机器学习
Machine Learning

北京航空航天大学计算机学院
School of Computer Science and Engineering, Beihang University
刘庆杰　陈佳鑫

2025年春季学期
Spring 2025

部分内容来源于C. Bishop和A. NG等人的课程以及互联网资源

第2章: 预备知识

Chapter 2: Basics of Machine Learning

内容提要

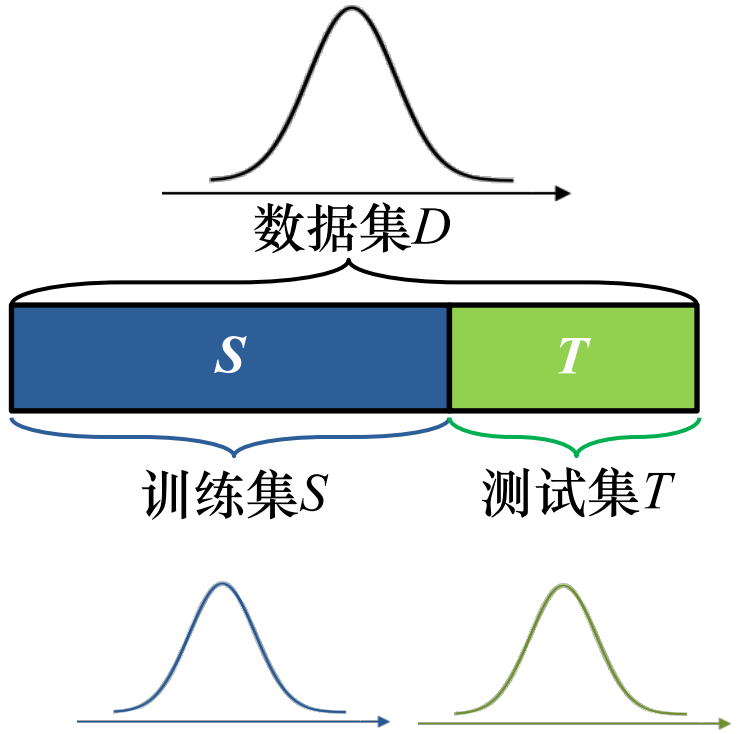
- 2.1 数据集的划分方法
- 2.2 模型的性能度量
- 2.3 概率统计基础
- 2.4 贝叶斯决策理论
- 2.5 参数化概率密度估计方法
- 2.6 非参数概率密度估计方法
- 2.7* 矩阵理论基础（拓展）

