第2章: 模型评估与选择

Chapter 2: Model Evaluation and Selection

张永飞

2023年9月12日

课前回顾

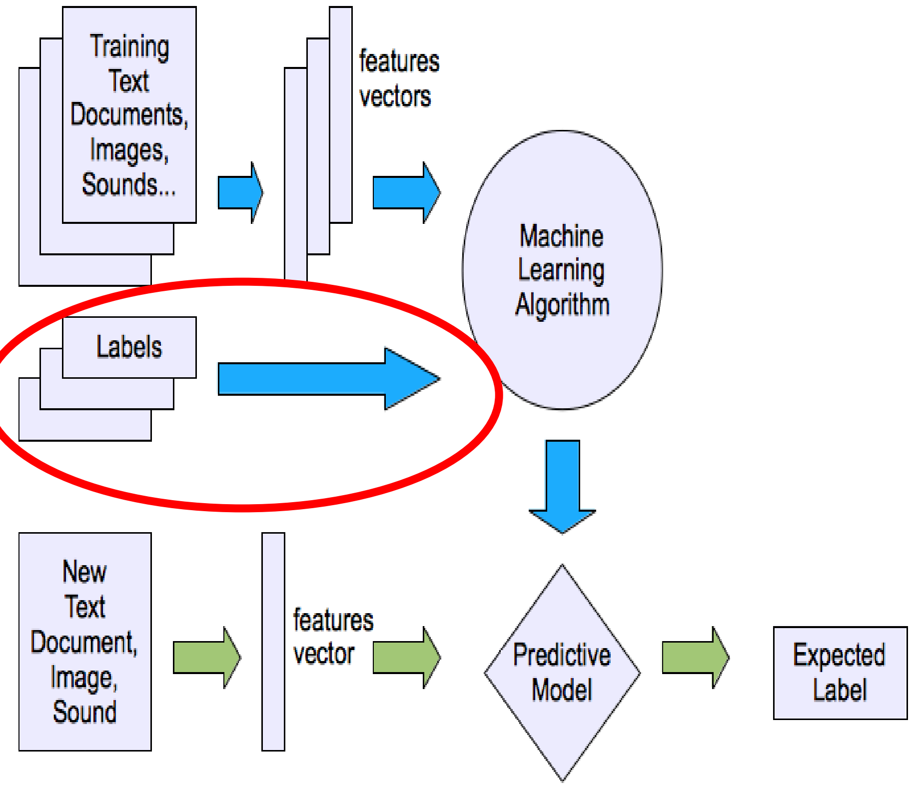
机器学习算法

机器学习主要问题

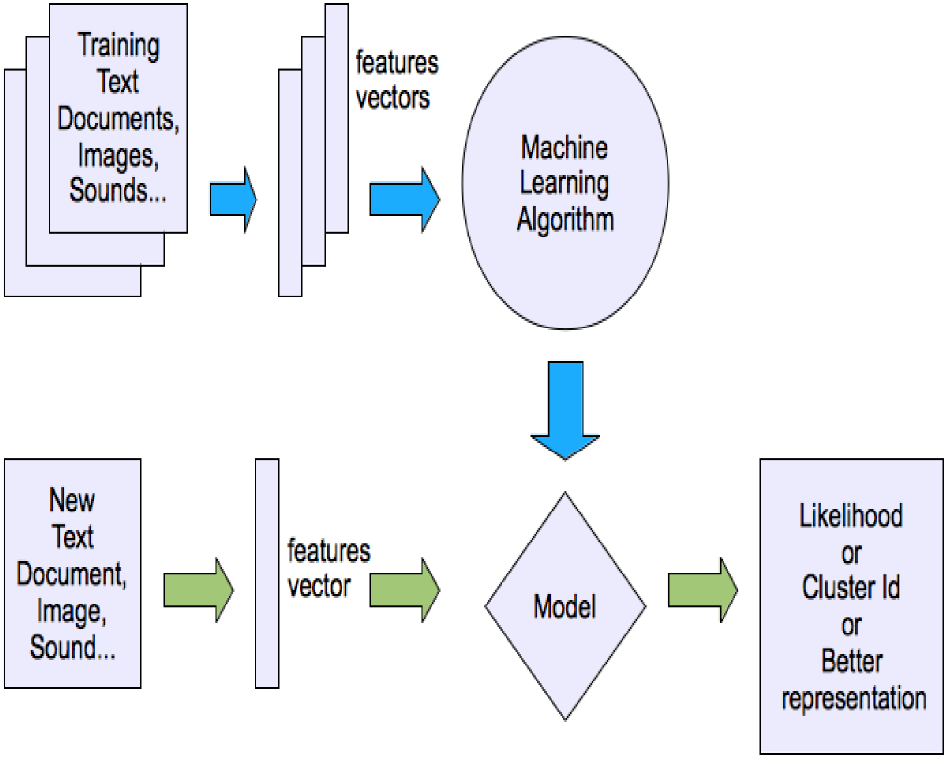
	Supervised Learning	Unsupervised Learning
Discrete	Classification or Categorization	Clustering
Continuous	Regression	Dimensionality Reduction

