

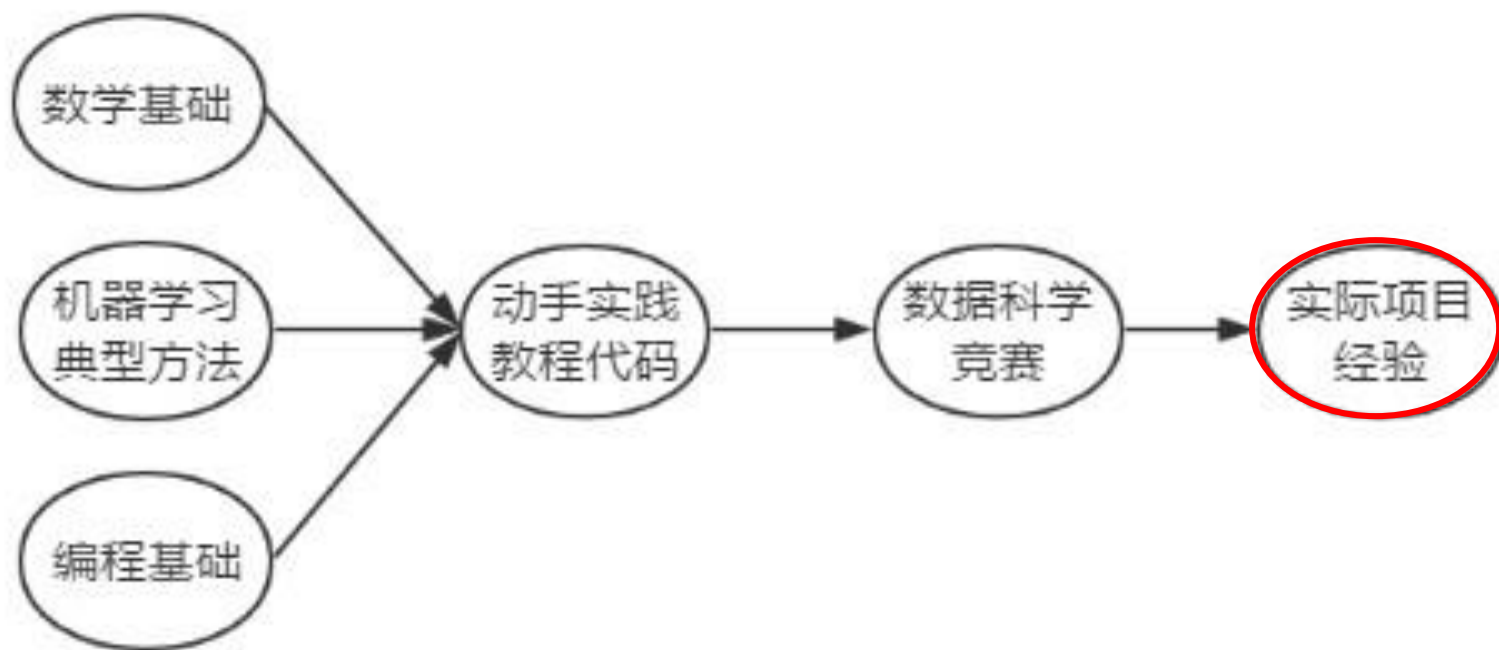
# 数据科学与数学基础知识

---

七月 算法 龙老师  
2016年6月25日

# 机器学习路线图

---



# 引言

---

- 这门课曾经考虑叫大数据挖掘
  - 海量数据，偏实践
- 机器学习理论有专门的课程
  - 《4月机器学习算法班》
- 机器学习数学基础有专门的课程
  - 《机器学习中的数学班》
- Python编程正在考虑开专门的课程