2.1 数据集的划分方法

- 数据集划分的基本要求
- 数据集的划分方法

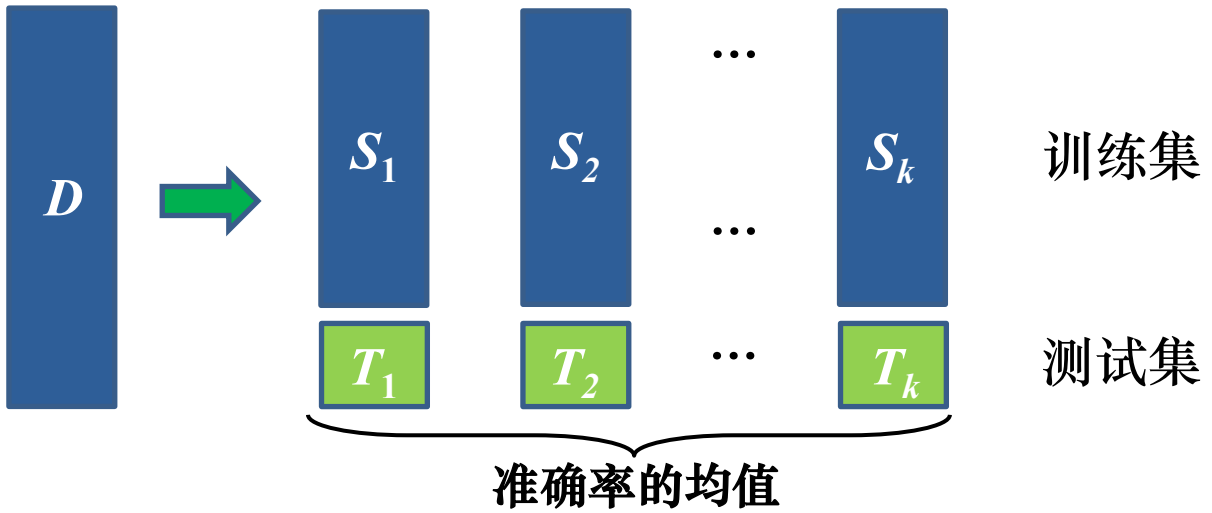
数据集划分的基本要求

- **模型评估**
  - 1. **数据集划分**：分为训练集和测试集两部分
    - 原则：测试集应尽可能与训练集**互斥**，测试样本不在训练集中出现
    - 目标：将数据集D划分为训练集S和测试集T两部分，在训练集上建立模型，在测试集上评估性能
    - 假设：测试样本从原样本真实分布中独立同分布采样得到
    - 方法：留出法、自助法、交叉验证法
  - 2. 性能度量：模型在测试集（新样本）上进行度量，也叫泛化性能



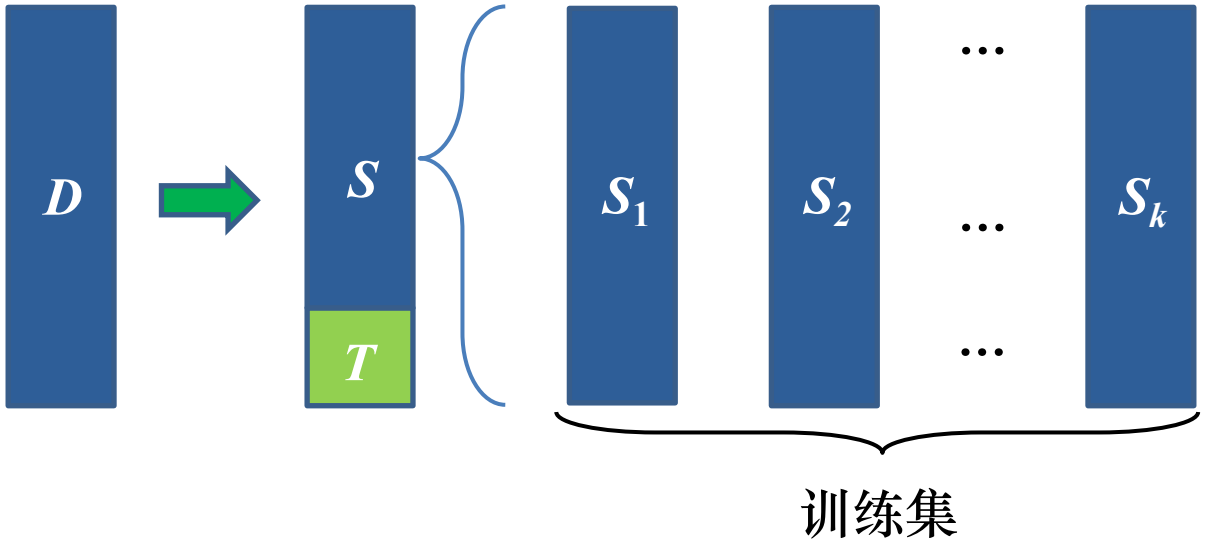
数据集的划分

- **留出/保持法(hold-out)：**
  - 在数据量充足的情况下，通常使用留出法
  - 在数据集D上通过随机采样的方式采得训练集，其余作为测试集（通常，2/3的数据分配到训练集）。使用训练集训练模型，使用测试集来估计泛化性能
- **随机子抽样(random sub-sampling)：**
  - 随机地将留出法重复k次，分出k组训练集和测试集，总准确率取每次计算准确率的均值



数据集的划分

- **自助法(bootstrapping):**
  - 与保留法相同是随机采样；不同在于自助法从初始样本D有放回地均匀抽样；每当选中一个样本，其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；



