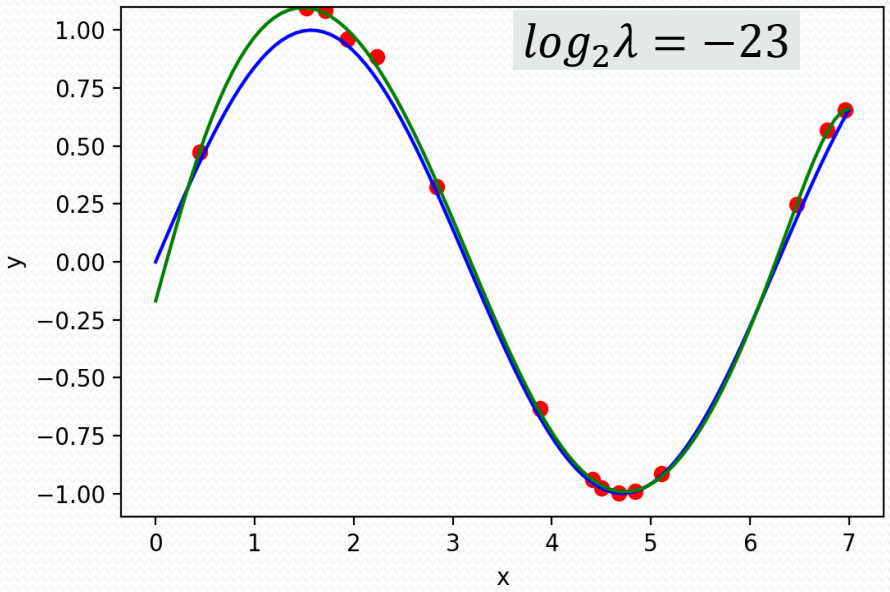
提高泛化能力：正则化

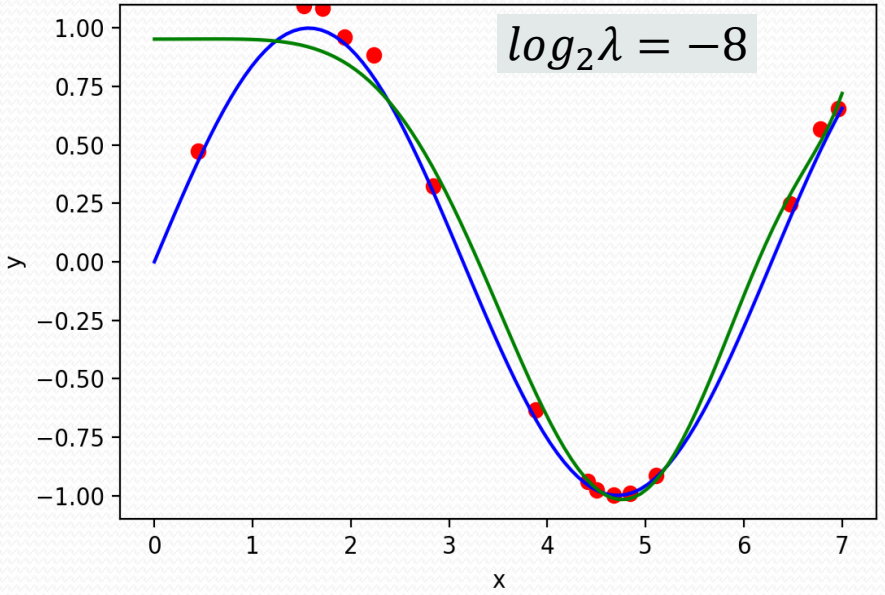
- 固定多项式的阶数$M = 9$及训练样本个数$N = 15$
- 在目标函数中加入关于参数的正则项
- 超参数：正则系数λ