# 主要内容

---

## □ 机器学习基础

- 机器学习的分类与一般思路

## □ 微积分基础

- 泰勒公式、导数与梯度

## □ 概率与统计基础

- 概率公式、常见分布、常见统计量

## □ 线性代数基础

- 矩阵乘法的几何意义



# 自我介绍

---

## □ 龙老师

- 数年互联网经验，专注ML/DM，[博客](#)：CSDN
- 现在某厂负责海量数据下的用户画像和智能营销相关项目

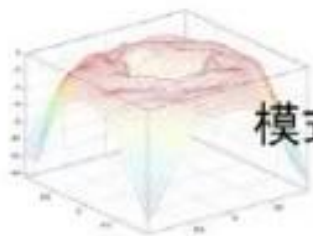
## □ 项目经验：

- 用户画像、智能营销策略
- 网络安全机器学习
- 自然语言处理相关项目



# 机器学习

---



模式识别

计算机视觉



数据挖掘



机器学习

语音识别



统计学习

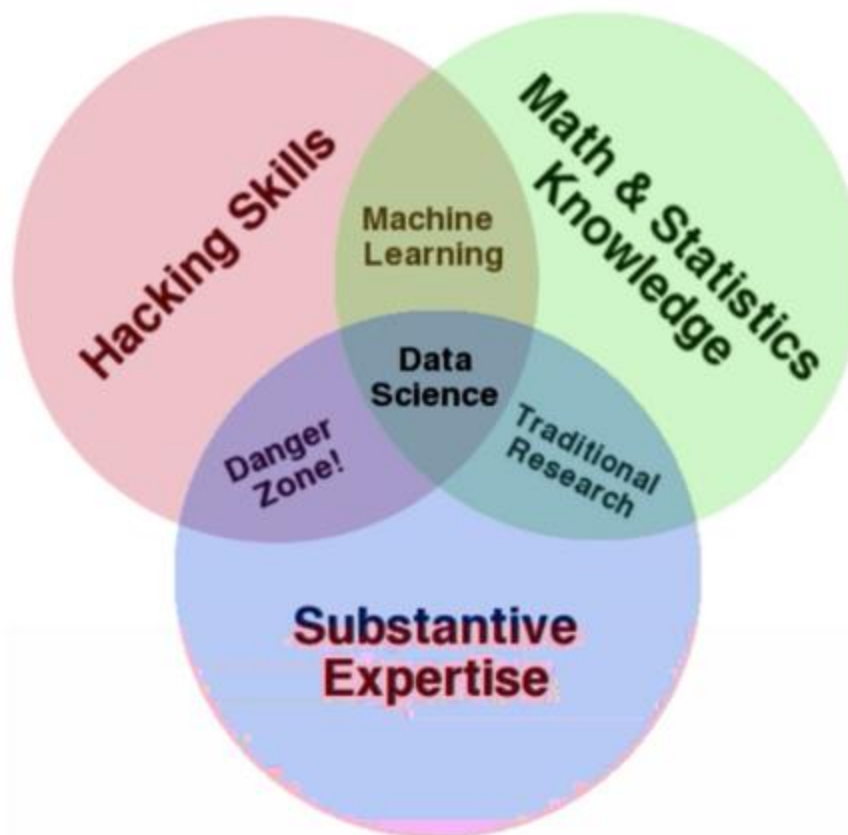


自然语言处理



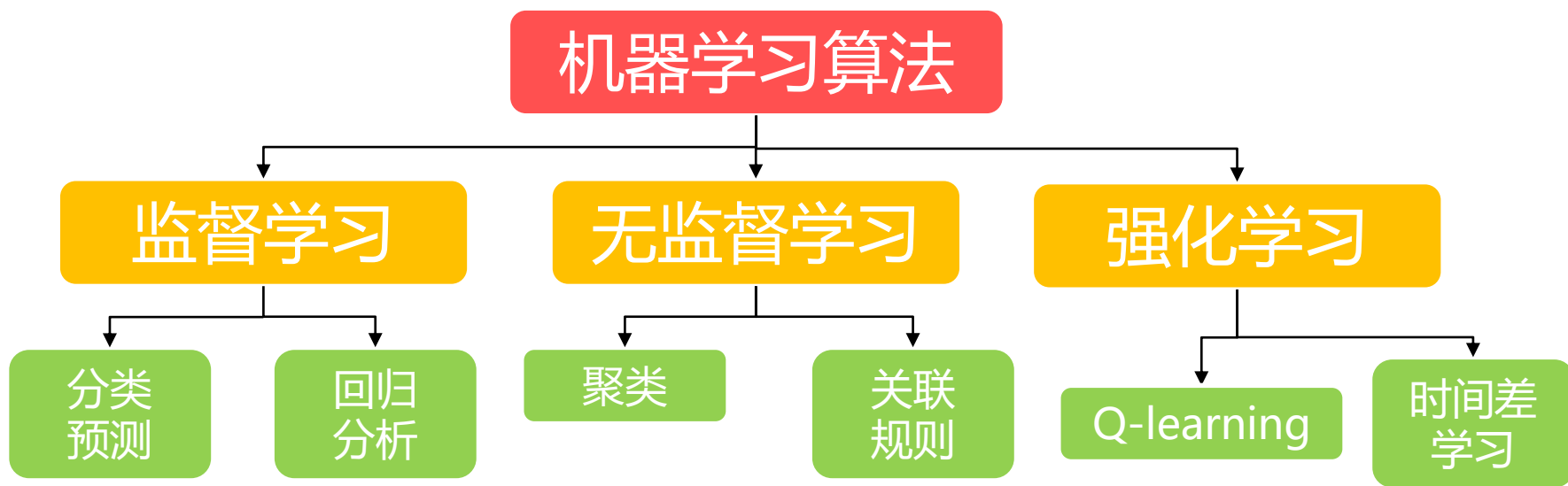
# 机器学习

---



# 机器学习分类

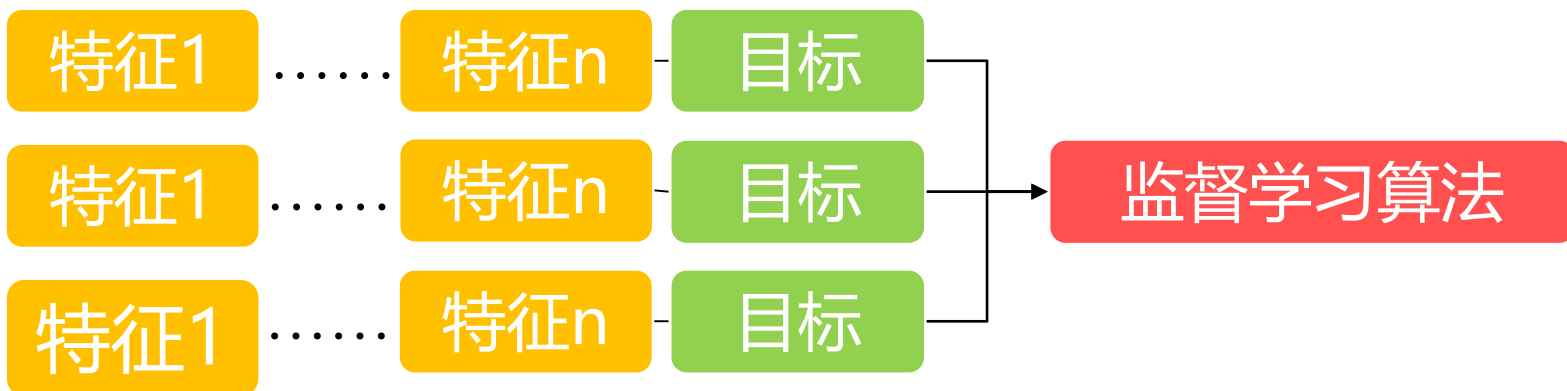
- 监督学习：例如用户点击/购买预测、房价预测
- 无监督学习：例如邮件/新闻聚类
- 强化学习：例如动态系统以及机器人控制





# 监督学习

训练集

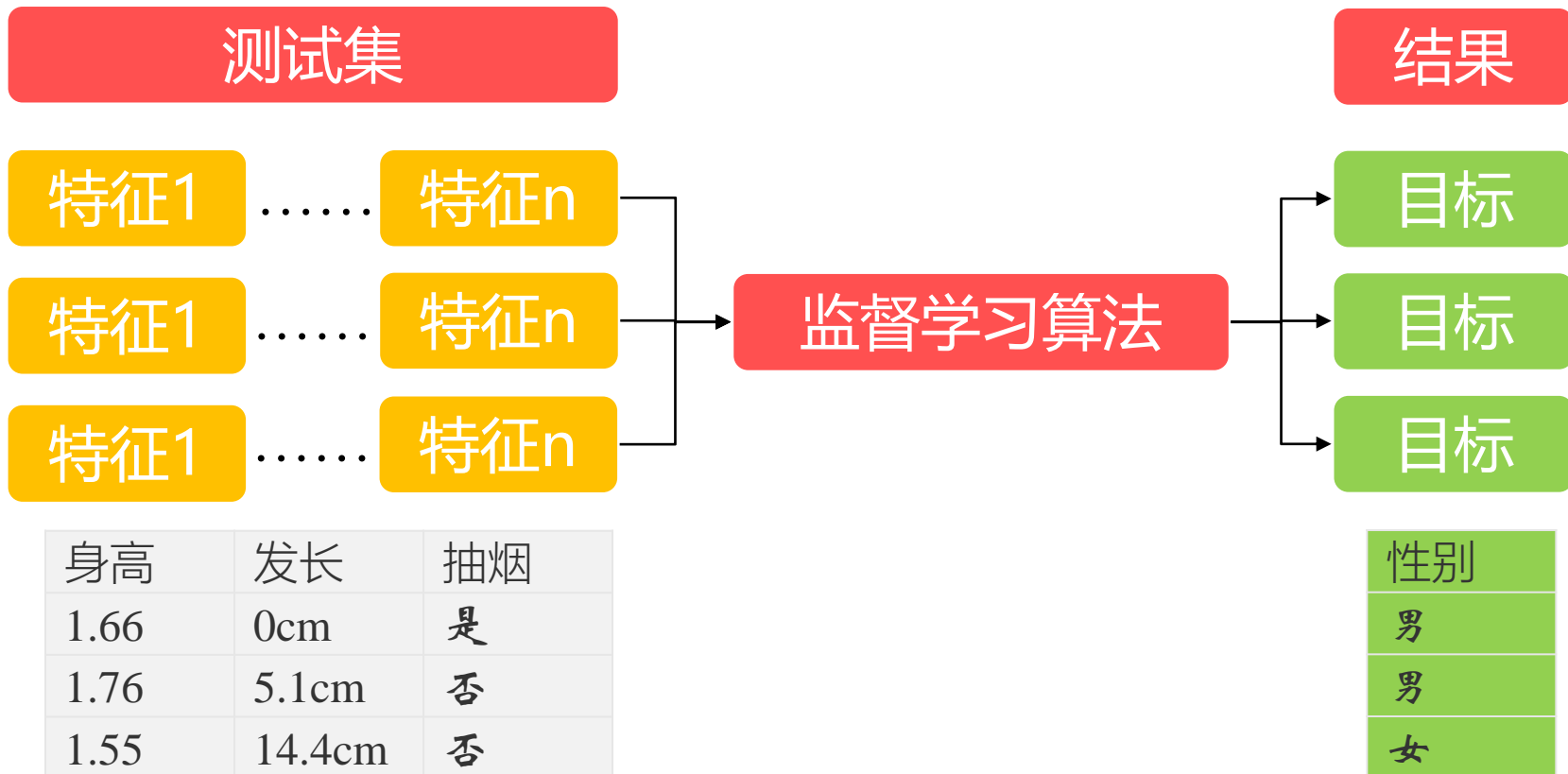


身高	发长	抽烟	性别
1.88	1.4cm	是	男
1.66	15.3cm	否	女
1.78	22.6cm	否	女

监督学习算法：训练/学习



# 监督学习



监督学习算法：预测



# 无监督学习

训练集

特征1

.....

特征n

特征1

.....

特征n

特征1

.....