监督/非监督学习算法流程

● 监督学习



● 无监督学习



分类问题-图像分类

- 训练过程

Training Images

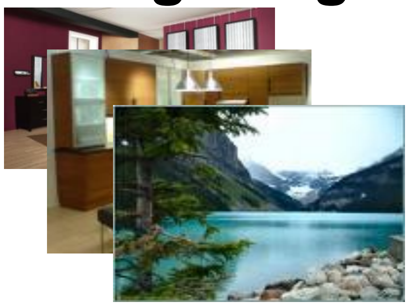


Image Features

Training Labels

Classifier Training

Trained Classifier

- 测试过程

Test Image

Image Features

Trained Classifier

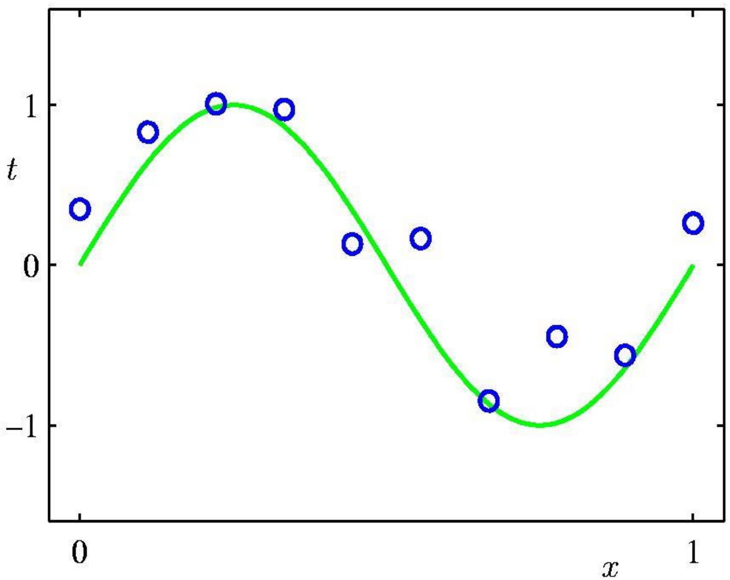
Prediction

Outdoor

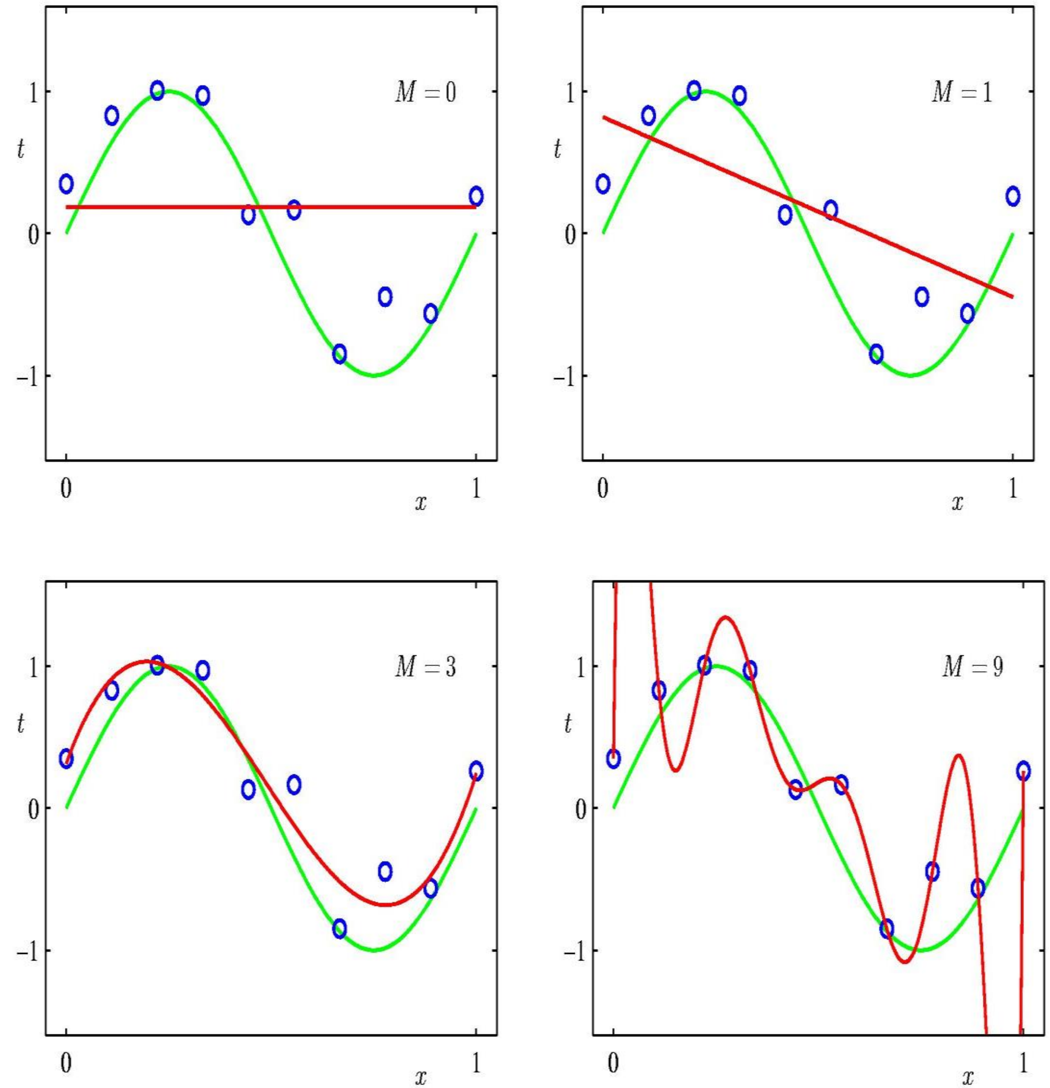
回归问题-曲线拟合

- **曲线拟合**

$\sin(2\pi x)$

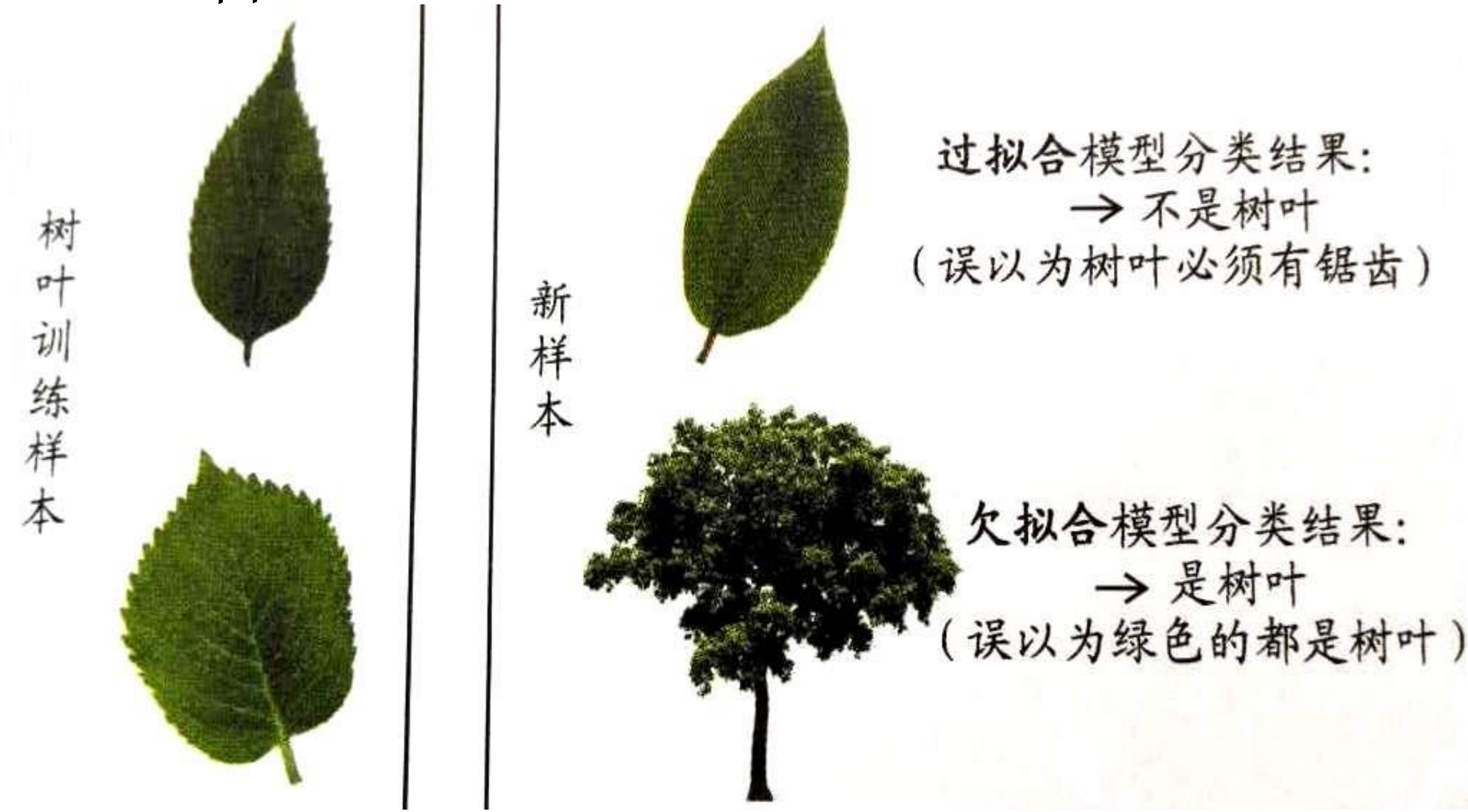


$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



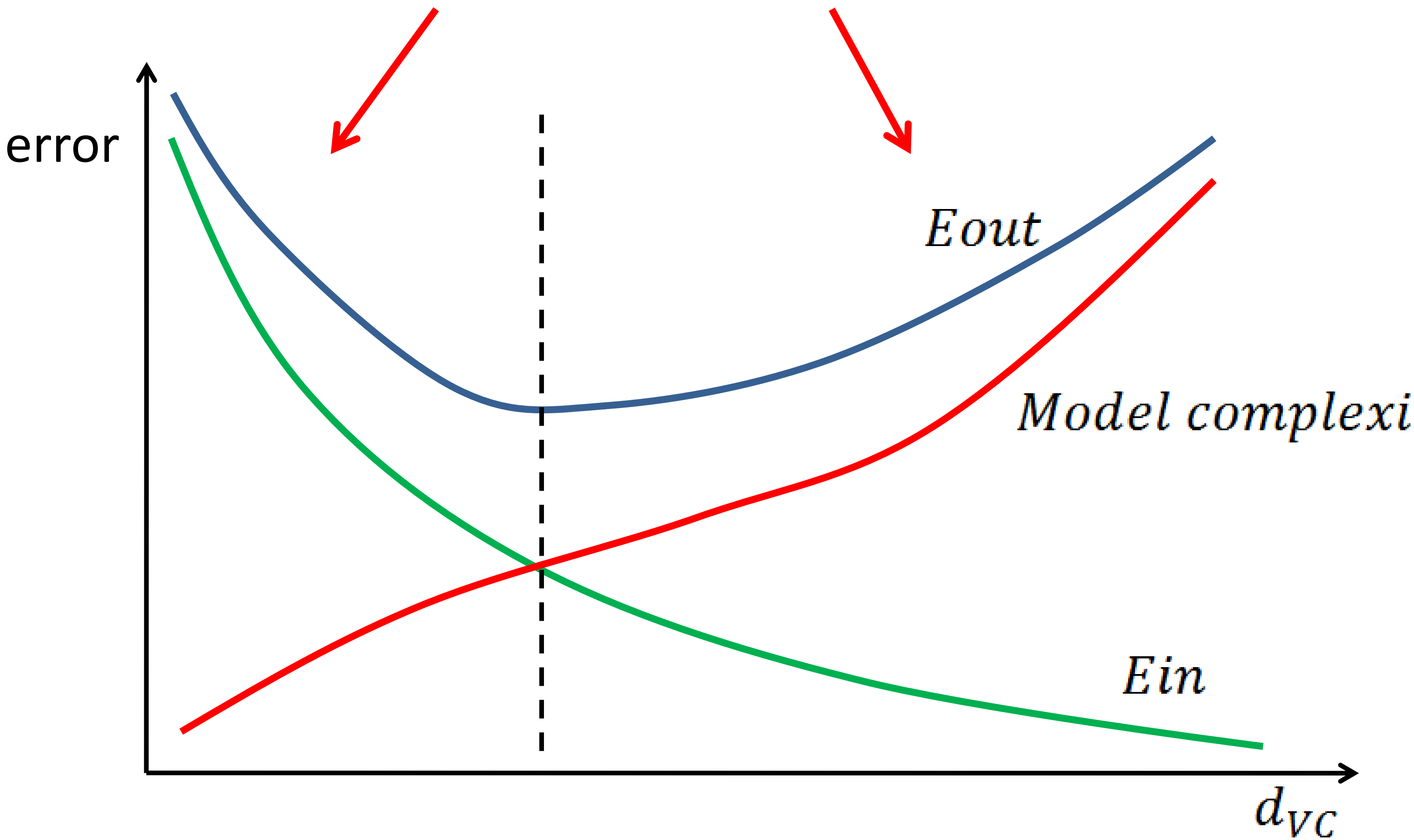
过拟合vs.欠拟合

- P24 图2.1



机器学习目标

Under-fitting VS. Over-fitting (fixed *N*)



第2章: 模型评估与选择

Chapter 2: Model Evaluation and Selection

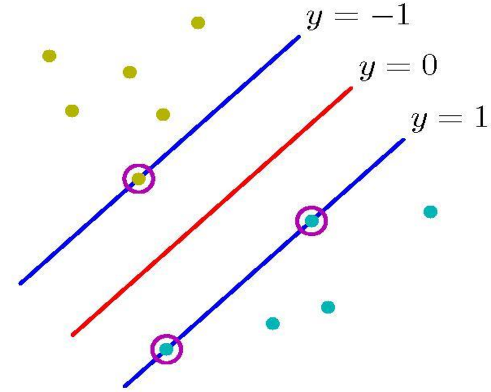
张永飞

2023年9月12日

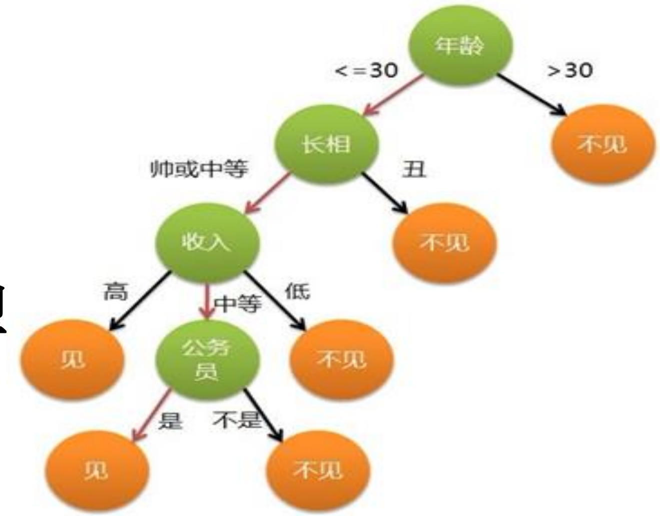
模型评估与选择

$$\gamma = \frac{2}{\|\omega\|}$$

- **模型性能有差异**
  - **同一问题，多种算法/模型**
    - **分类问题：贝叶斯决策、决策树、SVM…**



$$x \in \omega_k \quad iff \;\; k = \arg\max_i \{P(\omega_i | \mathbf{x})\}$$



  - **同一算法/模型，不同参数配置**
    - **例如：由不同训练数据得到的模型**
      训练的不同阶段

误 差

- **误差**(error)：算法/模型的实际预测输出与样本的真实输出之间的差异
- **训练误差/经验误差**(training/empirical error)：学习器在训练集上的误差
- **泛化误差**(generalization error)：学习器在新样本上的误差

模型评估与选择

- **目标**：min(泛化误差)

- **问题**：新样本未知

- **测试误差(testing error)**：学习器在测试集上的误差

- **目标**：min(泛化误差) → min(测试误差)