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||_2^2$$

提高泛化能力：正则化

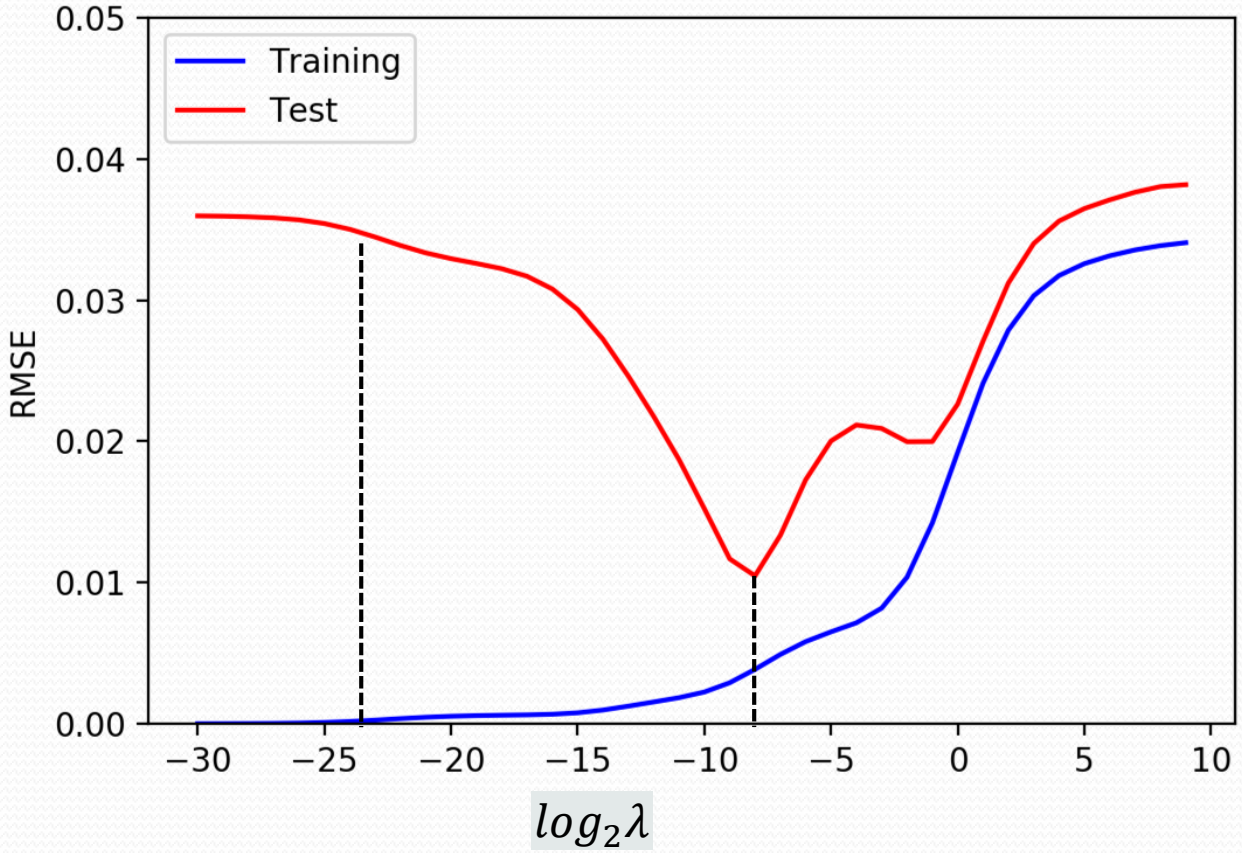
- 通过调节正则系数，降低过拟合的程度





提高泛化能力：正则化

- 通过调节正则系数，降低过拟合的程度



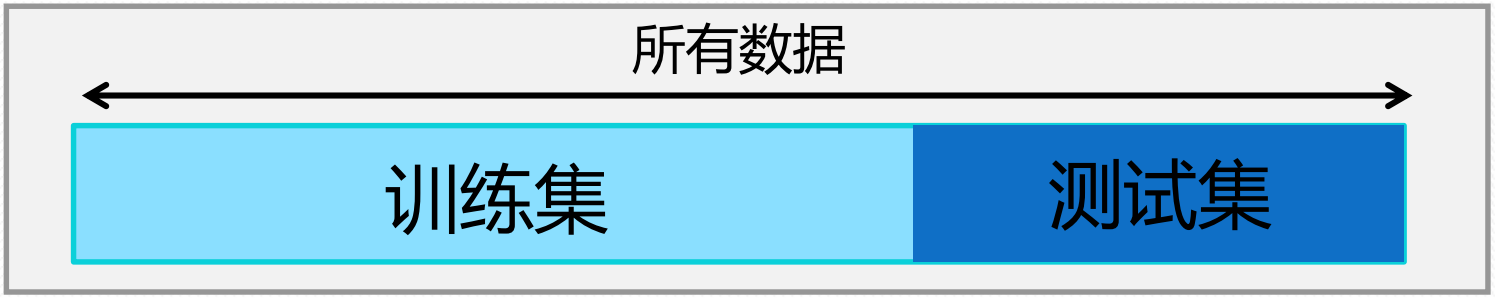
调参

- 几乎每个机器学习算法都有超参数。
- 涉及到泛化能力、可调整的超参数主要有： M， λ。
- 如何选取合适的超参数?
  - ✓ 需要依据泛化误差，但又不能基于测试集。
  - ✓ 所以，从训练集中划分出一个验证集（validation set），基于验证集调整选择超参数。