数据集的划分

- **自助法(bootstrapping):**
  - **样本在$|D|$次采样中始终不被采到的概率:**

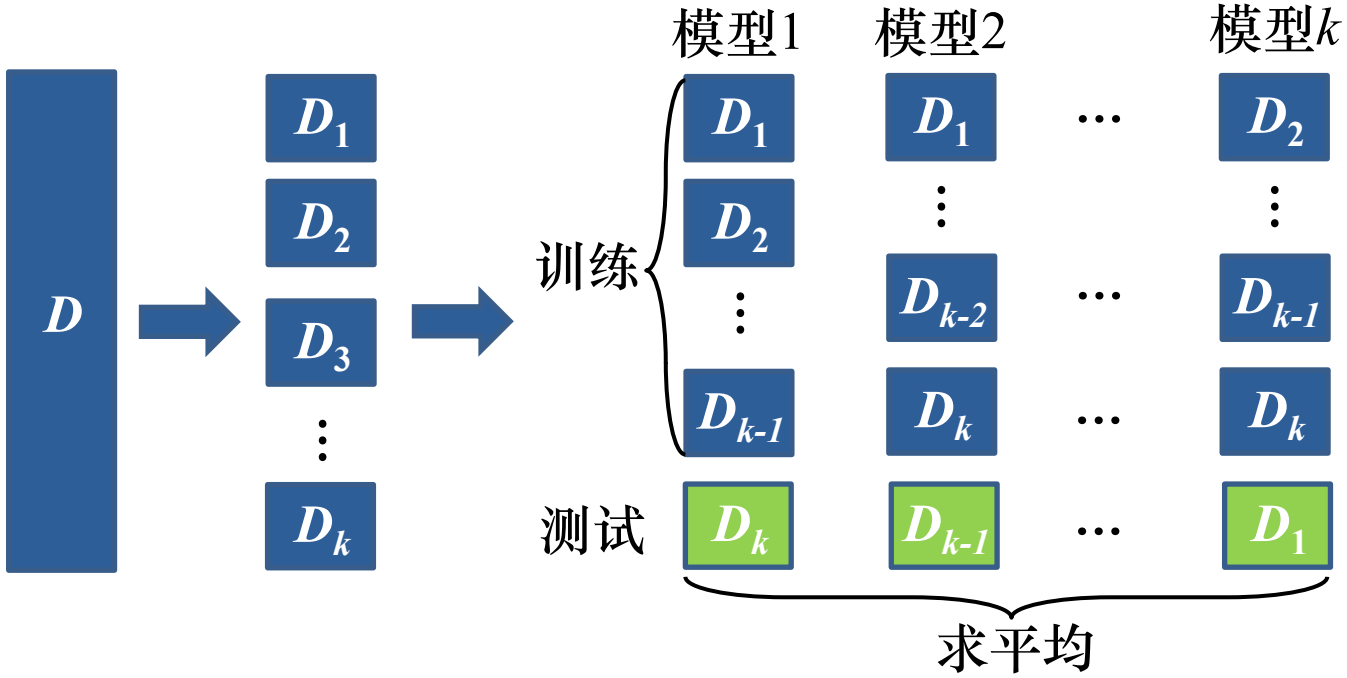
$$\lim_{|D|\to\infty}\left(1-\frac{1}{|D|}\right)^{|D|} \mapsto \frac{1}{e}\approx 0.368$$

  - **将未被采到的样本作为测试集**
  - **集成学习中最具代表性的Bagging算法就是基于自助法采样设计的**

    【具体细节在第14.3节Bagging算法中详细讲解】

数据集的划分

- **(k折)交叉验证(k-fold cross-validation)**
  - 在数据量相对较少的情况下，通常使用交叉验证法
  - 在将初始数据集D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，取k次测试结果的均值
- **留一法(leave-one-out)**
  - 将每个样本作为一“折”
  - k折交叉验证法的特殊情况