模型评估与选择

- 目标：min(泛化误差) → min(测试误差)

- 模型评估与选择
  - 1. 对数据集进行划分，分为训练集和测试集两部分
  - 2. 在训练集上训练得到模型
  - 3. 对模型在测试集上面的泛化性能进行度量
  - 4. 基于测试集上的泛化性能，依据假设检验来推广到全部数据集上面的泛化性能(延伸自学教材2.4-2.5)

数据集划分

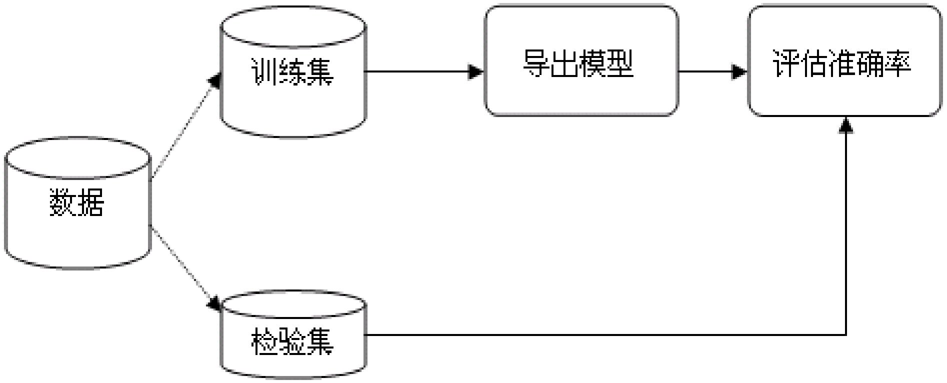
- 数据集划分
  - 目标：将数据集D划分为训练集S和测试集T两部分，在训练集上训练模型，然后在测试集上评估其性能
  - 原则：测试集应尽量与训练集互斥；即测试样本尽量不在训练集中出现，未在训练过程中使用
  - 示例：练习，考试

数据集划分

- 数据集划分
  - 目标：将数据集D划分为训练集S、验证集V和测试集T三部分，在训练集上训练模型，在验证集上调整模型超参数，并对模型的能力（是否过拟合）进行初步评估和选择，在验证集上然后在测试集上评估其性能
  - 原则：测试集、验证集应尽量与训练集互斥；即验证样本、测试样本尽量不在训练集中出现，未在训练过程中使用
  - 示例：练习，作业（月考、模考），考试（高考）

数据集划分

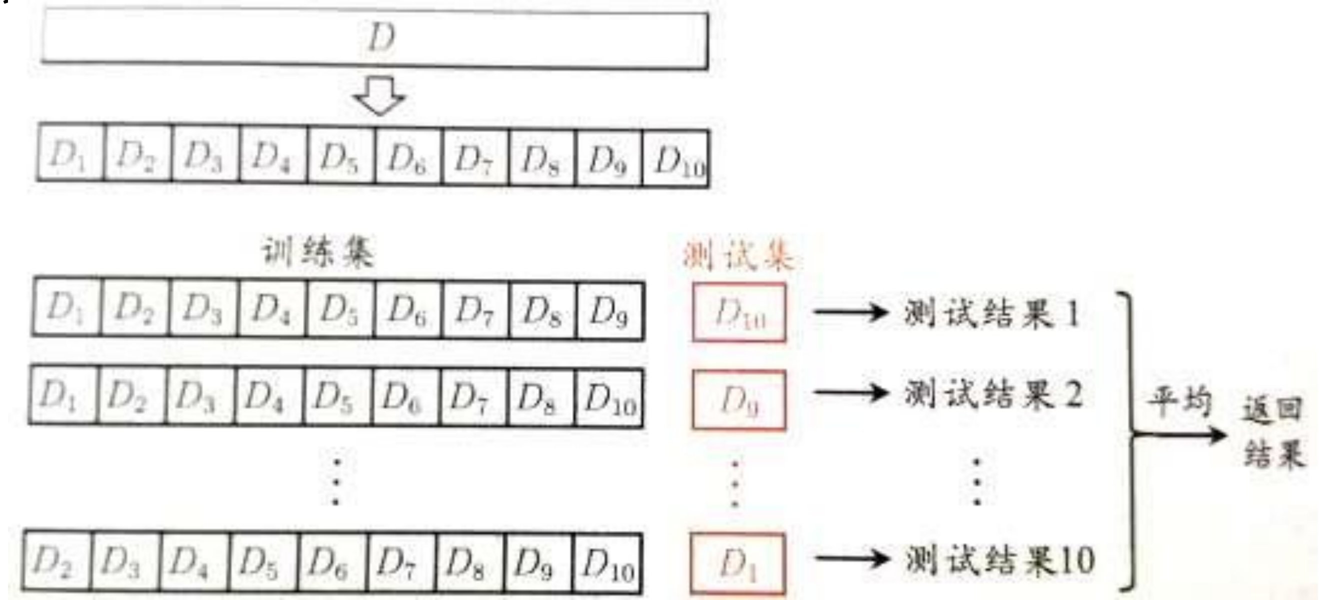
- **保持/留出法(hold-out)** ：给定数据随机地划分到两个独立的集合：训练集和测试集。通常，2/3的数据分配到训练集，其余1/3分配到测试集。使用训练集导出模型，用测试集来估计泛化误差
  - **优点**：简单
  - **缺点**：受数据划分影响大



- **随机子抽样(random sub-sampling)**：保持方法的一种变形；随机地选择训练集和测试集，将保持方法重复k次，总准确率估计取每次迭代准确率的平均值

数据集划分

- *k*折交叉验证(k-fold cross-validation)：初始数据数据被划分成 *k* 个大小相似、互不相交的子集/"折"。训练和测试 *k* 次；在第 *i* 次迭代，第 *i* 折用作测试集，其余的子集都用于训练学习，取 *k* 次测试结果的均值
- 与保持法和随机子抽样法不同，这里每个样本用于训练的次数相同，并且用于检验一次



数据集划分

- **留一法(leave-one-out):** 是k折交叉确认的特殊情况，其中k设置为初始样本数。用k-1个样本作为训练集，每次只给检验集“留出”一个样本，由此设计一个模型。从k个样本中选k-1个样本有k中选择，所以可用不同的大小为k-1训练样本重复进行k次
- **优点**：训练集比数据集只少一个样本，比较准确
- **缺点**：由于要设计 $k=|D|$ 个不同的模型并对其进行比较，当$|D|$较大时，这种方法计算量很大

数据集划分

- **自助法(bootstrapping):** 从初始样本D中有放回均匀抽样；即每当选中一个样本，它等可能地被再次选中并再次添加到训练集中；采样|D|次后，即可获取大小为|D|的训练样本集；没有进入训练集的数据样本形成测试集
- 样本在|D|次采样中始终不被采到的概率：
- $$\lim_{|D|\to\infty}\left(1-\frac{1}{|D|}\right)^{|D|} \mapsto \frac{1}{e} \approx 0.368 \quad \left(e=\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^{x} \right)$$
- **优势**：可产生多个不同训练样本集；对于小数据集，自助法效果胜过K折交叉验证；能从初始数据集中产生多个不同的训练集，这对集成学习等方法有很大的好处
- **缺点**：改变了数据集分布，会引入估计偏差

性能度量

- 回归任务
  - 均方误差(Mean Squared Error)

$$E(f;D)=\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2$$

  - 更一般情况：对于数据分布 D 和概率密度函数 $p(\cdot)$，均方误差可描述为：

$$E(f;D)=\int_{x\sim D}(f(x)-y)^2\,p(x)dx$$

其中：

- f：训练的学习器
- D：初始样本集，$D=\{(x_1, y_1), (x_2, y_2),\ldots,(x_n, y_n)\}$
- y_i：样本输入x_i的真实标记

性能度量

- 分类任务
  - 错误率： $$E(f;D)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{II}(f(x_i)\neq y_i)$$
  - 精　度： $$acc(f;D)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{II}(f(x_i)=y_i)=1-E(f;D)$$

其中：

- f：训练的学习器　　　　$II(.)$：指示函数
- D：初始样本集，$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$
- y_i：样本输入x_i的真实标记

性能度量

- 分类任务
  - 更一般情况：对于数据分布 D 和概率密度函数 $p(\cdot)$，错误率和精度可分别描述为：
  - 错误率：$$E(f;D)=\int_{x\sim D}\mathbb{I}(f(x)\neq y)p(x)dx$$
  - 精　度：$$acc(f;D)=\int_{x\sim D}\mathbb{I}(f(x)=y)p(x)dx$$
$$=1-E(f;D)$$

性能度量

- 错误率和精度：
  - 优点：理解直观，计算简单
  - 问题：
    - 数据类别不均衡时，占比大类别成为影响准确率的最主要因素
    - 仅能评估是否正确分类，无法提供更详细评估
  - 示例1：西瓜分类
    - 无法评估“挑出的西瓜中有多少比例是好瓜？”
    - 无法评估“所有好瓜中有多少比例被挑了出来？”
  - 示例2：信息检索
    - 无法评估“检索出的信息中有多少是用户感兴趣的？”
    - 无法评估“用户感兴趣信息中有多少被检索出来了？”

精度/查准率(precision)
召回率/查全率(Recall)

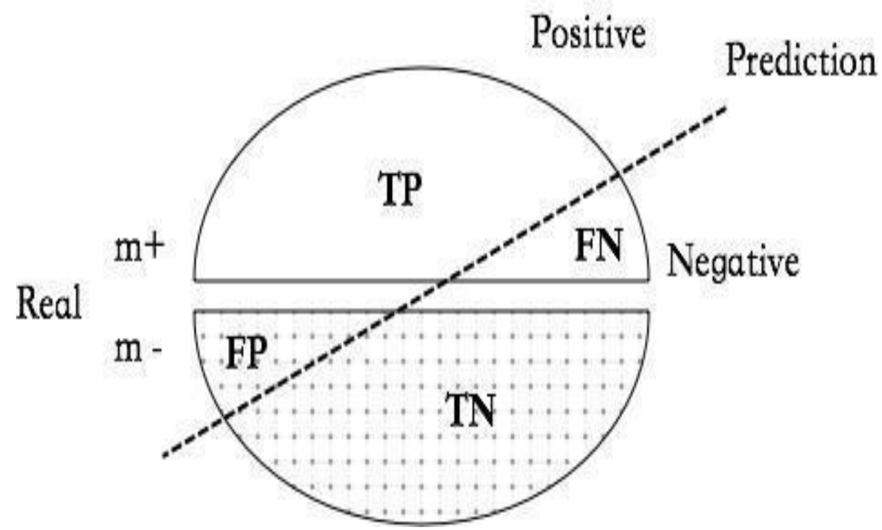
性能度量

- **混淆矩阵(Confusion Matrix)**：用来作为分类规则特征的表示，它包括了每一类的样本个数，包括正确的和错误的分类
- 主对角线给出了每一类正确分类的样本的个数，非对角线上的元素则表示未被正确分类的样本个数
- 对于m类的分类问题，误差可能有m^2-m种