- 1.8 评估方法与性能指标

留出法（Hold-out）

- **随机划分**：将数据集随机分为两组：训练集和测试集。利用训练集训练模型，然后利用测试集评估模型的量化指标。
- **取统计值**：为了克服单次随机划分带来的偏差，将上述随机划分进行若干次，取量化指标的平均值（以及方差、最大值等）作为最终的性能量化评估结果。

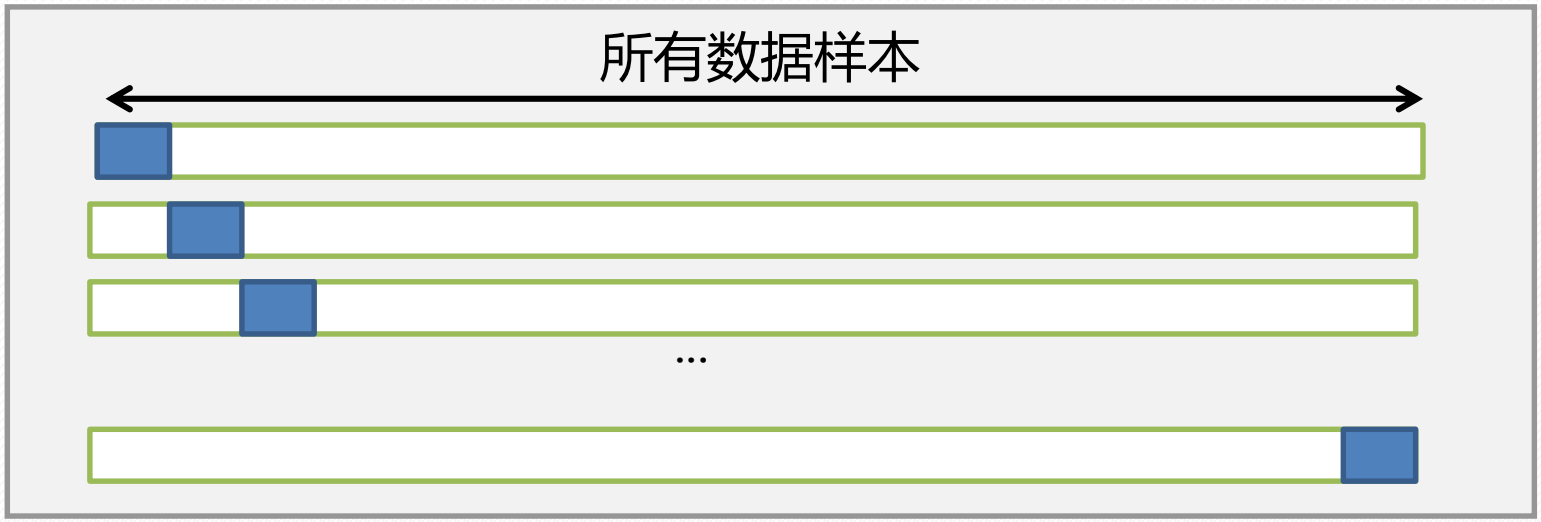


K折交叉验证（K-Folds Cross Validation）

- 将数据集分割成K个子集，从其中选取单个子集作为测试集，其他$K-1$个子集作为训练集。
- 交叉验证重复K次，使得每个子集都被测试一次；将K次的评估值取平均，作为最终的量化评估结果。

留一验证（leave-one-out cross-validation）

- 每次只取数据集中的一个样本做测试集，剩余的做训练集。
- 每个样本测试一次，取所有评估值的平均值作为最终评估结果。
- 等同于K折交叉验证，K为数据集样本总数。



- 对于回归任务：测试误差。
- 二类分类：真阳性（TP），假阳性（FP），真阴性（TN），假阴性（FN）
- 多类分类：依次以单个类作为正类，其余为负类。

	预测为正/阳性	预测为负/阴性
真值为正/阳性	True Positive(TP)	False Negative(FN)
真值为负/阴性	False Positive(FP)	True Negative(TN)