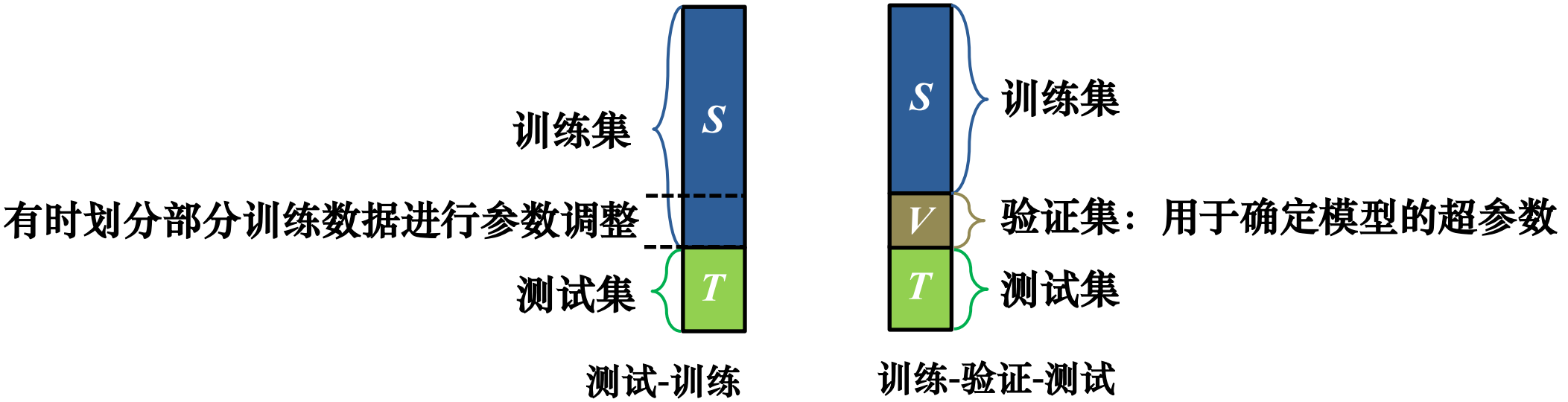


本节小结

方法	适用场景	方差	偏差
留出法	数据量较大	训练、测试集的分布不同时，会导致模型方差较高	当数据量不足时，会导致模型偏差较高
自助法	数据量较小	模型方差较低	改变了初始分布，会导致模型偏差较高
交叉验证法	数据量较小	模型方差较低	模型偏差较低

扩展知识

- **训练/测试集（Train-Test）和训练/验证/测试集（Train-Val-Test） 划分:**



  - 在ImageNet、COCO等大规模数据集上常使用“训练-验证-测试”划分

- **训练集和测试集的一些限制:**

  - 一些任务（例如时间序列）在划分数据集时要求确保测试集中的数据时间上永远不会早于训练集

2.2 模型的性能度量

- 模型的性能度量
- 分类任务性能度量
- 回归任务性能度量

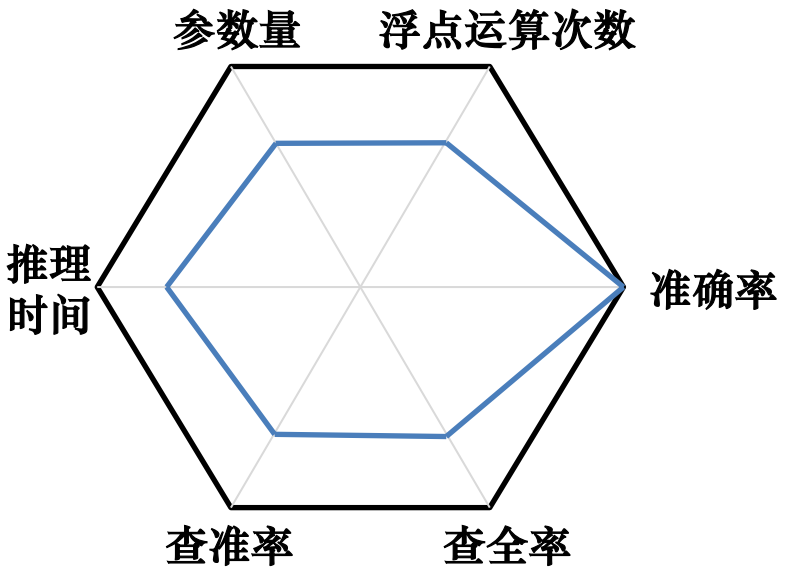
模型的性能度量

- 模型评估
  - 1. 数据集划分：分为训练集和测试集两部分
  - 2. **性能度量**：性能度量：模型在测试集（新样本）上进行度量，也叫泛化性能
    - 常用性能度量：错误率和精度（分类任务）
    - 仅能评估是否正确分类，无法提供更全面的评估