性能度量

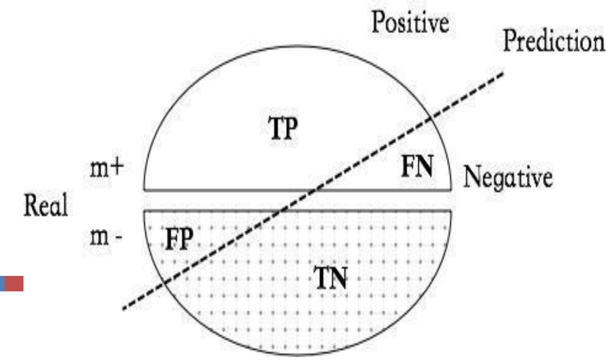
- **混淆矩阵-两类：** 仅有正、负样本2类，用P和N(或1和0)来表征：

	预测结果		
真实类别		正例P	负例P
	正例P	真正例TP	假负例FN
	负例P	假正例FP	真负例TN



- TP： 被分类器正确分类的正元组；期望为P，分类为P：称为真正
- TN：被分类器正确分类的负元组； 期望为N，分类为N：称为真负
- FP：被错误标记为正元组的负元组； 期望为N，分类为P：称为假正
- FN：被错误标记为负元组的正元组。期望为P，分类为N：称为假负

性能度量



- 准确率(识别率)：评估分类器正确识别正、负样本的能力

$$accuracy = \frac{TP + TN}{P + N}$$

- 错误率：评估分类器错误识别正、负样本的能力

$$ErrorRate = \frac{FP + FN}{P + N}$$

- 真阳性率（TPR）：评估分类器正确识别正样本的能力

$$SN = \frac{TP}{P} = \frac{TP}{TP + FN}$$

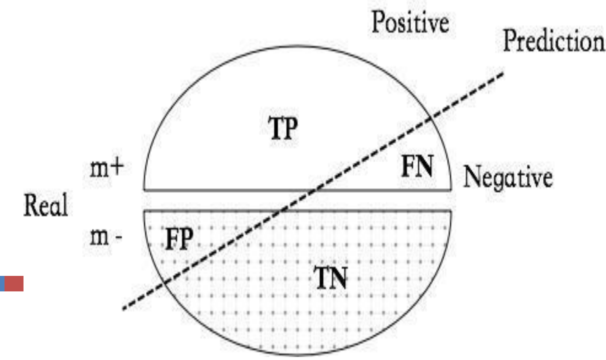
敏感性(sensitivity)

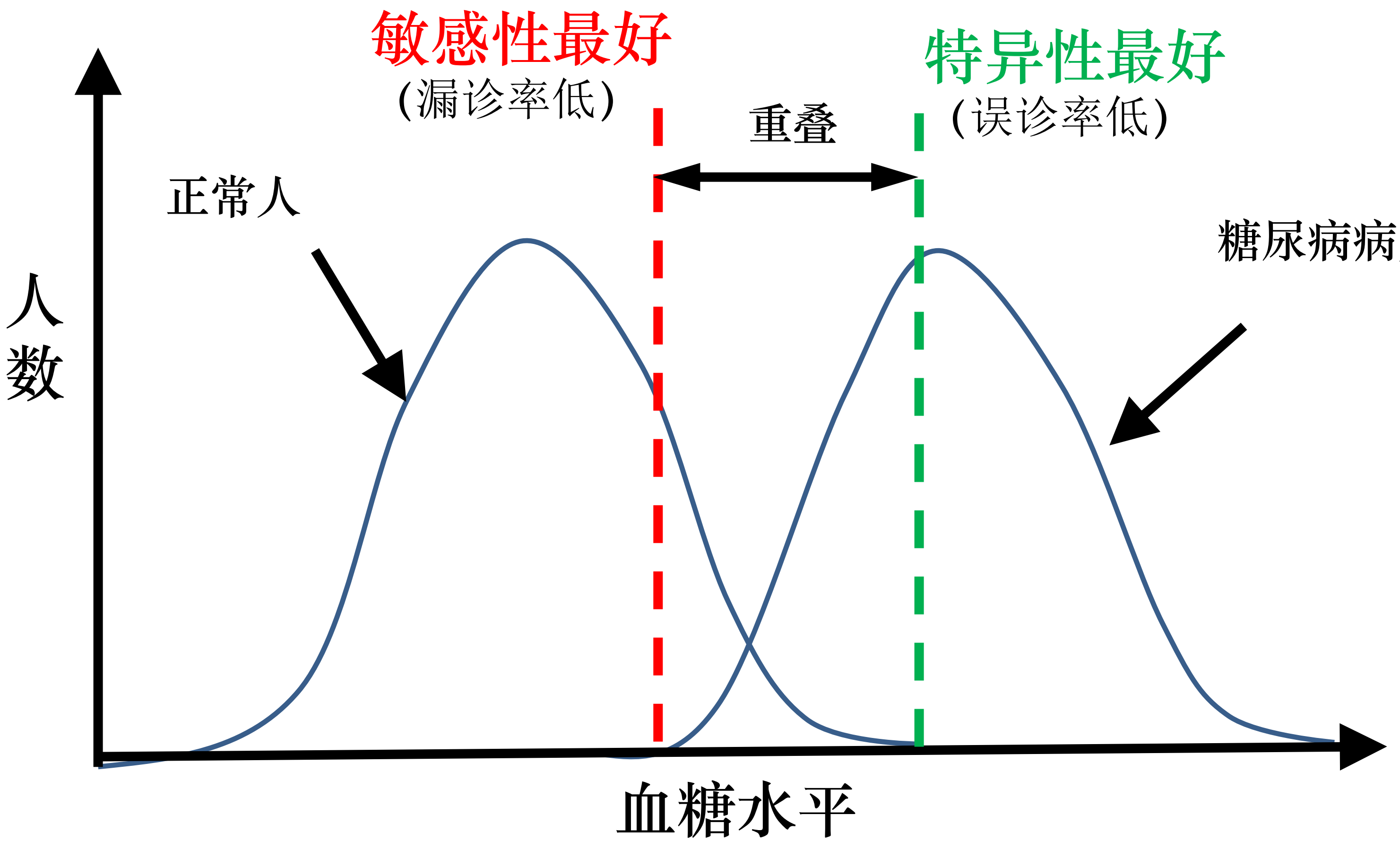
- 真阴性率（TNR）：评估分类器正确识别负样本的能力

$$SP = \frac{TN}{N} = \frac{TN}{TN + FP}$$

特异性(specificity)

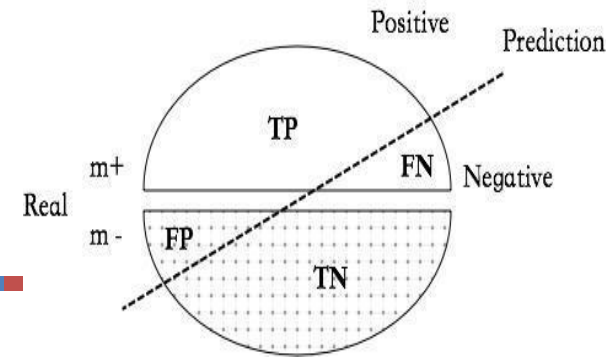
性能度量





新冠：密接、次密接；核酸

性能度量



- **精度/查准率(precision)：评估预测正样本中的真正样本**

$$percision = \frac{TP}{TP + FP}$$

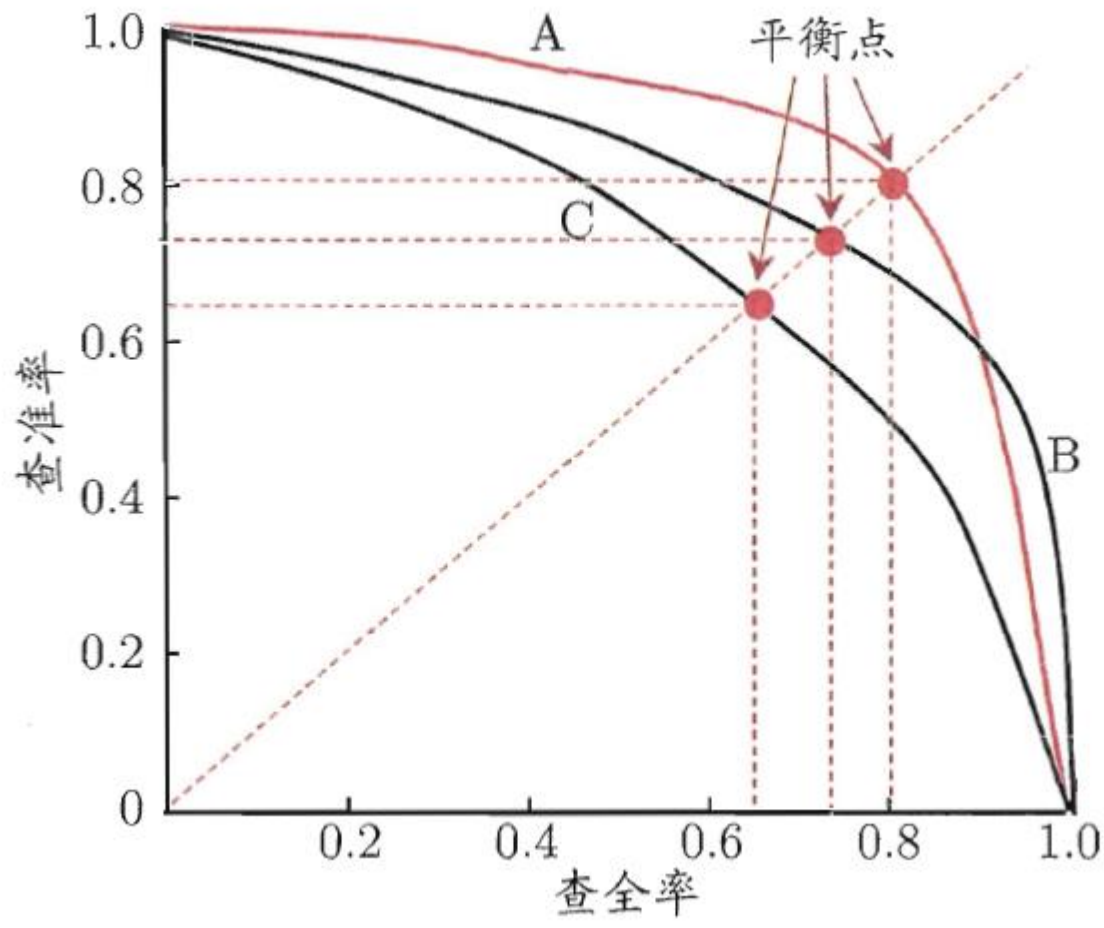
- **召回率/查全率(Recall)：评估分类器正确识别正样本的能力，等价于敏感性**

$$recall = \frac{TP}{TP + FN}$$

- **查准率和查全率互相矛盾。查准率高，则查全率低；反之亦然**
- **示例：挑西瓜**
  - **若想好瓜尽可能多选出来，则增大选瓜数量；极限，选上所有西瓜，则查全率最高(1)，但查准率较低**
  - **若想选出的瓜中好瓜比例高，则只选有把握的瓜，但会漏掉不少好瓜；即准率高了，但查全率较低**