- 准确度（Accuracy）：将阳性和阴性综合起来度量识别正确的程度。
  - ✓ 如果阳性和阴性样本数量比例失衡，该指标很难度量识别性能。
  - ✓ 例子：阳性/阴性=5/95，模型把所有样本都判断为阴性。

	预测为正/阳性	预测为负/阴性
真值为正/阳性	True Positive(TP)	False Negative(FN)
真值为负/阴性	False Positive(FP)	True Negative(TN)
$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$		

- 精度（Precision）：预测为阳性样本的准确程度。在信息检索领域，也称作查准率。
- 召回率（Recall）：也称作敏感度（sensitivity），全部阳性样本中被预测为阳性的比例。在信息检索领域也称作查全率。

	预测为正/阳性	预测为负/阴性	指标
真值为正/阳性	True Positive(TP)	False Negative(FN)	$\text{Recall}=\frac{TP}{TP+FN}$
真值为负/阴性	False Positive(FP)	True Negative(TN)	$\text{Specificity}=\frac{TN}{TN+FP}$
$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$	$\text{Precision}=\frac{TP}{TP+FP}$		$\text{F1 Score}=\frac{2*Recall*Precision}{Recal+Precision}$

- 精度高、同时召回率也高，说明模型性能越好。
- 但是，在有些情况下，精度和召回率是矛盾的。
  - ✓ 例子：阳性/阴性=50/50，模型只识别出来一个样本为阳性，其余被识别为阴性。此时，precision=1/(1+0)=100%，recall=1/(1+49)=2%.

	预测为正/阳性	预测为负/阴性	指标
真值为正/阳性	True Positive(TP)	False Negative(FN)	$\text{Recall}=\frac{TP}{TP+FN}$
真值为负/阴性	False Positive(FP)	True Negative(TN)	$\text{Specificity}=\frac{TN}{TN+FP}$
$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$	$\text{Precision}=\frac{TP}{TP+FP}$		$\text{F1 Score}=\frac{2*Recall*Precision}{Recal+Precision}$

F-Score

- 通过加权平均，综合precision和recall：

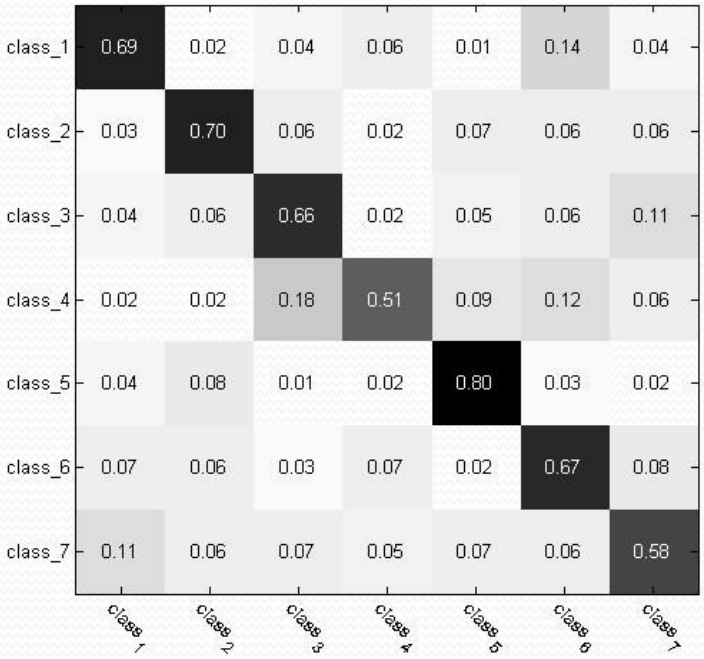
$$F = \frac{(a^2+1) \times precision \times recall}{a^2 \times precision + recall}$$

- 设置$a = 1$，得到F1-score：

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

定义

- 矩阵的列代表预测值，行代表真值。
- 矩阵中每个元素的值是根据每个测试样本的预测值和真值得到的计数统计值。
- 对角线元素的值越大，表示模型性能越好。



样本序号	真值	分类器输出结果
1	C1	C2
2	C1	C1
3	C1	C1
4	C1	C1
5	C1	C1
6	C1	C1
7	C1	C1
8	C1	C1
9	C1	C1
10	C1	C1
11	C2	C1
12	C2	C3
13	C2	C1
14	C2	C3
15	C2	C3
16	C2	C2
17	C2	C3
18	C2	C3
19	C2	C3
20	C2	C2