示例1：发动机合格检测

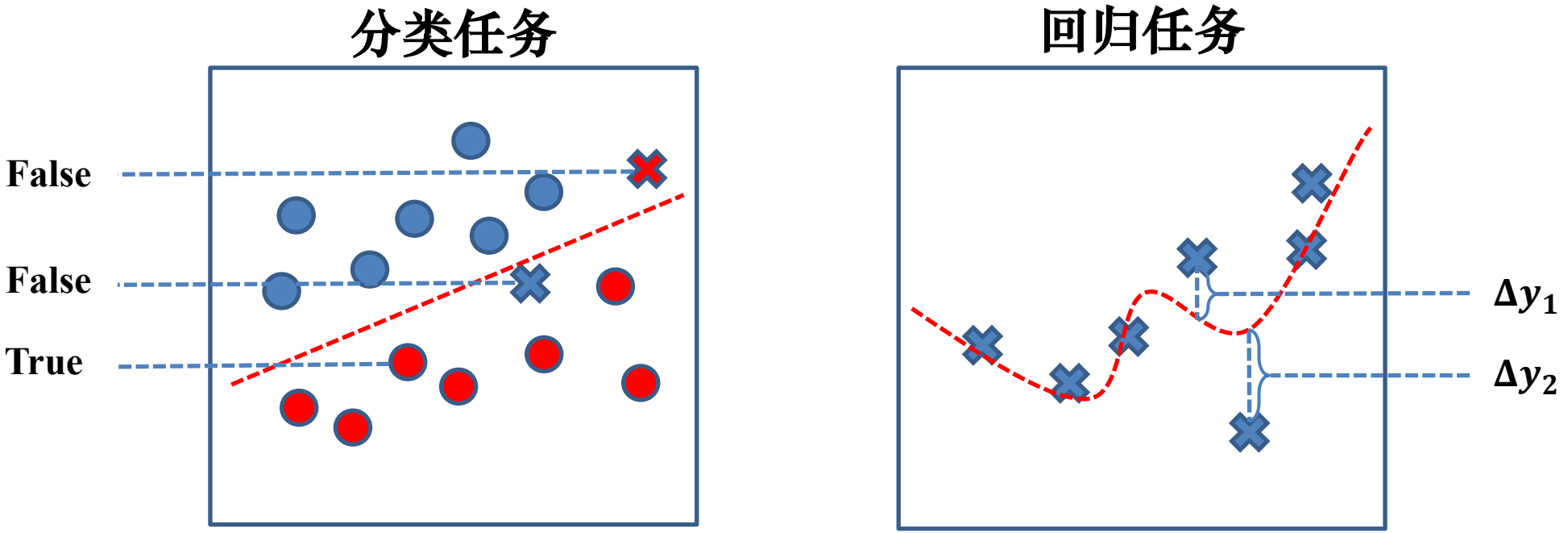
精度可以评估“检测出有多少发动机是合格的”

无法评估“检测出的合格发动机有多少是真正合格的”



模型的性能度量

- 分类和回归任务的不同指标



- 对于单一样本，分类任务的输出是正确或者错误（二分类），仅定性评价分类是否正确
- 对于单一样本，回归任务的输出是连续数值，需要定量反馈指标来评估数值预测的精度

分类任务性能度量

- 错误率/精度仅能粗略评估模型
  - 只能评估“检测出有多少发动机是合格的”
  - 无法评估更细节的问题，如
    - 有多少检测合格的是真正合格的
    - 有多少检测不合格的是真正不合格的
    - 有多少检测合格的是实际不合格的
    - 有多少检测不合格的是实际合格的
  - 引入展现更多细节的相关要素：
    - TP：被分类器正确分类的正样本；真值为T，预测为T：称为真正
    - TN：被分类器正确分类的负样本； 真值为F，预测为F：称为真负
    - FP：被错误标记为正元组的负样本；真值为F，预测为T：称为假正
    - FN：被错误标记为负元组的正样本；真值为T，预测为F：称为假负

分类任务性能度量

- 混淆矩阵

		预测结果	
真实结果		正例	反例
	正例	真正例TP (True Positive)	假负例FN (False Negative)
	反例	假正例FP (False Positive)	真负例TN (True Negative)