性能度量- P-R曲线

- **定义**：以查全率R为横轴，查准率P为纵轴，根据模型预测结果对样本进行排序，把最可能是正样本个体排在前面，而后面的则是模型认为最不可能为正例的样本，再按此顺序逐个把样本作为正例进行预测并计算出当前的查准率和查全率得到的曲线



性能度量- F分数

- **F1度量**：查准率和查全率的调和平均，推荐系统常用

$$\frac{1}{F1}=\frac{1}{2}\left(\frac{1}{precision}+\frac{1}{recall}\right)$$

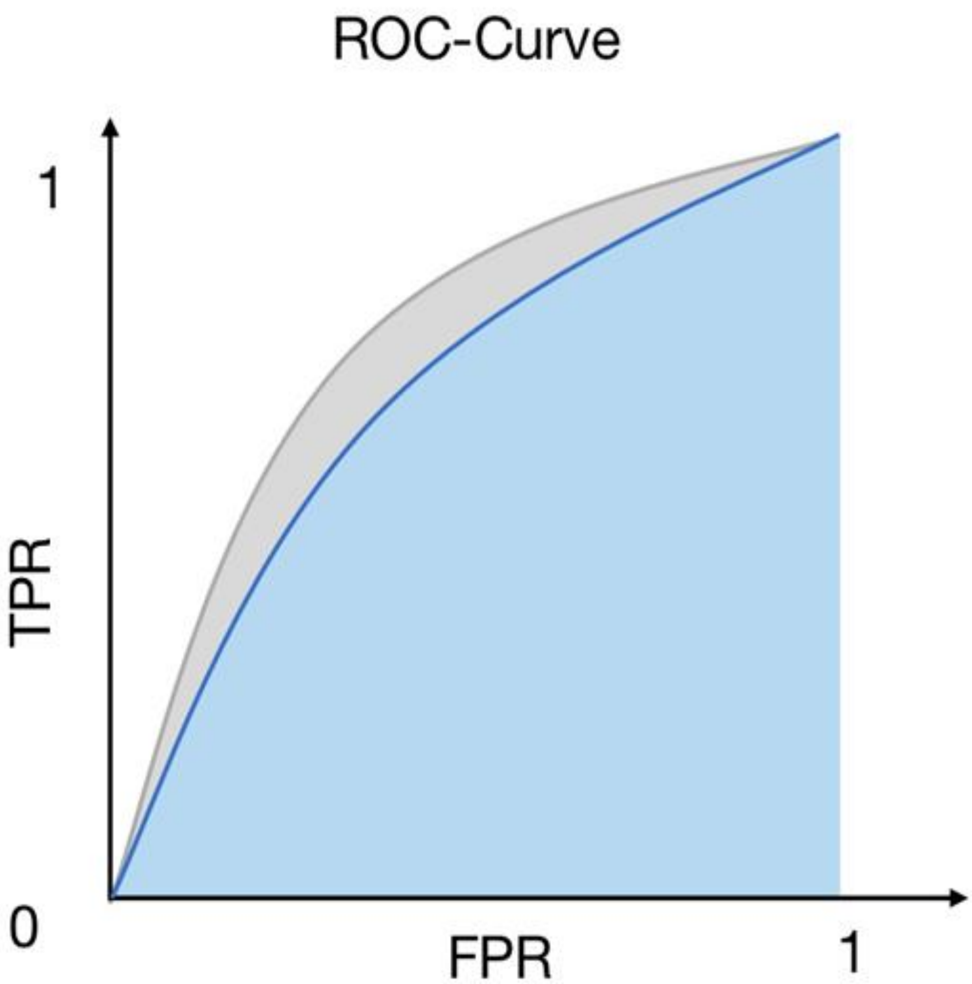
$$F1=\frac{2}{1/precision+1/recall}=\frac{2\times precision\times recall}{precision+recall}$$

- **F_β度量**： F1度量的一般形式，利用参数β控制查全率对查准率的相对重要性；β=1时，退化为F1；β>1时，查全率有更高大影响；β<1时，查准率有更高大影响

$$F_\beta=\frac{(1+\beta^2)\times precision\times recall}{\beta^2\times precision+recall}$$

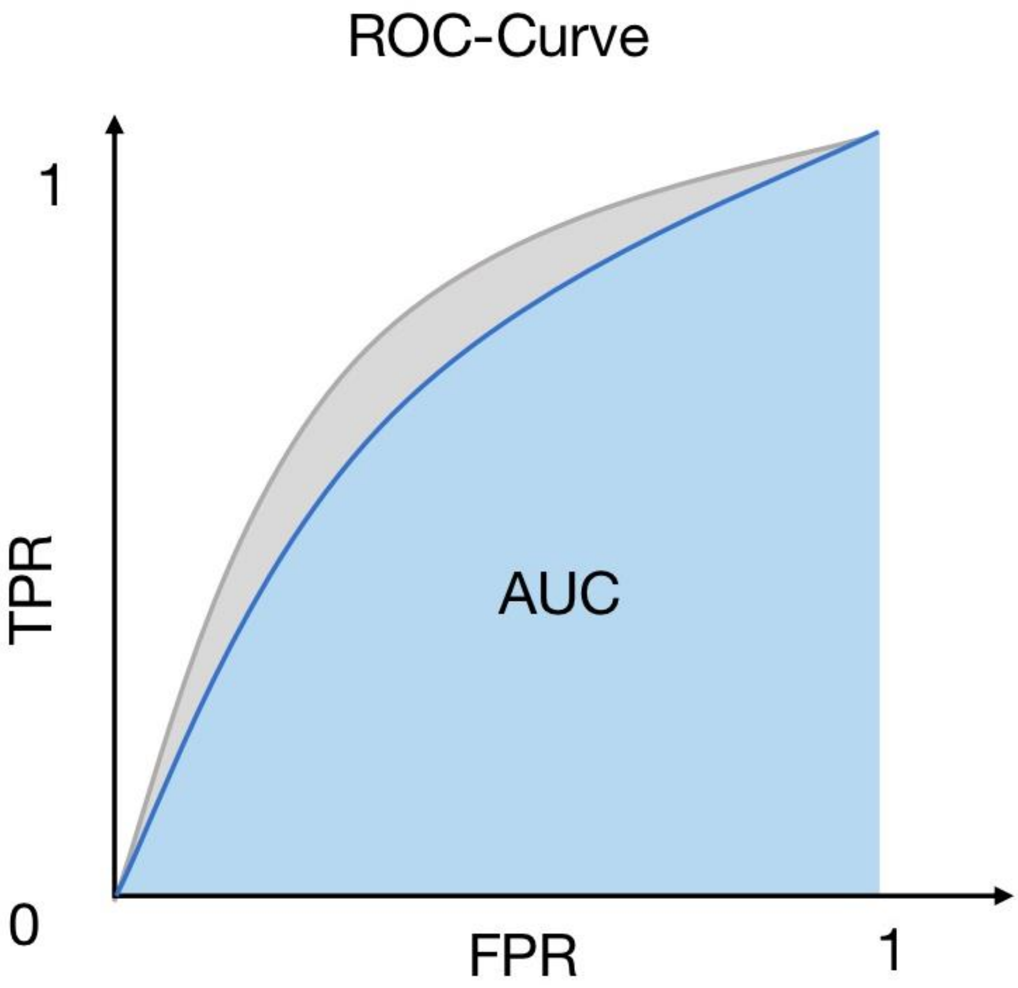
性能度量- ROC曲线

- **定义**：
  - 受试者工作特征曲线（receiver operating characteristic curve，简称ROC曲线）
  - 将预测结果按照预测为正类概率值排序
  - 将阈值由1开始逐渐降低，按此顺序逐个把样本作为正例进行预测，每次可以计算出当前的FPR，TPR值
  - 以TPR为纵坐标，FPR为横坐标绘制图像



性能度量- AUC

- **定义**：
  - AUC(Area Under Curve)即指曲线下面积占总方格的比例
  - 有时不同分类算法的ROC曲线存在交叉，因此很多时候用AUC值作为算法好坏的评判标准
  - 面积越大，表示分类性能越好



代价敏感性能度量

- **代价矩阵：** 描述不同错误的不同代价/风险

	预测结果	
真实结果	正例	反例
正例	0	Cost_{FN}
反例	Cost_{FP}	0

- **代价敏感错误率：**

$$E(f;D)=\frac{1}{d}\left(\sum_{\mathbf{x}_i\in D^+}\mathbb{I}(f(\mathbf{x}_i)\neq y_i)\times cost_{\text{FN}}\right.$$

$$\left.+\sum_{\mathbf{x}_i\in D^-}\mathbb{I}(f(\mathbf{x}_i)\neq y_i)\times cost_{\text{FP}}\right)$$

数学基础回顾

概率论基础

张永飞

概率统计

- **概率 (Probability)**

  对随机事件发生可能性大小的度量

- **联合概率 (Joint Probability)**

  A和B共同发生的概率，称事件A和B的联合概率，记作P(A, B)或P(A∩B)

- **条件概率 (Conditional Probability)**

  事件A已发生的条件下，事件B发生的概率，记作P(B|A)

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

概率统计

- **独立事件 (Independent Events)**

  事件A(或B)是否发生对事件B(或A)的发生概率没有影响，则称A和B为相互独立事件

- **条件独立 (Conditional Independence)**

  在给定C的条件下，若事件A和B满足：

$$P(A,B|C) = P(A|C) \cdot P(B|C)$$

或：

$$P(A|B,C) = P(A|C)$$

则称在给定C的情况下A和B独立

概率统计

- **乘法公式**
  - **设A，B为任意事件**

$$
\begin{aligned}
P(A,B) &= P(A|B)\cdot P(B) \\
&= P(B|A)\cdot P(A)
\end{aligned}
$$