样本序号	真值	分类器输出结果
21	C3	C1
22	C3	C1
23	C3	C3
24	C3	C3
25	C3	C3
26	C3	C3
27	C3	C3
28	C3	C1
29	C3	C3
30	C3	C3
31	C4	C1
32	C4	C2
33	C4	C2
34	C4	C1
35	C4	C3
36	C4	C2
37	C4	C1
38	C4	C2
39	C4	C3
40	C4	C2

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

统计分类器分类结果正确的个数为 18 个

Accuracy=分类正确个数/样本总个数=18/40=45%

统计各个类的分类情况：

C1：分为 C1：9 个，分为 C2 ：1 个，分为 C3 ：0 个 ，分为 C4： 0 个

C2：分为 C1：2 个，分为 C2 ：2 个，分为 C3 ：6 个 ，分为 C4： 0 个

C3：分为 C1：3 个，分为 C2 ：0 个，分为 C3 ：7 个 ，分为 C4： 0 个

C4：分为 C1：3 个，分为 C2 ：5 个，分为 C3 ：2 个 ，分为 C4： 0 个

	Predict				
actual		C1	C2	C3	C4
	C1	9	1	0	0
	C2	2	2	6	0
	C3	3	0	7	0
	C4	3	5	2	0

P(C1)=9/(9+2+3+3)=52.94% R(C1)=9/(9+1+0+0)=90%

P(C2)=2/(1+2+0+5)=25% R(C2)=2/(2+2+6+0)=20%

P(C3)=7/(0+6+7+2)=46.67% R(C3)=7/(3+0+7+0)=70%

P(C4)=0 R(C4)=0

F1=2*P*R/(P+R)

F1(C1)=66.67% F1(C2)=22.22% F1(C3)=56.00% F1(C4)=0

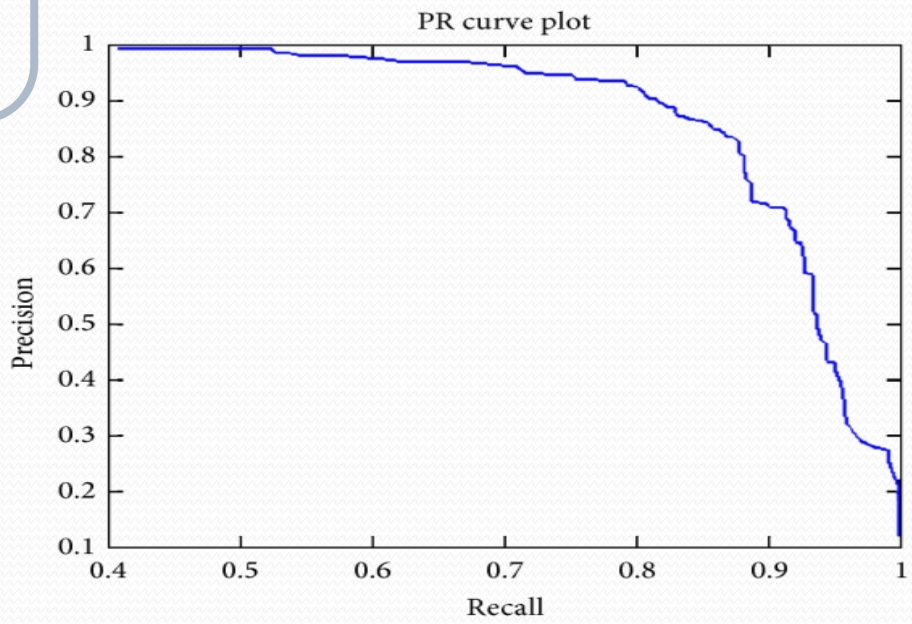
	Predict				
actual		C1	C2	C3	C4
	C1	9	1	0	0
	C2	2	2	6	0
	C3	3	0	7	0
	C4	3	5	2	0

曲线度量

- 上述性能指标都是基于分类器输出，是一个确定的离散标签。
  - ✓ 因此，在指标空间，性能评估只是一个点。
- 有些分类器在输出端会以数值形式表达：概率值、score值。
  - ✓ 因此，可以设置若干个关于输出值的阈值，不同的阈值可以代表不同的应用任务，得到多个评估值，从而可以在指标空间画出一条曲线，从而得到评估指标的期望表征。