- TP、FN、FP、TN共同组成了混淆矩阵
- 主对角线元素表示正确分类个数，非对角线上的元素表示未被正确分类个数

分类任务性能度量

- **查准率(Precision)**
  - 评估预测正样本中的真正样本 $precision = \frac{TP}{TP + FP}$
- **查全率(Recall)**
  - 评估真正样本中被正确预测的样本 $recall = \frac{TP}{TP + FN}$
- **查准率和查全率相互制约，在实际使用需要平衡**

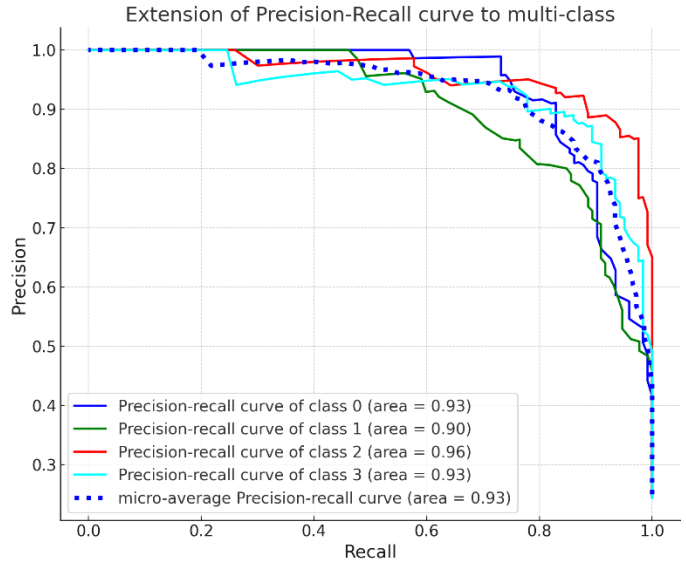
示例：发动机合格检测

若提高模型的合格阈值，模型评价的合格发动机变少，查准率升高，查全率降低

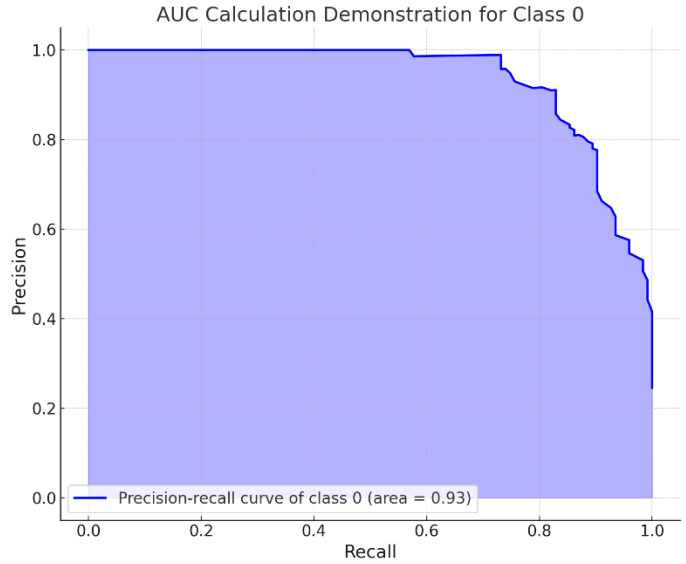
若降低模型的合格阈值，查全率上升，查准率降低，极限情况查全率达到1

分类任务性能度量

- **Precision-Recall（P-R）曲线**
  - 查全率Recall为横轴，查准率Precision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，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查准率和查全率得到的曲线



- **Area Under the PR Curve（AUC）曲线下面积**
  - AUC-PR 是指PR曲线下面积，用来量化模型的分类性能，AUC值的范围是0到1，AUC值越接近1，模型的性能越好



分类任务性能度量

- 在医学诊断、生物统计学、检测系统、安全系统等领域
  - 准确率
    - 评估分类器正确识别正/负样本的能力 $accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{(TP + FN) + (TN + FP)}$
  - 错误率
    - 评估分类器错误识别正/负样本的能力 $ER = \frac{FP + FN}{P + N} = \frac{FP + FN}{(TP + FN) + (TN + FP)}$
  - 敏感性(Sensitivity)
    - 评估分类器正确识别正样本的能力，与查全率相同 $SN = \frac{TP}{P} = \frac{TP}{\mathrm{TP} + FN}$
  - 特异性(Specificity)
    - 评估分类器正确识别负样本的能力，查准率衡量的是分类器正确识别正样本的能力 $SP = \frac{TN}{N} = \frac{TN}{\mathrm{TN} + FP}$