  - **推广到n个事件的情况：**

$$
P(A_1A_2\ldots A_n) = P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1A_2)\cdot\ldots\cdot P(A_n|A_1A_2\ldots A_{n-1})
$$

$$
= \prod_{i=1}^{n} P(A_i|A_1A_2\ldots A_{i-1})
$$

概率统计

- **全概率公式 (Law of Total Probability)**
  - **设A1,A2,…,An两两互不相容，且**

$$B \subset A_1 + A_2 + \ldots + A_n$$

即B的发生总是与A1, A2,…,An之一同时发生，则对于事件B，有

$$P(B) = \sum_{k=1}^{n} P(A_k) P(B \mid A_k)$$

知因求果

贝叶斯公式

- **贝叶斯公式 (Bayes' Theorem)** 知果求因
  - 设$A_1,A_2,\ldots,A_n$两两互不相容，且 $B\subset A_1+A_2+\ldots+A_n$

即B的发生总是与$A_1,A_2,\ldots,A_n$之一同时发生，则在B已经发生的条件下，A_k的条件概率为

乘法公式

$$\boxed{P(A_k\mid B)=\frac{\boxed{P(A_kB)}}{P(B)}}=\frac{\boxed{P(A_k)P(B\mid A_k)}}{\boxed{\sum_{i=1}^{n}P(A_i)P(B\mid A_i)}}$$

乘法公式

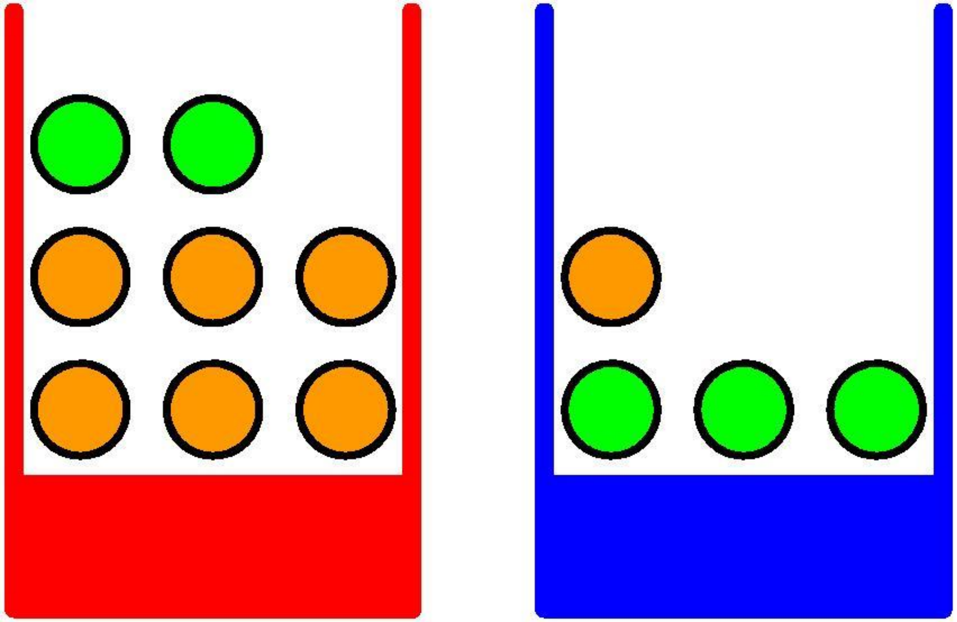
全概率公式

其中$P(A_k)$为先验概率；$P(A_k|B)$ 为后验概率

贝叶斯公式给出了“结果”事件B已经发生的条件下，“原因”事件A的条件概率，对结果的任何观测都将增加我们对原因事件A的真正分布的知识

概率练习

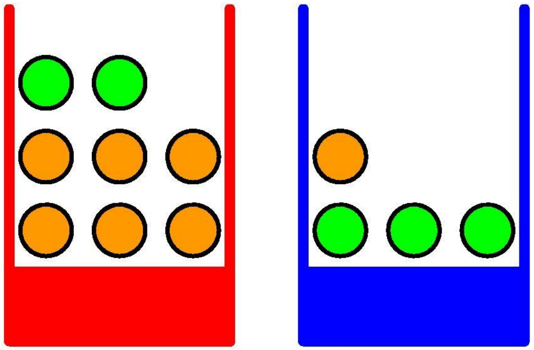
- 苹果和桔子
  - 选箱子事件变量B
  - 选水果事件变量F



$$P(B=r)=4/10$$
$$P(B=b)=6/10$$

概率练习

- 苹果和桔子



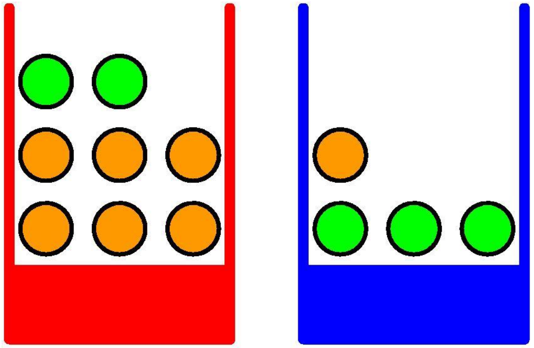
$$P(B=r)=4/10$$
$$P(B=b)=6/10$$

1.取到苹果的概率?

2.如果取到桔子，来自红箱子的概率?

概率练习

- 苹果和桔子



$$P(B=r)=4/10$$
$$P(B=b)=6/10$$

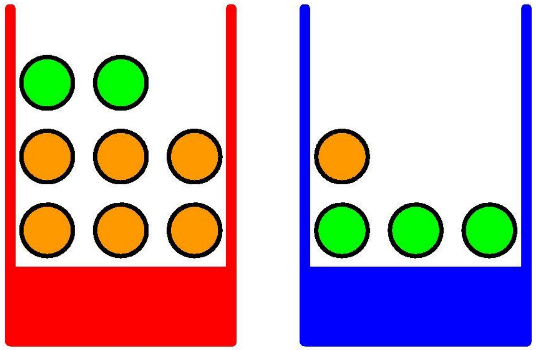
1.取到苹果的概率?

$$P(F=a|B=r)=1/4$$
$$P(F=o|B=r)=3/4$$
$$P(F=a|B=b)=3/4$$
$$P(F=o|B=b)=1/4$$

$$P(F=a)=P(F=a|B=r)P(B=r)+P(F=a|B=b)P(B=b)$$
$$=1/4\times 4/10+3/4\times 6/10=11/20$$

概率练习

- 苹果和桔子



$$P(B=r)=4/10$$
$$P(B=b)=6/10$$

2.如果取到桔子，来自红箱子的概率?

$$P(B=r|F=o)=\frac{P(F=o|B=r)P(B=r)}{P(F=o)}$$
$$=3/4\times 4/10\times 20/9=2/3$$

课外复习

概率论基础（完整版）

张永飞

随机事件

- **随机试验**
- **定义：** 为了研究随机现象，就要对研究对象进行观测或试验，这种观测或试验统称为随机试验

- **随机事件**
- **定义：** 一定条件下，可能发生也有可能不发生的试验结果称为随机事件，简称事件，通常用大写字母A, B等表示
- **两个极端情况：** 必然事件、不可能事件

随机事件

- **事件的关系**

1.事件的包含 $A \subset B$

2.事件的相等 $A = B$

3.事件的积（交） $A \cap B$

4.互不相容（互斥）事件 $A \cap B = \Phi$

5.事件的和（并） $A \cup B$

6.对立事件 $\overline{A}$

7.差事件 $A - B$

随机事件

- **事件间的运算**

1.交换律 $A \cup B = B \cup A$; $A \cap B = B \cap A$

2.结合律 $A \cup (B \cup C) = (A \cup B) \cup C$

$A \cap (B \cap C) = (A \cap B) \cap C$

3.分配律 $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

4.对偶原则 $\overline{A \cup B} = \overline{A} \cap \overline{B}$; $\overline{A \cap B} = \overline{A} \cup \overline{B}$

随机事件的频率

- 如果在相同的条件下，进行了n次试验。若随机事件A在这n次试验中发生了$r_n(A)$次,则比值

$$\frac{r_n(A)}{n}$$

称为事件A在这n次试验中发生的**频率**，记作$f_n(A)$，即

$$f_n(A) = \frac{r_n(A)}{n}$$

随机事件频率的特性

- 对任一事件A，0≤ f_n(A) ≤1；
- 对必然事件S，f_n(S)＝1；而 $f_n(\phi)=0$
- 可加性：若事件A、B互不相容，即A∩B＝ϕ，则 f_n(A ∪ B)＝ f_n(A) ＋f_n(B)。

一般地，若事件A1， A2 ，…， An两两互不相容，则

$$f_n\left(\bigcup_{i=1}^{n} A\right) = \sum_{i=1}^{n} f_n(A_i)$$

随机事件的概率

- **概率的统计定义**
  - 设随机事件A在n次重复试验中发生的次数为n_A，若当试验次数n很大时，频率n_A/n稳定地在某一数值p的附近摆动，且随着试验次数n的增加，其摆动的幅度越来越小，则称数p为随机事件A的概率，记为P(A)=p
  - 由定义，显然有

$$0 \le P(A) \le 1, \quad P(S)=1, \quad P(\phi)=0$$

随机事件的概率

● 概率的性质

① 非负性：$0 \le P(A) \le 1$

② 规范性：$P(S)=1$，$P(\varphi)=0$

③ 有限可加性：即若事件A1，A2，…，An两两互不相容，

则必有$P(A_1 \cup A_2 \cup \ldots \cup A_n)= P(A_1)+ P(A_2) +\ldots+ P(A_n)$

④ 单调不减性：设A，B是两个事件，则$P(A-B)=P(A)-P(AB)$

若A包含B，则AB=B，$P(A-B)=P(A)-P(B)$，且$P(A) \ge P(B)$

⑤ 互补性：对任一事件A，有 $P(\overline{A})=1-P(A)$

随机事件的概率

● 概率的性质

⑥ 加法公式：对任意两个事件A，B，有

$$P(A\cup B)=P(A)+P(B)-P(AB)$$

若A、B互斥，则 $P(A\cup B)=P(A)+P(B)$

⑦ 完备性：$A_1,A_2,\dots,A_n$两两互不相容，则有

$$P\left(\bigcup_{i=1}^{n}A_i\right)=\sum_{i=1}^{n}P(A_i)$$

⑧ 若$A_1,A_2,\dots,A_n$为完备事件集，则进一步有 $\sum_{i=1}^{n}P(A_i)=1$

⑨ 可分性：对任意两事件A，B，有 $P(A)=P(AB)+P(A\overline{B})$

随机事件的概率

● 概率-示例

例：某人外出旅游两天，据天气预报，第一天降水概率为0.6，第二天为0.3，两天都降水的概率为0.1，试求：

(1)“第一天下雨而第二天不下雨”的概率P(B)