特征n

领型

袖长

材质

圆领

长袖

全棉

圆领

短袖

全棉

V领

短袖

纤维

立领

短袖

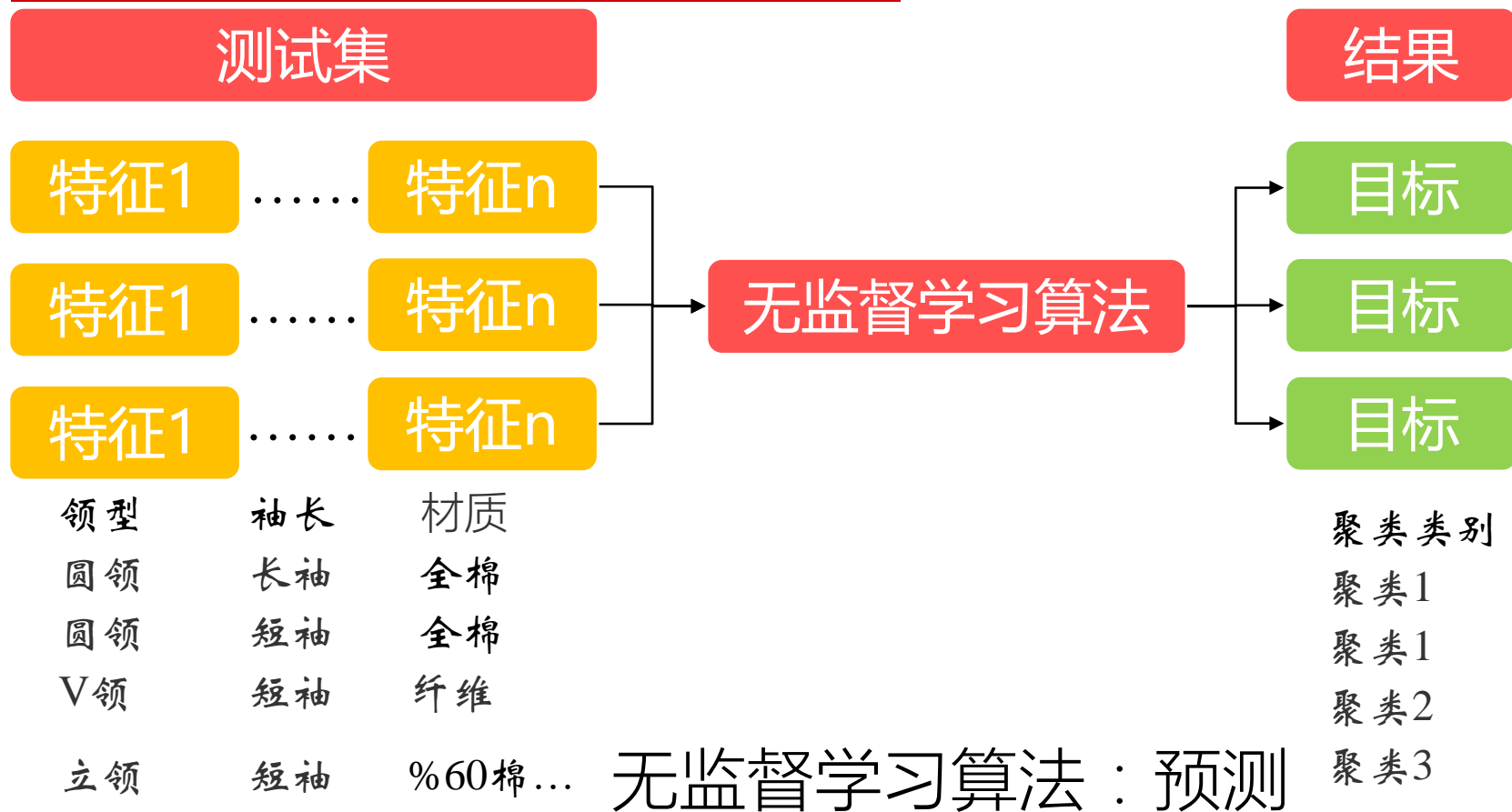
%60棉...