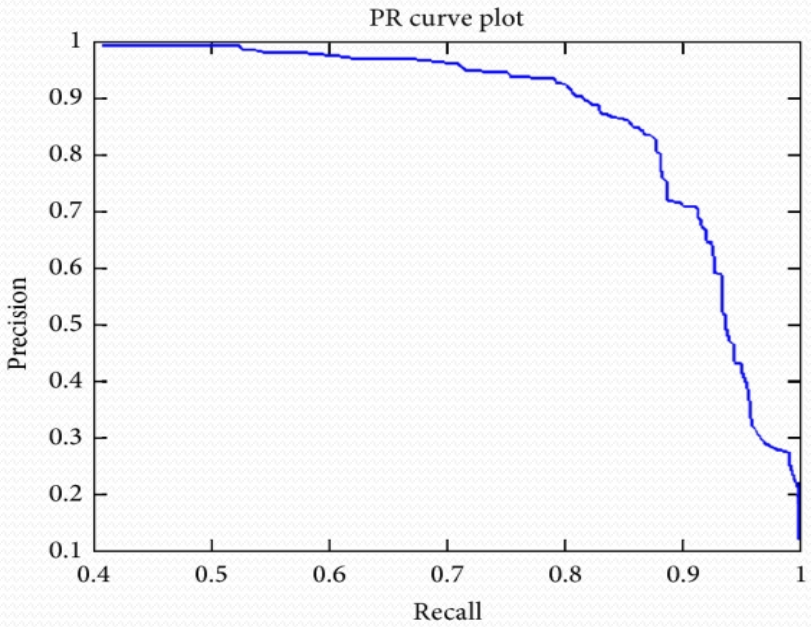
PR曲线定义

- 横轴：召回率（recall）
- 纵轴：精度（precision）
- 理想性能：右上角（1,1）处。
- PR（Precision-Recall Curve）曲线越往右上凸，说明模型的性能越好。



PR曲线绘制方法

- 根据模型的预测数值，对样本进行从高到低排序，排在前面的样本是正例的可能性更高。
- 按此顺序逐个样本作为正例进行预测（或设置阈值截断正例和负例），则每次可以计算一个召回率和精度。
- 将这些值连成（拟合）一条曲线。