(2)“第一天不下雨而第二天下雨”的概率P(C)

(3)“至少有一天下雨”的概率P(D)

(4)“两天都不下雨”的概率P(G)

(5)“至少有一天不下雨”的概率P(F)

随机事件的概率

- **概率-示例**
- 解 设Ai表示事件“第i天下雨”，i=1，2，由题意

  $P(A_1)=0.6$，$P(A_2)=0.3$，$P(A_1A_2)=0.1$

（1） $$B=A_1\overline{A}_2=A_1-A_1A_2$$

$$P(B)=P(A_1-A_1A_2)=P(A_1)-P(A_1A_2)=0.6-0.1=0.5$$

（2） $$P(C)=P(A_2-A_1A_2)=P(A_2)-P(A_1A_2)=0.3-0.1=0.2$$

随机事件的概率

- **概率-示例**

(3) $D = A_1 \cup A_2$

$$P(D) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1A_2)$$
$$= 0.6 + 0.3 - 0.1 = 0.8$$

(4) $G = \overline{A_1}\,\overline{A_2} = \overline{A_1 \cup A_2}$

$$P(G) = P(\overline{A_1 \cup A_2}) = 1 - P(A_1 \cup A_2) = 1 - 0.8 = 0.2$$

(5) $P(F) = P(\overline{A_1 \cap A_2}) = 1 - P(A_1A_2) = 1 - 0.1 = 0.9$

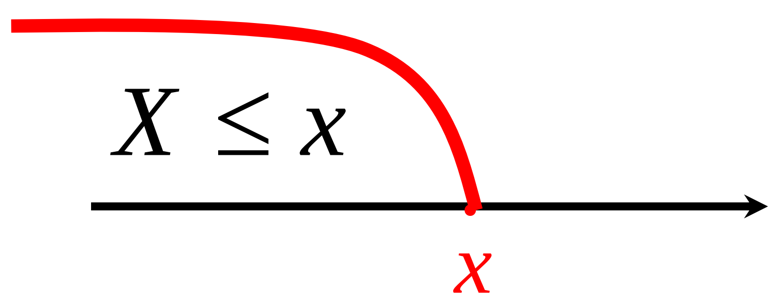
随机变量

- **定义**
- 对于随机试验E，Ω是其样本空间。如果对每一个样本点w，都对应着一个实数X(w)，则称Ω上的实值函数X(w)为随机变量，简记为X

概率分布函数

- **定义**
- 对于给定随机变量X，其取值X(w)不超过实数*x*的事件的概率P(X ≤*x*)是*x*的函数，称为X的概率分布函数，简称为分布函数，记为F(x)
- 即X的分布函数为

$$F(x) = P(X \le x), \qquad x \in (-\infty, +\infty)$$



概率分布函数

- **性质**
  - 单调不减　　$若 a<b, 则 F(a)\le F(b)$
  - 非负有界　　$0\le F(x)\le 1, (-\infty<x<+\infty), 且$

$$\lim_{x\to-\infty} F(x)=F(-\infty)=0,$$

$$\lim_{x\to+\infty} F(x)=F(+\infty)=1$$

  - 右连续　　$F(x+0)=F(x)$

P(a<x ≤b)=F(b)-F(a)

概率分布密度

- **离散型随机变量及其分布**
- 设x_k(k=1,2,…)是离散型随机变量X所取的一切可能值，p_k是X取值x_k的概率，则称

$$P(X=x_k)=p_k, \qquad k=1,2,\ldots$$

为离散型随机变量X的概率分布或分布律

X	x_1	x_2	…	x_k	…
p_k	p_1	p_2	…	p_k	…

概率分布密度

- **连续型随机变量及其分布**
- 设随机变量X的分布函数为$F(x)$，如果存在非负函数$f(x)$，使得对任意的实数x，都有

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

- 则称X为连续型随机变量，$f(x)$称为X的概率密度函数，简称为概率密度或分布密度

概率分布密度

- 概率密度的性质

$(1)\ f(x) \geq 0;$

$(2)\int_{-\infty}^{+\infty} f(x)dx = 1;$

$(3)P(a < X \leq b) = \int_{a}^{b} f(x)dx;$

$(4)P(X = x) = 0;$

$(5)F(x)$是连续函数,若$f(x)$在x_0连续,有

$$F'(x_0) = f(x_0).$$

概率分布密度

- 常见随机变量及其分布
- **0-1分布**
  - 若随机变量X只可能取0和1两个值，其概率分布为
    $P(X=1)=p$，$P(X=0)=1-p\quad(0<p<1)$
    则称X服从参数为p的0-1分布
- **二项分布**
  - 若随机变量X的概率分布为

$$P_n(k)=P(X=k)=C_n^k p^k q^{n-k},$$

$$k=0,1,2,\ldots,n;0<q=1-p<1,$$

  - 称X服从参数为n和p的二项分布，记作$X\sim B(n,p)$

概率分布密度

- 常见随机变量及其分布
- **泊松分布**
  - 若随机变量X的概率分布为

$$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k=0,1,2,\ldots,$$

  - 其中常数$\lambda>0$,则称X服从参数为λ的泊松分布，记作$X \sim P(\lambda)$

概率分布密度

- 常见随机变量及其分布
- **几何分布**
  - 在独立试验序列中，若一次贝努利试验中某事件A发生的概率为$P(A)=p$，只要事件A不发生, 试验就不断地重复下去，直到事件A发生，试验才停止。设随机变量X为直到事件A发生为止所需的试验次数，X的概率分布为

$$P(X=k)=(1-p)^{k-1}p, \qquad k=1,2,\ldots,$$

则称X服从参数为p的几何分布，记作$X\sim G(p)$.

概率分布密度

- 常见随机变量及其分布
- **均匀分布**
  - 若随机变量X的概率密度为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b,\\ 0, & \text{其他},\end{cases}$$

  - 则称X服从区间$[a,b]$上的均匀分布,记作$X\sim U[a,b]$.

$$F(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & a\le x<b,\\ 1, & x\ge b.\end{cases}$$

概率分布密度

- 常见随机变量及其分布
- **指数分布**
- 若随机变量X的概率密度为

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0,\\ 0, & x\le 0,\end{cases}$$

- 其中,$\lambda>0$为常数,则称X服从参数为λ的指数分布,记作$X\sim E[\lambda]$.

$$F(x)=\begin{cases}1-e^{-\lambda x}, & x>0,\\ 0, & x\le 0.\end{cases}$$

概率分布密度

- 常见随机变量及其分布
- **正态分布**
- 若随机变量X的概率密度为

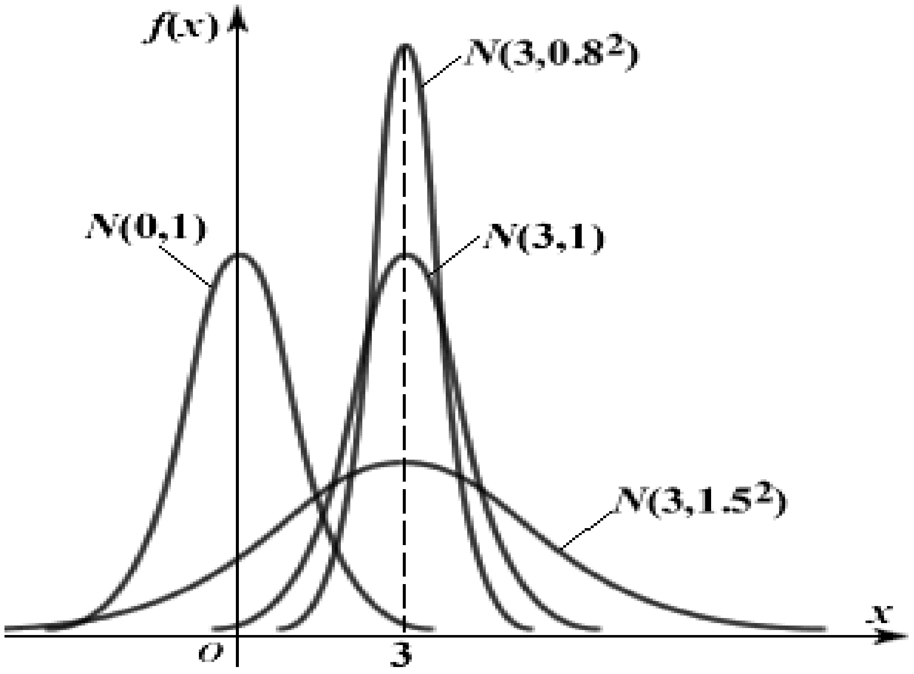
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad -\infty<x<\infty$$

- 其中μ和σ都是常数,$\sigma>0$，则称X服从参数为μ和σ^2的正态分布.记作$X\sim N(\mu,\sigma^2)$

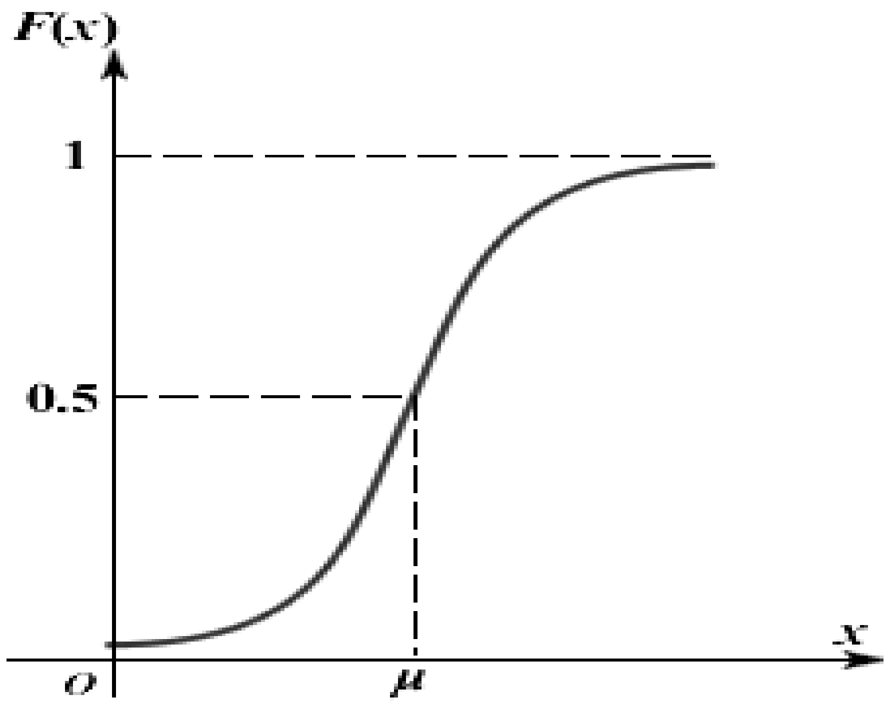
$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt,\quad -\infty<x<\infty$$