无监督学习算法

无监督学习算法：训练/学习

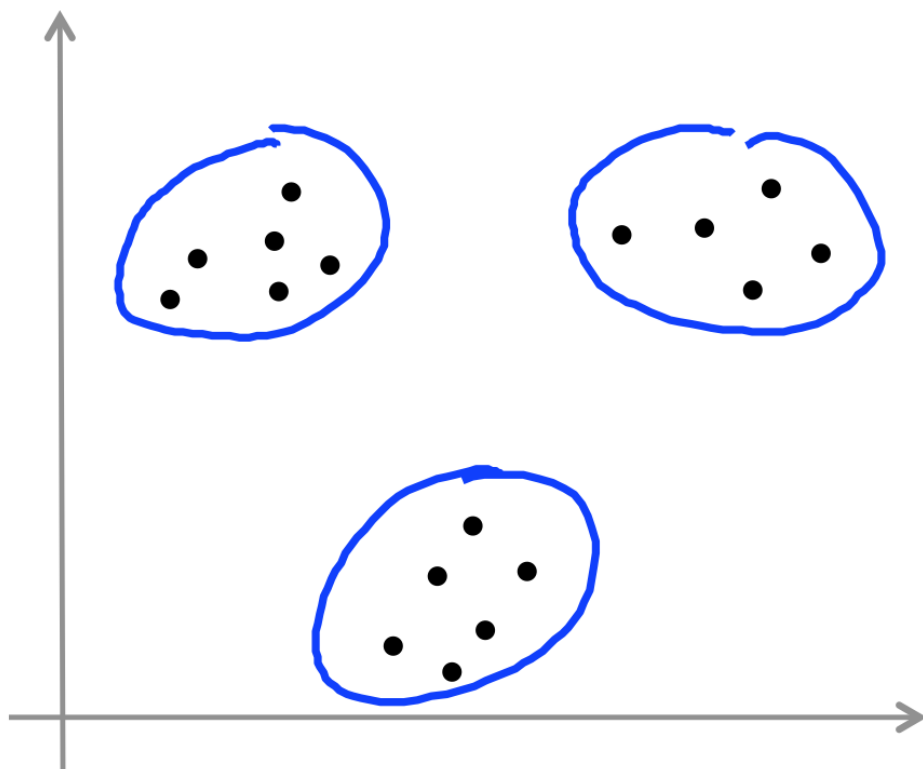


# 无监督学习

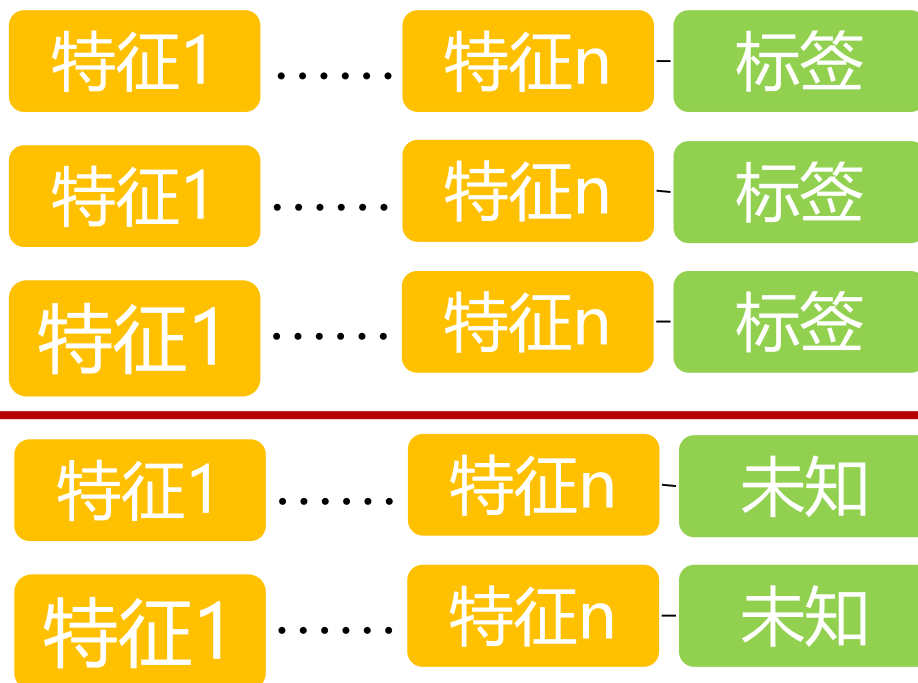


# 无监督学习

---



# 机器学习的分类



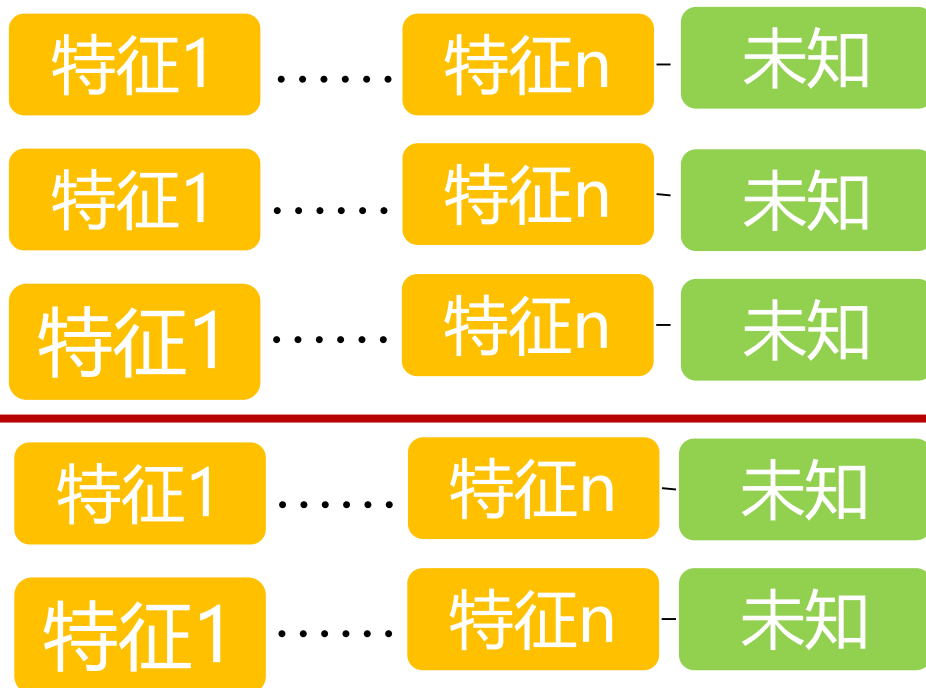
训练

监督学习算法

预测



# 机器学习的分类



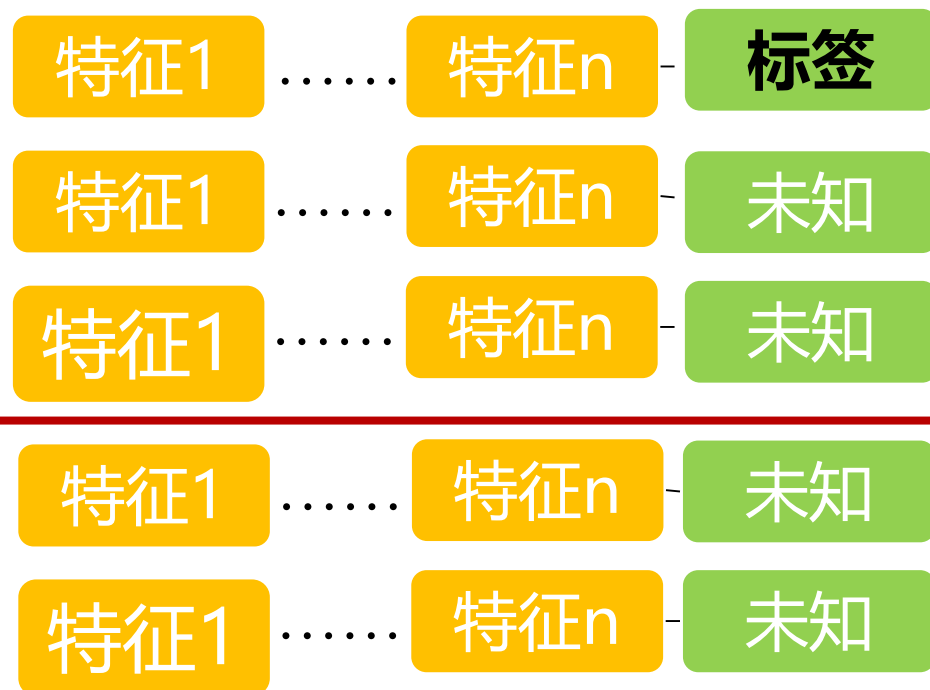
训练

无监督学习算法

预测



# 机器学习的分类



训练

半监督学习算法

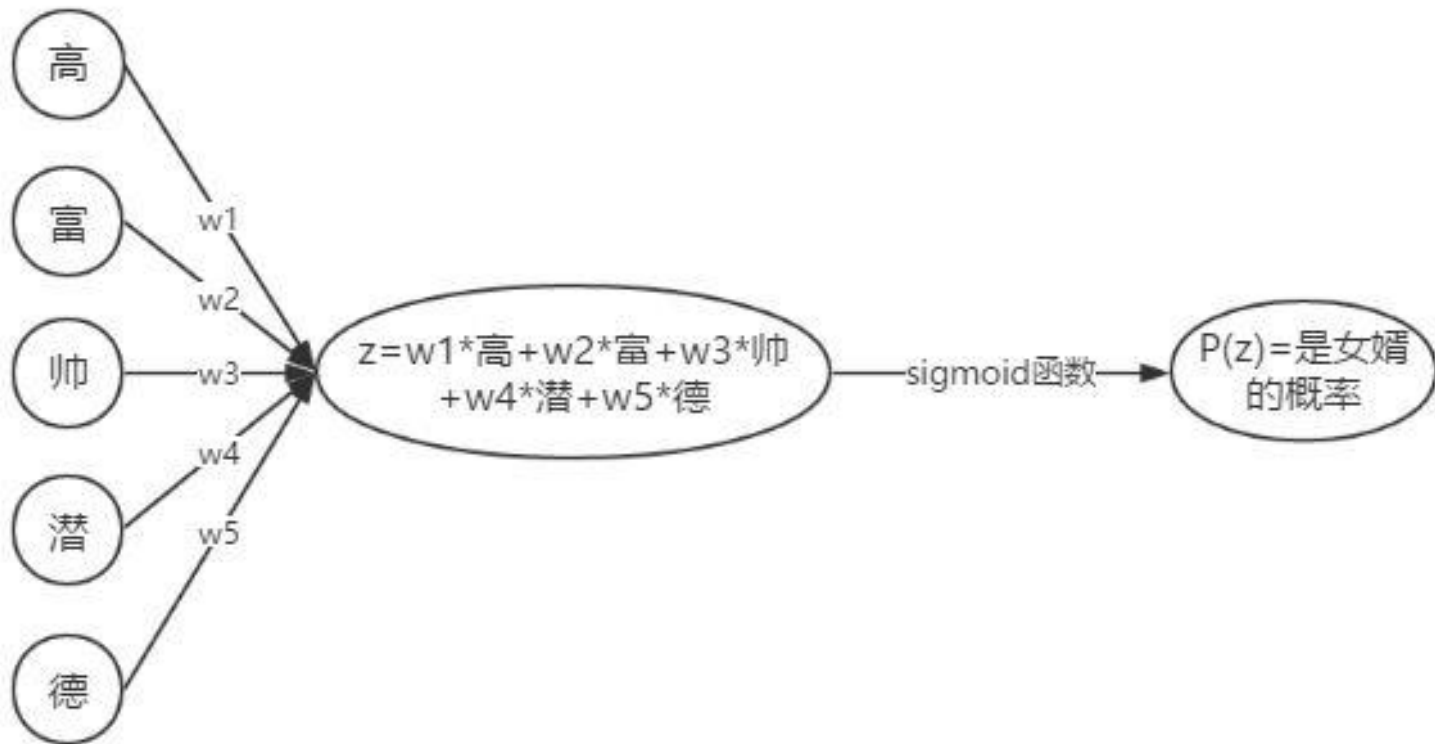
预测





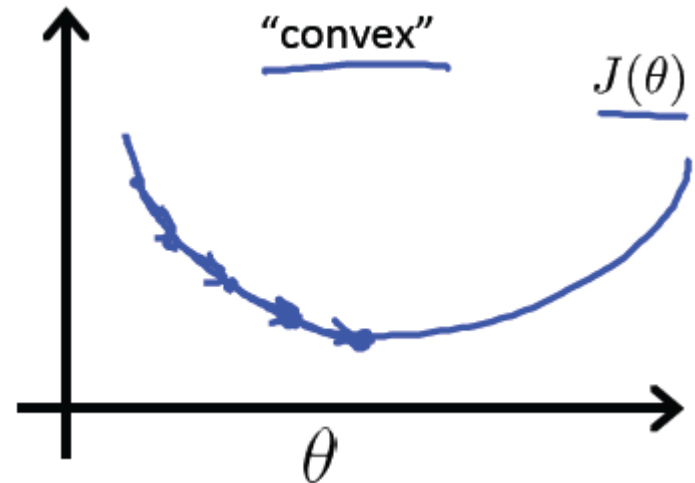
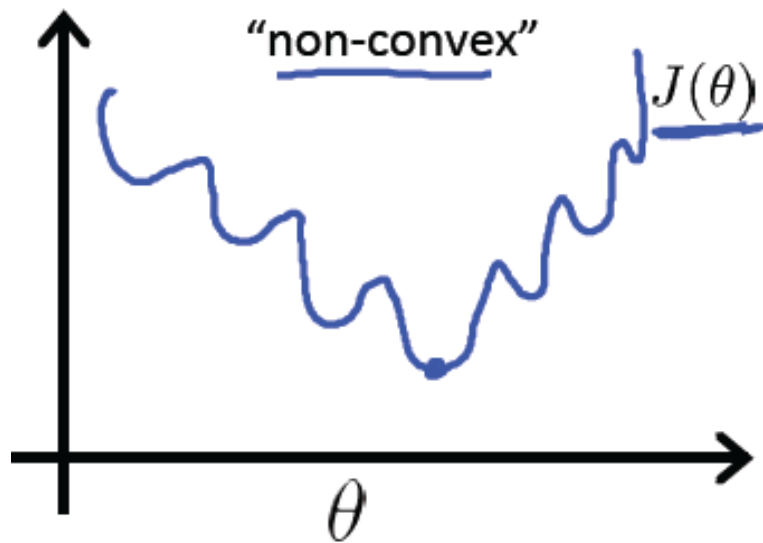
# 机器学习的一般思路

## □ 得分函数



# 机器学习的一般思路

## □ 损失函数的最优化问题

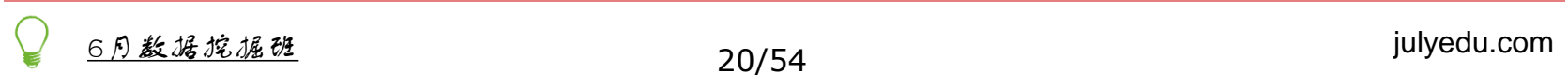


# 算法一览

## Machine Learning Algorithms *(sample)*

	<u>Unsupervised</u>	<u>Supervised</u>
<u>Continuous</u>	<ul style="list-style-type: none"><li>• Clustering &amp; Dimensionality Reduction<ul style="list-style-type: none"><li>○ SVD</li><li>○ PCA</li><li>○ K-means</li></ul></li></ul>	<ul style="list-style-type: none"><li>• Regression<ul style="list-style-type: none"><li>○ Linear</li><li>○ Polynomial</li></ul></li><li>• Decision Trees</li><li>• Random Forests</li></ul>
<u>Categorical</u>	<ul style="list-style-type: none"><li>• Association Analysis<ul style="list-style-type: none"><li>○ Apriori</li><li>○ FP-Growth</li></ul></li><li>• Hidden Markov Model</li></ul>	<ul style="list-style-type: none"><li>• Classification<ul style="list-style-type: none"><li>○ KNN</li><li>○ Trees</li><li>○ Logistic Regression</li><li>○ Naive-Bayes</li><li>○ SVM</li></ul></li></ul>



[illegible]

# 相关资料

---

- ❑ Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006