ROC曲线（Receiver-operating-characteristic curve）

- 横轴：False positive rate (FPR)，度量所有阴性样本中被错误识别为阳性的比率。FPR=1-specificity。

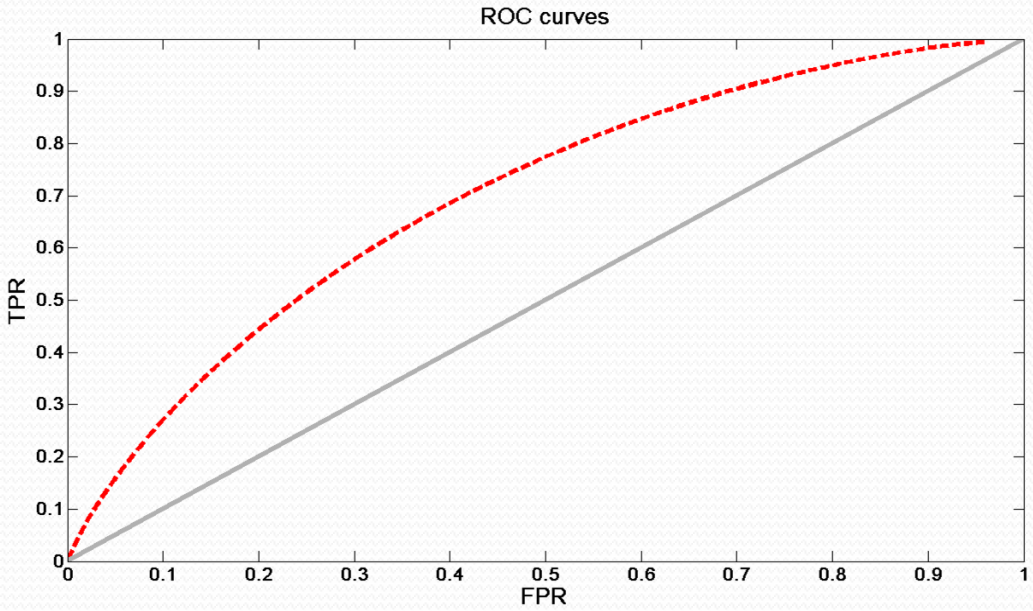
$$FPR = \frac{FP}{FP + TN}$$

- 纵轴：True positive rate (TPR)，即recall。度量所有阳性样本被识别为阳性的比例。

真值为负/阴性	False Positive(FP)	True Negative(TN)	Specificity= $\frac{TN}{TN+FP}$

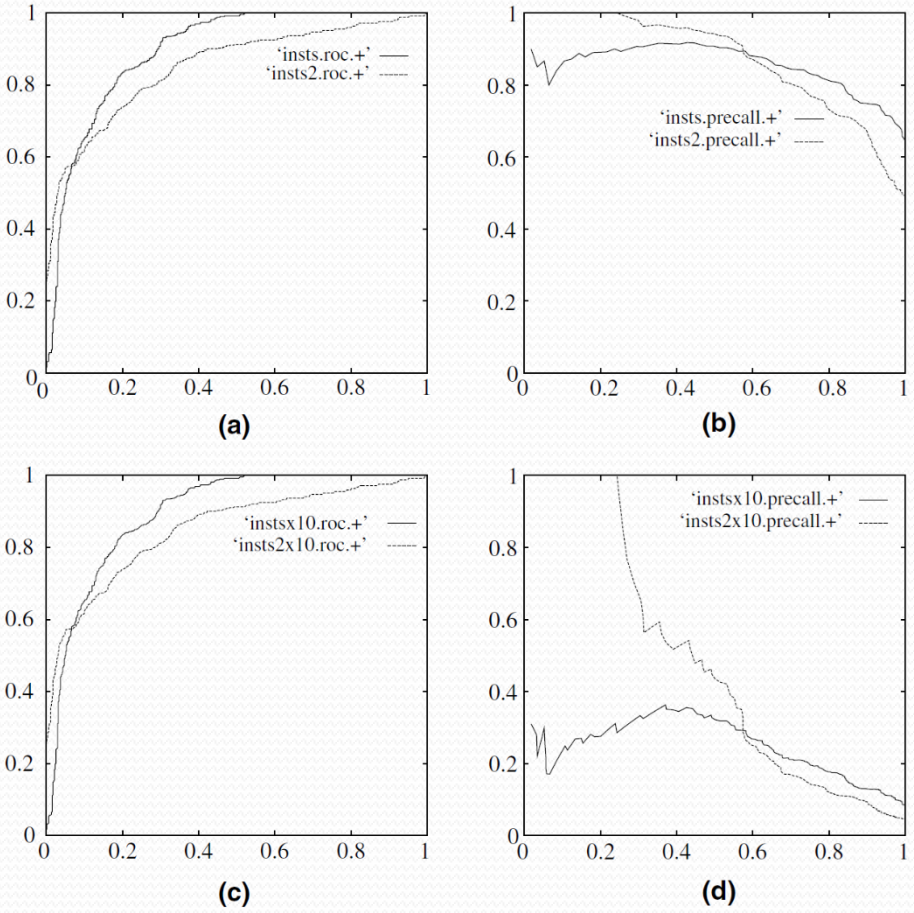
ROC曲线

- 理想性能：左上角(0,1)处。
- ROC曲线越往左上凸，说明模型的性能越好。
- 对角线：随机识别的ROC曲线。
- 绘制方法：与PR曲线相似。



PR与ROC曲线

- ROC曲线：对于各类别之间样本分布比例不敏感，因为FPR和TPR各自只跟真值为负或真值为正的样本相关。
- PR曲线：对于各类别样本分布比例敏感，因为precision同时和真值正负的样本都相关。