概率分布密度

- 常见随机变量及其分布
- **正态分布**



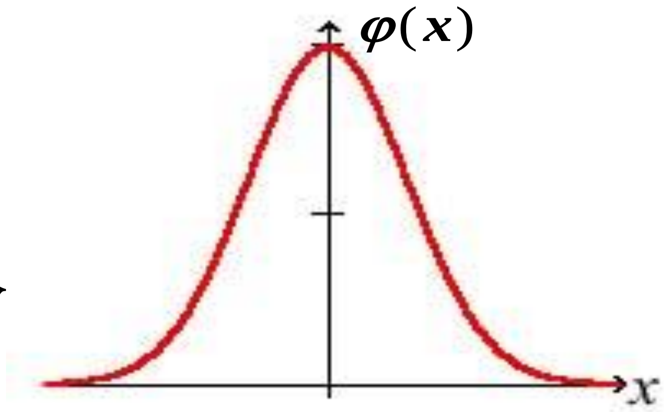
正态分布的概率密度图形



正态分布的分布函数图形

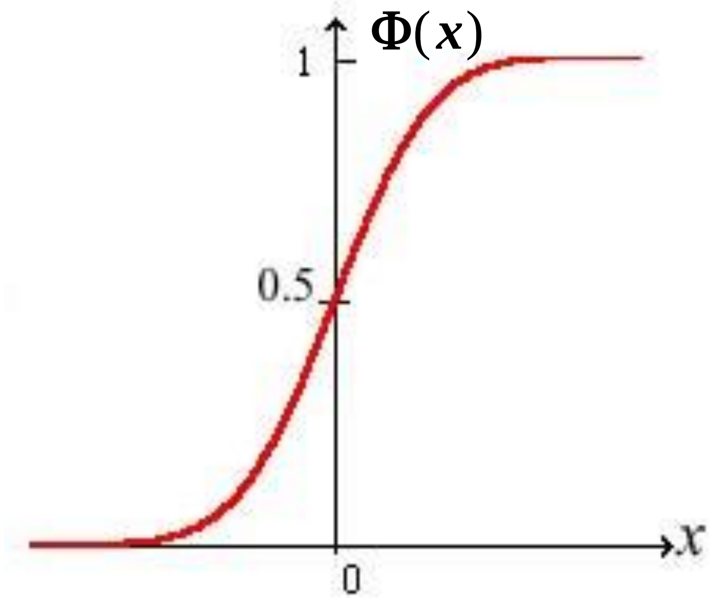
概率分布密度

- 常见随机变量及其分布
- **标准正态分布**
- μ=0, σ=1时的正态分布称为标准正态分



$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\quad -\infty<x<\infty$$

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$



二维随机变量及其概率

- **二维随机变量**：对于随机试验E，Ω是其样本空间。X(w)和Y(w)是定义在样本空间Ω上的两个随机变量，由它们构成的向量(X,Y)称为二维随机变量或二维随机向量
- **联合分布函数**：设(X,Y)是二维随机变量，对于任意实数x, y，称二元函数

$$F(x,y)=P(X\leq x,Y\leq y)$$

为二维随机变量(X,Y)的联合分布函数，简称分布函数

二维随机变量及其概率

- **联合概率密度**：设二维随机变量 (X,Y)的分布函数为$F(x,y)$，如果存在非负函数$f(x,y)$使得对任意的实数x, y,都有

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dvdu$$

- 则称$f(x,y)$ 为连续型随机变量(X,Y)的联合概率密度函数，简称联合概率密度或联合分布密度

条件概率

- 设 A、B是随机试验E的两个随机事件，且$P(A)>0$，则称

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

为已知事件A发生条件下，事件B发生的条件概率

- 若事件 A、B相互独立，则

$$P(AB)=P(A)P(B)$$

乘法公式

- 设A，B为任意事件，

$$P(AB)=P(A)P(B|A)$$

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

- 推广到n个事件的情况：

$$P(A_1A_2\ldots A_n)=P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1A_2)\cdot\ldots\cdot P(A_n|A_1A_2\ldots A_{n-1})$$

$$=\prod_{i=1}^{n}P(A_i|A_1A_2\ldots A_{i-1})$$

全概率公式

- 设A1,A2,⋯,An两两互不相容，且

$$B \subset A_1 + A_2 + \ldots + A_n$$

即B的发生总是与A1, A2,⋯,An之一同时发生，则对于事件B，有

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B \mid A_k)$$

贝叶斯公式

- 设$A_1,A_2,\ldots,A_n$两两互不相容，且

$$B \subset A_1 + A_2 + \ldots + A_n$$

即B的发生总是与$A_1,A_2,\ldots,A_n$之一同时发生，则在B已经发生的条件下，A_k的条件概率为

$$P(A_k \mid B) = \frac{P(A_k B)}{P(B)} = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}$$

其中$P(A_k)$为先验概率；$P(A_k|\mathrm{B})$ 为后验概率

随机变量的数字特征

- **也称为随机变量的统计特征或统计量**
  - 数学期望
  - 方差
  - 矩
  - 协方差
  - 相关系数

数学期望

- **数学期望的定义——离散型**
  - 设X是离散型随机变量，它的分布律是:

$$P(X = x_k) = p_k, \qquad k = 1,2,\ldots$$

如果级数$\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛,则称

$$E(X) = \sum_{k=1}^{\infty} x_k p_k$$

为X的数学期望

数学期望

- **数学期望的定义——连续型**
  - 设X是连续型随机变量，其密度函数为 $f(x)$，如果

$$\int_{-\infty}^{\infty} x f(x)dx$$

绝对收敛，则定义X的数学期望为

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

数学期望

- **随机变量函数的数学期望**
  - 设Y是随机变量X的连续函数，$Y=g(X)$，则

$$E(Y)=E[g(X)]=\begin{cases}\sum_{k=1}^{\infty} g(x_k)p_k, & X\text{离散型}\\ \int_{-\infty}^{\infty} g(x)f(x)dx, & X\text{连续型}\end{cases}$$

  - 当X为离散型时，$P(X=x_k)=p_k$；
  - 当X为连续型时，X的密度函数为$f(x)$.
  - 数学期望描述了随机变量的平均值

数学期望

- **数学期望的性质**

1. $E(aX+b)=aE(X)+b$;

$\Rightarrow$ $E(aX)=aE(X)$ $\quad$ $E(b)=b$

2. $E(X+Y) = E(X)+E(Y)$;

$\Rightarrow$ $$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E(X_i)$$

$$E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E(X_i)$$

3. 设X、Y相互独立，则 $E(XY)=E(X)E(Y)$。

方差

- **方差的定义**

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

$$= \begin{cases} \sum_{k=1}^{\infty}[x_k - E(X)]^2 p_k & X\text{离散型} \\ \int_{-\infty}^{\infty}[x - E(X)]^2 f(x)dx & X\text{连续型} \end{cases}$$

$$D(X)=E(X^2)-[E(X)]^2$$

- 方差描述了随机变量取值距离其平均值(即数学期望)的分散程度

方差

- **方差的性质**

1. $D(aX+b)=a^2D(X)$；

$\Rightarrow$ $D(aX)=a^2D(X)$

$D(b)=0$

$D(-X)=D(X)$

方差

- **方差的性质**
- 2. 若X、Y相互独立, $D(X+Y) = D(X)+D(Y)$;

⇒

$$若X,Y相互独立, D(aX+bY)=a^2D(X)+b^2D(Y)$$

$$若X,Y相互独立, D(X-Y)=D(X)+D(Y)$$

$$若X_1,X_2,\ldots,X_n相互独立, D[\sum_{i=1}^{n}a_iX_i]=\sum_{i=1}^{n}a_i^2D(X_i)$$

方差

- **方差的性质**
- 3.切比雪夫不等式

设随机变量X有数学期望μ和方差σ^2，则对于任给$\varepsilon>0$,有

$$P\{|X-\mu|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2}$$

即 $$P\{|X-\mu|<\varepsilon\}\geq 1-\frac{\sigma^2}{\varepsilon^2}$$

随机变量“标准化”

- **标准化**

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}}$$

称X^*为X的标准化随机变量

$$E(X^*)=0, \qquad D(X^*)=1$$

常见分布的数学期望与方差

- 若$X \sim$ 0-1分布，那么$E(X)=p$， $D(X)=p(1-p)$;
- 若$X \sim B(n,p)$，那么$E(X)=np$，$D(X)= np(1-p)$;
- 若$X \sim P(\lambda)$，那么$E(X)=\lambda$，$D(X)=\lambda$;
- 若$X \sim G(p)$，那么$E(X)=1/p$，$D(X)=(1-p)/p^2$;
- 若$X \sim U[\mathrm{a},\mathrm{b}]$, 那么$E(X)=(a+b)/2$, $D(X)=(\mathrm{b}-\mathrm{a})^2/12$;
- 若$X \sim E(\lambda)$, 那么$E(X)=1/\lambda$, $D(X)=1/\lambda^2$;
- 若$X \sim N(\mu,\sigma^2)$, 那么$E(X) =\mu$, $D(X) =\sigma^2$.

矩

- $E(X^k)$——X的k阶原点矩
- $E\{[(X-E(X)]^k\}$——X的k阶中心矩

- $E(X)$——X的1阶原点矩
- $D(X)$—— X的2阶中心矩

协方差与相关系数

- **协方差**

设(X,Y)为二维随机变量，若

$$E\{[X-E(X)][Y-\mathrm{E}(Y)]\}$$

存在，则称其为X和Y的协方差，记为$\mathrm{cov}(X,Y)$。

- **相关系数/标准协方差**

设$D(X)>0$, $D(Y)>0$，称

$$\rho_{XY}=\frac{\mathrm{cov}(X,Y)}{\sqrt{D(X)D(Y)}}$$

为随机变量X和Y的相关系数

相关系数

- 性质

1 $\text{cov}(X,Y)=\text{cov}(Y,X)$

2 $\text{cov}(aX,bY)=ab\,\text{cov}(X,Y)$ a,b是常数

3 $\text{cov}(X_1+X_2,Y)=\text{cov}(X_1,Y)+\text{cov}(X_2,Y)$

4 $|\rho|\leq 1 \Longleftrightarrow$ 存在常数$a,b(a\neq 0)$,使$P(Y=aX+b)=1$

5 X与Y相互独立 $\Longrightarrow$ X与Y不相关