- ❑ Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012
- ❑ 李航, 统计学习方法, 清华大学出版社, 2012
- ❑ 周志华, 机器学习, 清华大学出版社, 2016
- ❑ Machine Learning, Andrew Ng, coursera
- ❑ 机器学习基石/技术, 林轩田, coursera



# 高等数学回顾

---

□ 微积分之：两边夹定理/夹逼定理

□ 当  $x \in U(x_0, r)$  时，有  $g(x) \leq f(x) \leq h(x)$  成立，  
并且  $\lim_{x \rightarrow x_0} g(x) = A$ ， $\lim_{x \rightarrow x_0} h(x) = A$ ，那么

$$\lim_{x \rightarrow x_0} f(x) = A$$



# 导数



- 简单的说，导数就是曲线的斜率，是曲线变化快慢的反应
- **二阶导数**是斜率变化快慢的反应，表征曲线的**凸凹性**
  - 在GIS中，往往一条二阶导数连续的曲线，我们称之为“**光顺**”的。
  - 还记得高中物理老师时常念叨的吗：**加速度的**方向总是指向轨迹曲线凹的一侧



# 常用函数的导数

---

$$C' = 0$$

$$(x^n)' = nx^{n-1}$$

$$(\sin x)' = \cos x$$

$$(\cos x)' = -\sin x$$

$$(a^x)' = a^x \ln a$$

$$(e^x)' = e^x$$

$$(\log_a x)' = \frac{1}{x} \log_a e$$

$$(\ln x)' = \frac{1}{x}$$

$$(u + v)' = u' + v'$$

$$(uv)' = u'v + uv'$$





# Taylor公式 – Maclaurin公式

---

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_n(x)$$

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + o(x^n)$$



# Taylor公式的应用1

□ 数值计算：初等函数值的计算(在原点展开)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \cdots + (-1)^{m-1} \frac{x^{2m-1}}{(2m-1)!} + R_{2m}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + R_n$$

□ 在实践中，往往需要做一定程度的变换。

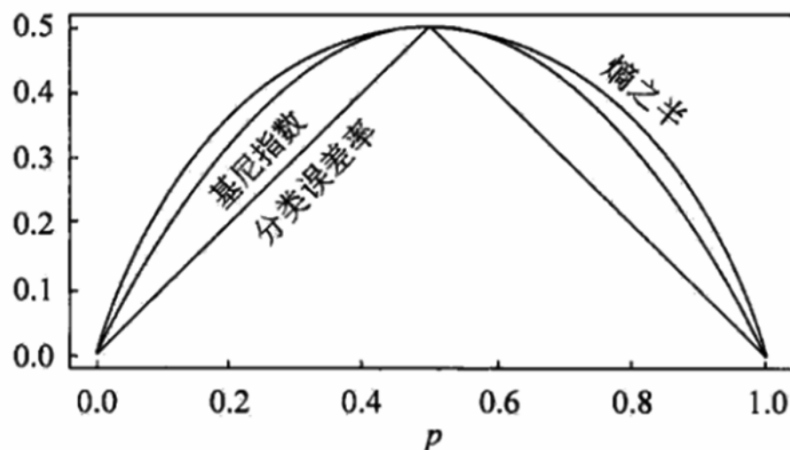


# Taylor公式的应用2

□ 考察基尼指数的图像、熵、分类误差率三者之间的关系

■ 将 $f(x)=-\ln x$ 在 $x=1$ 处一阶展开，忽略高阶无穷小，得到 $f(x) \approx 1-x$

$$\begin{aligned} H(X) &= -\sum_{k=1}^K p_k \ln p_k \\ &\approx \sum_{k=1}^K p_k (1-p_k) \end{aligned}$$