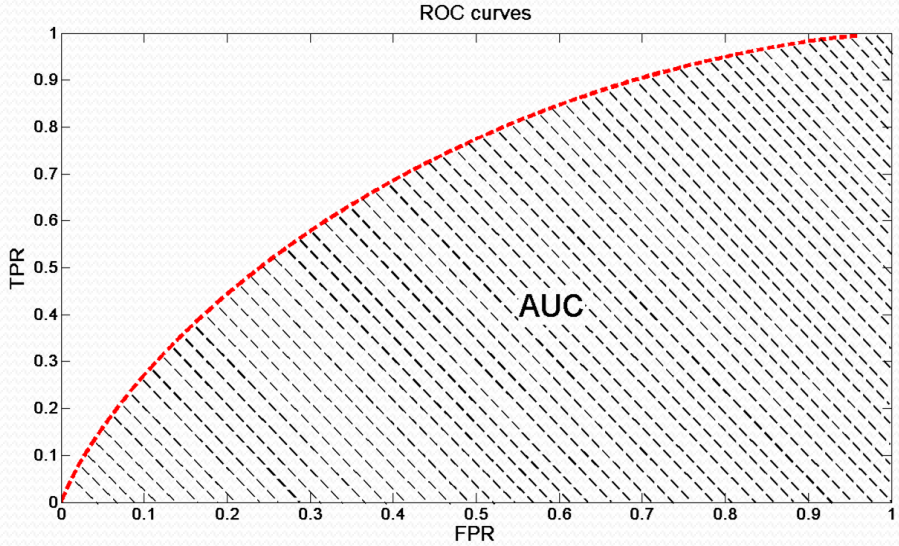


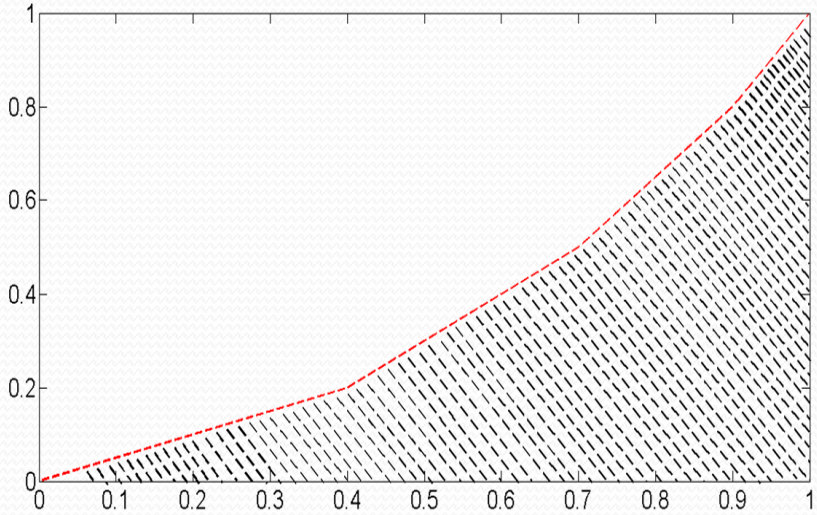
ROC曲线　　PR曲线

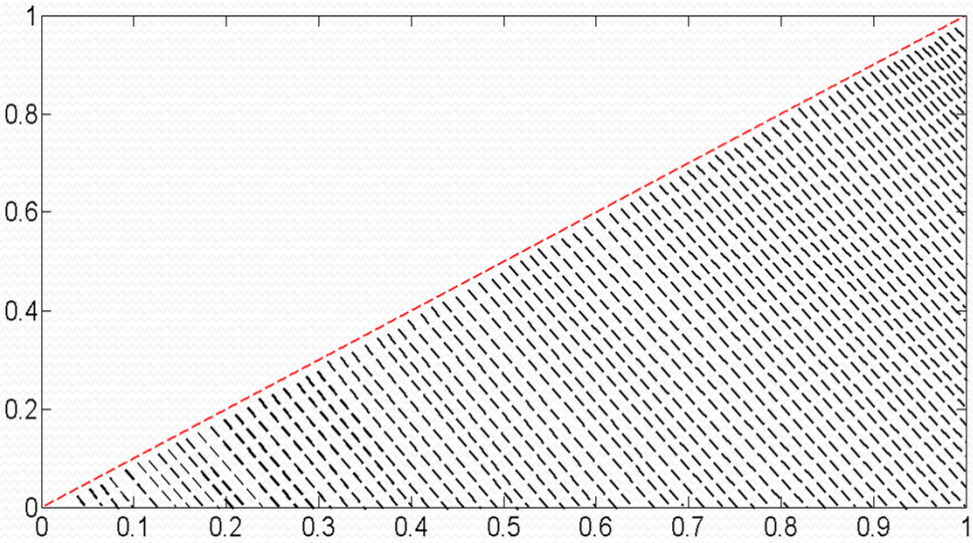
AUC曲线定义与特性

- 曲线下方面积(Area under the Curve, AUC)：将曲线度量所表达的信息浓缩到一个标量表达。
  - ✓ AUC = 1：是完美分类器，
  - ✓ 0.5 < AUC < 1：优于随机猜测。这个模型妥善设定阈值的话，能有预测价值。
  - ✓ AUC = 0.5：跟随机猜测一样，模型没有预测价值。
  - ✓ AUC < 0.5：比随机猜测还差。

AUC曲线


ROC curves
TPR
FPR
AUC
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9


0
0.2
0.4
0.6
0.8
1
0.1
0.3
0.5
0.7
0.9