■ 上述结论，在决策树章节中会进一步讨论



# 方向导数

---

- 如果函数 $z=f(x,y)$ 在点 $P(x,y)$ 是可微分的，那么，函数在该点沿任一方向 $L$ 的方向导数都存在，且有：

$$\frac{\partial f}{\partial l} = \frac{\partial f}{\partial x} \cos \varphi + \frac{\partial f}{\partial y} \sin \varphi$$

- 其中， $\psi$ 为 $x$ 轴到方向 $L$ 的转角。



# 梯度

- 设函数 $z=f(x,y)$ 在平面区域 $D$ 内具有一阶连续偏导数，则对于每一个点 $P(x,y) \in D$ ，向量

$$\left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

为函数 $z=f(x,y)$ 在点 $P$ 的梯度，记做 $\text{grad}f(x,y)$

- 梯度的方向是函数在该点变化最快的方向
  - 考虑一座解析式为 $z=H(x,y)$ 的山，在 $(x_0,y_0)$ 的梯度是在该点坡度变化最快的方向。
- 梯度下降法
  - 思考：若下山方向和梯度呈 $\theta$ 夹角，下降速度是多少？

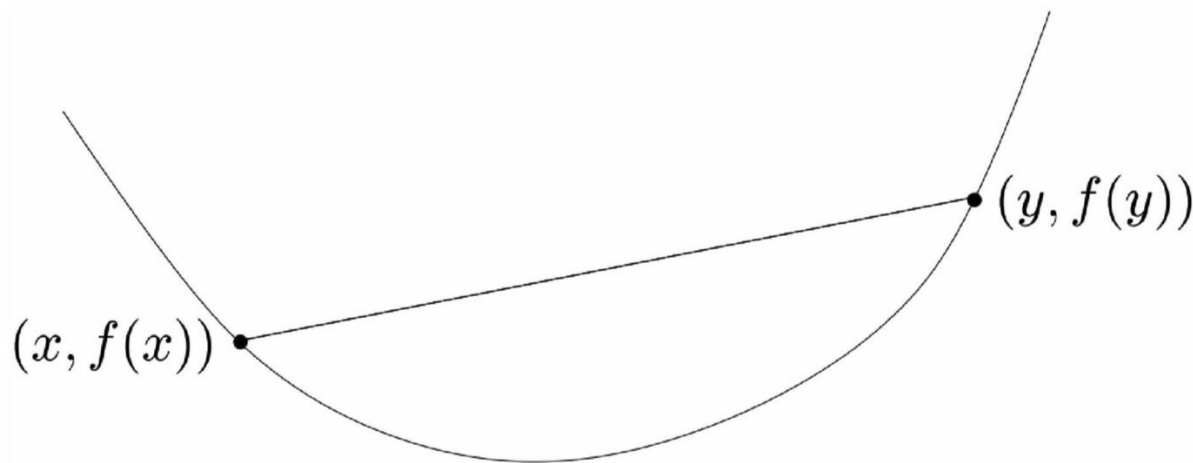


# 凸函数

□ 若函数 $f$ 的定义域 $\text{dom}f$ 为凸集，且满足

$\forall x, y \in \text{dom} f, 0 \leq \theta \leq 1$ ，有

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$



# 凸函数的判定

---

- 定理： $f(x)$ 在区间 $[a,b]$ 上连续，在 $(a,b)$ 内二阶可导，那么：
  - 若 $f''(x) > 0$ ，则 $f(x)$ 是凸的；
  - 若 $f''(x) < 0$ ，则 $f(x)$ 是凹的
  
- 即：一元二阶可微的函数在区间上是凸的，当且仅当它的二阶导数是非负的



# 凸函数

## □ 凸函数的表述

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

$f$  为凸函数，则有：

$$f(\theta_1 x_1 + \dots + \theta_n x_n) \leq \theta_1 f(x_1) + \dots + \theta_n f(x_n)$$

其中  $0 \leq \theta_i \leq 1, \theta_1 + \dots + \theta_n = 1$ .

□ 意义：可以在确定函数的凸凹性之后，对函数进行不等式替换。





# Jensen不等式：若f是凸函数

## □ 基本Jensen不等式

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$

## □ 若 $\theta_1, \dots, \theta_k \geq 0, \theta_1 + \dots + \theta_k = 1$

## □ 则 $f(\theta_1 x_1 + \dots + \theta_k x_k) \leq \theta_1 f(x_1) + \dots + \theta_k f(x_k)$

## □ 若 $p(x) \geq 0$ on $S \subseteq \text{dom } f, \int_S p(x) dx = 1$

## □ 则 $f\left(\int_S p(x)x dx\right) \leq \int_S f(x)p(x) dx$

$$f(\mathbf{E} x) \leq \mathbf{E} f(x)$$

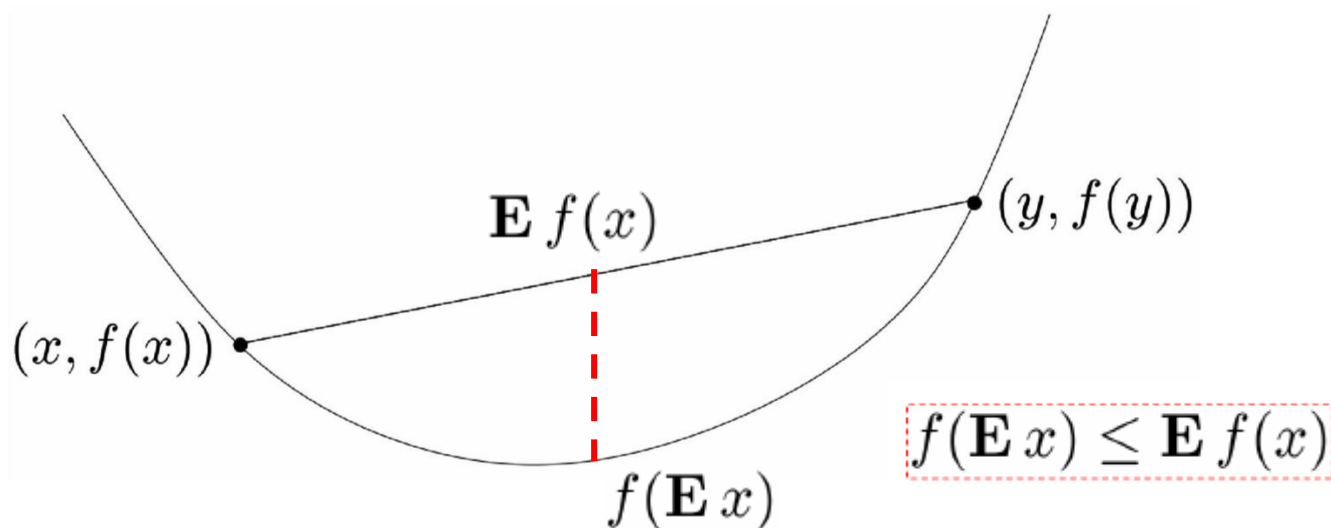


# 凸函数

□ 若函数 $f$ 的定义域 $\text{dom}f$ 为凸集，且满足

$\forall x, y \in \text{dom} f, 0 \leq \theta \leq 1$ ，有

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$



# 概率公式

---

□ 条件概率:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

□ 全概率公式:

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

□ 贝叶斯(Bayes)公式:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$



# 贝叶斯公式 $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

□ 给定某系统的若干样本 $x$ ，计算该系统的参数，即

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

- $P(\theta)$ : 没有数据支持下， $\theta$ 发生的概率：先验概率。
- $P(\theta|x)$ : 在数据 $x$ 的支持下， $\theta$ 发生的概率：后验概率。
- $P(x|\theta)$ : 给定某参数 $\theta$ 的概率分布：似然函数。

□ 例如：

- 在没有任何信息的前提下，猜测某人姓氏：先猜李王张刘……猜对的概率相对较大：先验概率。
- 若知道某人来自“牛家村”，则他姓牛的概率很大：后验概率——但不排除他姓郭、杨等情况。



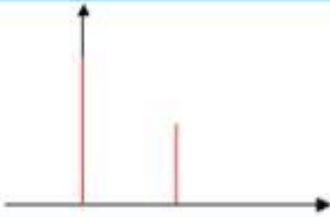

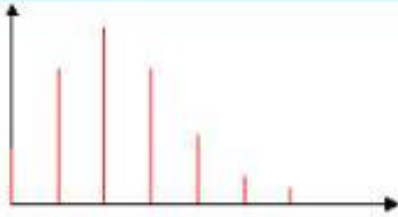
# 常见的概率分布

分 布	参 数	数学期望	方差
两点分布	$0 < p < 1$	$p$	$p(1 - p)$
二项分布	$n \geq 1,$ $0 < p < 1$	$np$	$np(1 - p)$
泊松分布	$\lambda > 0$	$\lambda$	$\lambda$
均匀分布	$a < b$	$(a + b)/2$	$(b - a)^2 / 12$
指数分布	$\theta > 0$	$\theta$	$\theta^2$
正态分布	$\mu, \sigma > 0$	$\mu$	$\sigma^2$



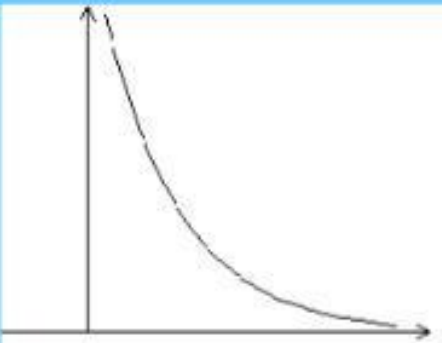
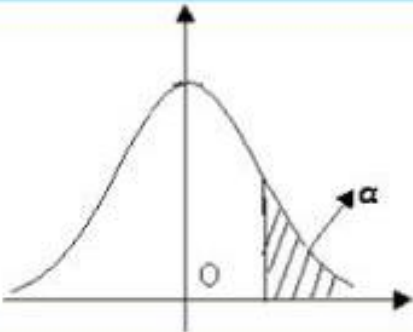

# 常见的概率分布

概率分布表

分布名称	概率与密度函数 $p(x)$	数学期望	方差	图形
贝努里分布 两点分布	$p_k = \begin{cases} q, & k=0 \\ p, & k=1 \end{cases}$ $0 < p < 1, \quad q = 1 - p$	$p$	$pq$	
二项分布 $b(k, n, p)$	$b(k, n, p) = \binom{n}{k} p^k q^{n-k}$ $k = 0, 1, \dots, n$ $0 < p < 1, \quad q = 1 - p$	$np$	$npq$	
泊松分布 $p(k, \lambda)$	$p(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0$ $k = 0, 1, 2, \dots, n$	$\lambda$	$\lambda$	



# 常见的概率分布

指数分布	$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$ $\lambda > 0, \text{ 常数}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$	
正态分布 高斯分布 $N(a, \sigma^2)$	$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-a)^2/2\sigma^2}$ $-\infty < x < \infty, a, \sigma > 0, \text{ 常数}$	$a$	$\sigma^2$	
均匀分布 $U[a, b]$	$p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{其他} \end{cases}$ $a < b, \text{ 常数}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	



# 概率与统计的关注点

## □ 概率论问问题的方式：

### 装箱问题

- 将12件正品和3件次品随机装在3个箱子中，每箱装5件，则每箱中恰有1件次品的概率是多少？

## □ 数理统计问问题的方式：

### 例：正态分布的矩估计

- 在正态分布的总体中采样得到 $n$ 个样本： $X_1, X_2, \dots, X_n$ ，估计该总体的均值和方差。





# 概率与统计的关注点

□ 根据是否已知整体进行区分：

## 装箱问题

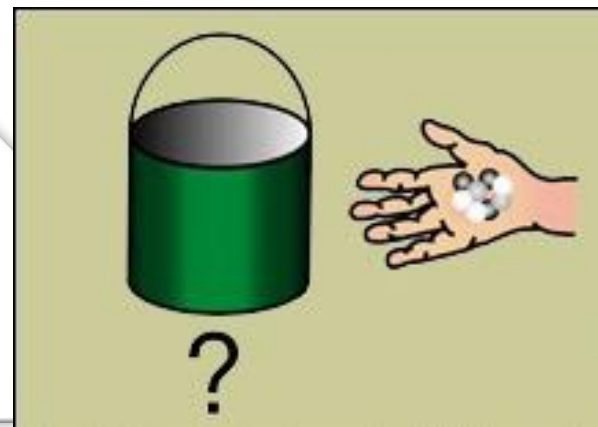
□ 将12件正品和3件次品随机装在3个箱子中，每箱装5件，则每箱中恰有1件次品的概率是多少？



□ 统计问题是概率问题的逆向工程：

## 例：正态分布的矩估计

□ 在正态分布的总体中采样得到 $n$ 个样本： $X_1, X_2, \dots, X_n$ ，估计该总体的均值和方差。



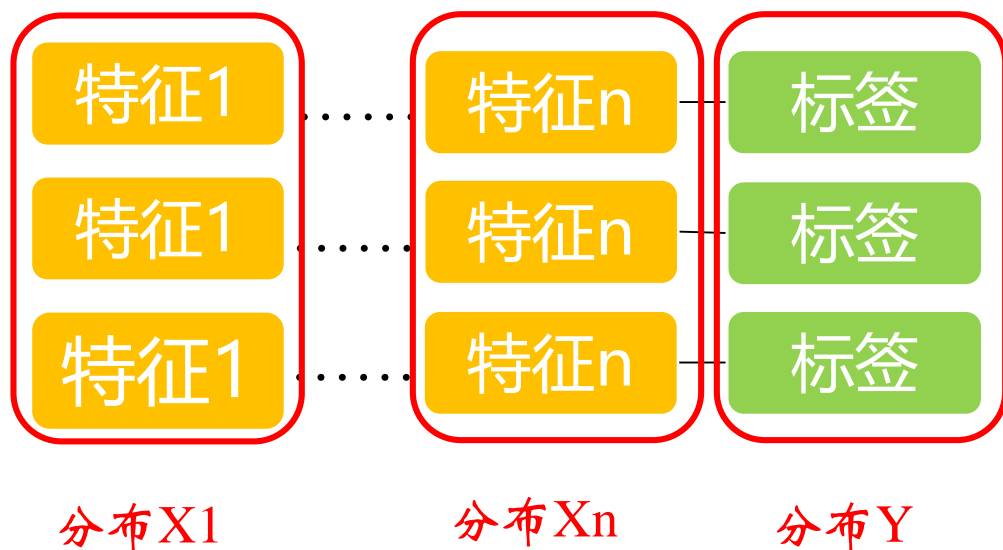
# 概率统计与机器学习的关系



监督学习算法



# 概率统计与机器学习的关系



□ 可基于各个分布的特性来评估模型和样本



# 概率统计与机器学习的关系

---

- 统计估计的是分布，机器学习训练出来的是模型，模型可能包含了很多分布。
- 训练与预测过程的一个核心评价指标就是模型的误差。
- 误差本身就可以是概率的形式，与概率紧密相关。
- 对误差的不同定义方式就演化成了不同损失函数的定义方式。
- 机器学习是概率与统计的进阶版本。（不严谨的说法）



# 重要统计量

---

□ 都是描述全局（整体）统计量

■ 期望

■ 方差

■ 协方差



# 期望

---

□ 离散型  $E(X) = \sum_i x_i p_i$

□ 连续型  $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

□ 即：概率加权下的“平均值”



# 方差

□ 定义

$$\text{Var}(X) = E\{[X - E(X)]^2\} = E(X^2) - E^2(X)$$

□ 无条件成立  $\text{Var}(c) = 0$

$$\text{Var}(X + c) = \text{Var}(X)$$

$$\text{Var}(kX) = k^2 \text{Var}(X)$$

□ X和Y独立

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

■ 此外，方差的平方根，称为标准差



# 协方差

---

□ 定义  $Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$

□ 性质:

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(aX + b, cY + d) = acCov(X, Y)$$

$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$





# A·x的几何意义

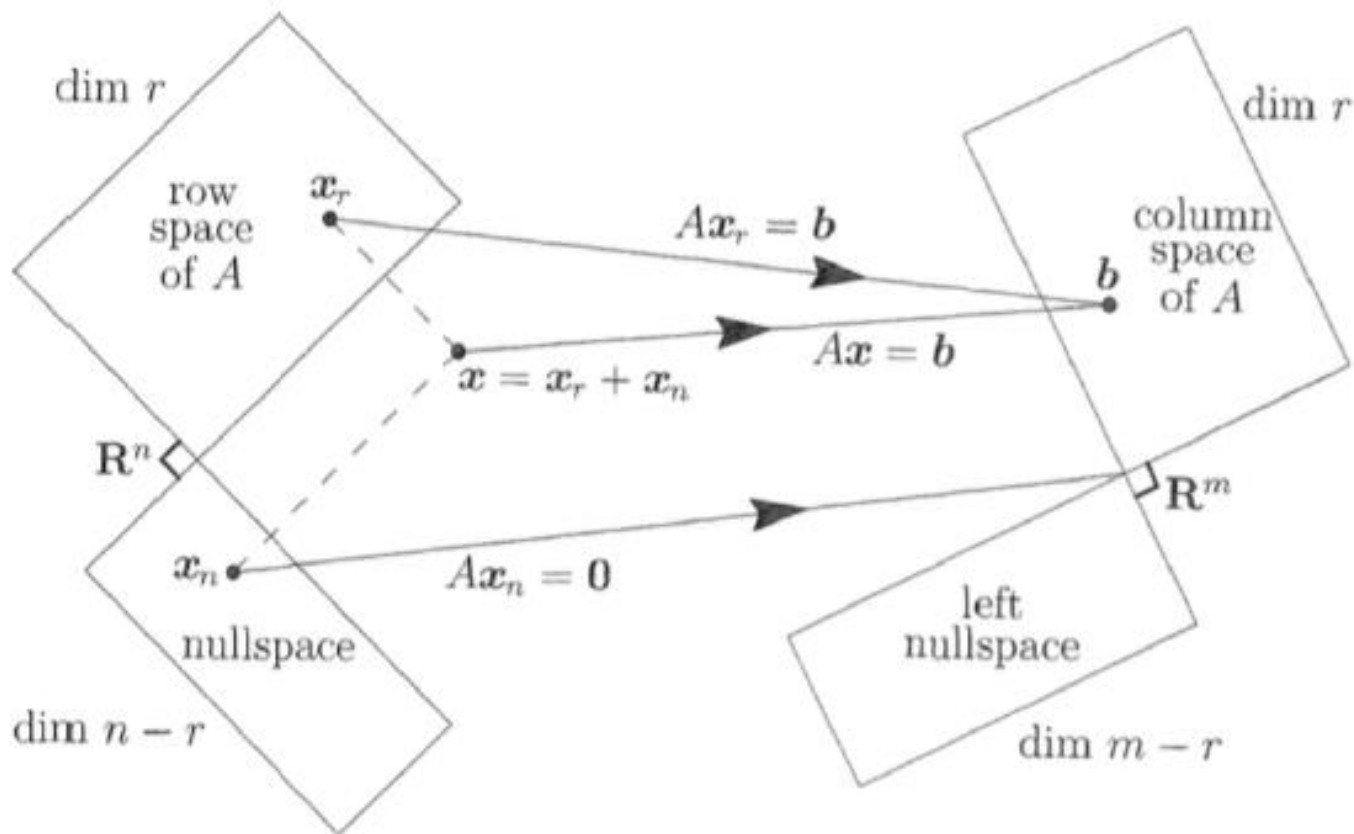
- 矩阵绝不是把一堆数用括号括起来！本课件中，假设 $\mathbf{A} \in \mathbb{R}^{m \times n}$ 。

$$\underbrace{\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}}_{\mathbf{A} \in \mathbb{R}^{2 \times 2}} \underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{\mathbf{x} \in \mathbb{R}^2} = \underbrace{\begin{bmatrix} 1 \\ 5 \end{bmatrix}}_{\mathbf{b} \in \mathbb{R}^2} \quad (1)$$

$$\underbrace{\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}}_{\mathbf{A} \in \mathbb{R}^{3 \times 3}} \underbrace{\begin{bmatrix} u \\ v \\ w \end{bmatrix}}_{\mathbf{x} \in \mathbb{R}^3} = \underbrace{\begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix}}_{\mathbf{b} \in \mathbb{R}^3} \quad (2)$$



# 四个基本子空间



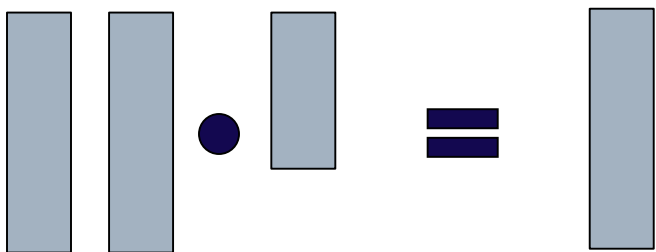
# 能看懂这张表就赚到了

	线性组合产生 (几何意义1)	正交性判断 (几何意义2)
列视图：以每一列为一个向量进行观察	$A x_n = y_m$ 列空间 (右乘) $r$	$y_m^T \cdot A = 0$ 左零空间 (左乘) (列补) $m-r$
行视图：以每一行为一个向量进行观察	$y_m^T \cdot A = x_n$ 行空间 (左乘) $r$	$A x_n = 0$ 零空间 (右乘) (行补) $n-r$

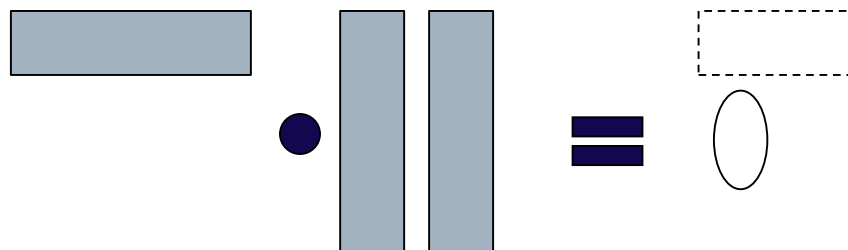


# $A \cdot x$ 的几何意义

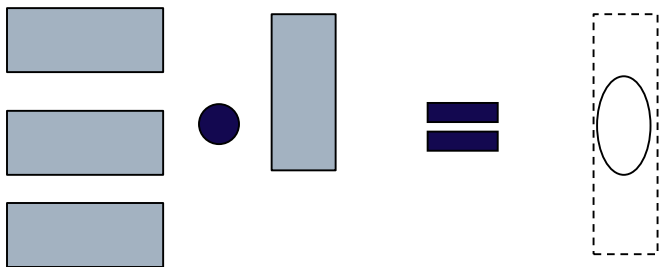
□ 产生列空间



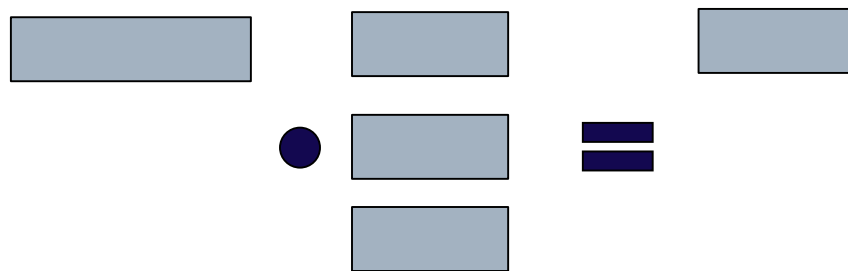
□ 列空间正交性



□ 行空间正交性



□ 产生行空间



# 矩阵乘法在计算中的优势

---

- ❑ 将很多for循环写成矩阵或者向量乘法的方式。
- ❑ 矩阵计算模块在底层有优化。
- ❑ Numpy进行矩阵运算很快：现场实例



---

感谢大家！

恳请大家批评指正！