分类任务性能度量

- **F_1度量**
  - 查准率和查全率的调和值，推荐系统常用

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

- **F_β度量**
  - F_1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

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

回归任务性能度量

- 平均绝对误差(Mean Absolute Error, MAE)

$$E(f;D)=\frac{1}{d}\sum_{i=1}^{d}|f(\mathbf{x}_i)-y_i|$$

  - MAE 衡量预测值与实际值之间的平均绝对差异
  - 特点：
    - 计算简单：直接测量预测值与实际值之间的绝对差异
    - 敏感性不足：对于一些相对较小的误差不够敏感，不足以揭示预测错误的严重性
    - 不可微分：MAE 在零误差点处不可微分，可能影响某些优化算法的性能

回归任务性能度量

- **均方误差(Mean Squared Error, MSE)**

$$E(f;D)=\frac{1}{d}\sum_{i=1}^{d}(f(\mathbf{x}_i)-y_i)^2$$

  - MSE衡量预测值与实际值之间的平均平方差异，平方差的计算使得误差被放大
  - 特点：
    - 对误差敏感：MSE 会放大误差，使其在优化时更具影响力
    - 可微分：MSE是可微的，这对很多优化算法（如梯度下降）是有利的
    - 对离群点敏感：MSE对离群点非常敏感，较大误差的平方会显著影响整体误差度量

回归任务性能度量

- **均方根误差(RMSE)**

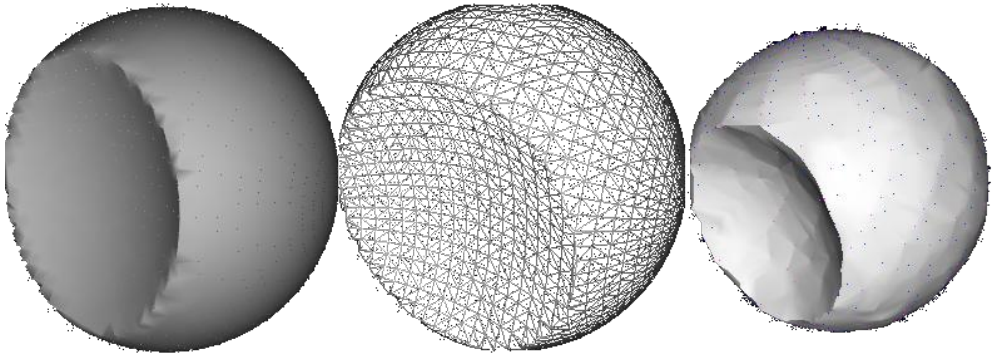
$$E(f;D)=\sqrt{\frac{1}{d}\sum_{i=1}^{d}(f(\mathrm{x}_i)-y_i)^2}$$

  - RMSE 是 MSE 的平方根，表示预测值与实际值之间的标准偏差
  - 特点：
    - 对误差敏感：与MSE类似，RMSE对误差有一定程度的敏感性
    - 依旧对离群点敏感：由于RMSE是MSE的平方根，仍然对离群点敏感，对于包含大量离群点的数据集，RMSE依旧不如 MAE 鲁棒

回归任务性能度量扩展

图像重建

点云重建



- 通常不直接对所有特征计算偏差，而是通过特征提取后计算特征间相似性
- 结构相似性指数（Structural Similarity Index, SSIM）
  - 一种感知为主的图像质量评估指标，用于衡量重建图像与原始图像在结构信息上的相似性
- 感知相似性（Learned Perceptual Image Patch Similarity, LPIPS）
  - 衡量高维特征空间中的相似性，提取特征后在特征空间计算距离，更接近人类视觉系统的感知

2.3 概率统计基础

- 一些基本概念
- 贝叶斯公式

一些基本概念

- **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B已发生的条件下，事件A发生的概率，记作P(A|B)

一些基本概念

- **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条件独立，记为$A \perp\!\!\!\perp B|C$

一些基本概念

- **乘法原理**

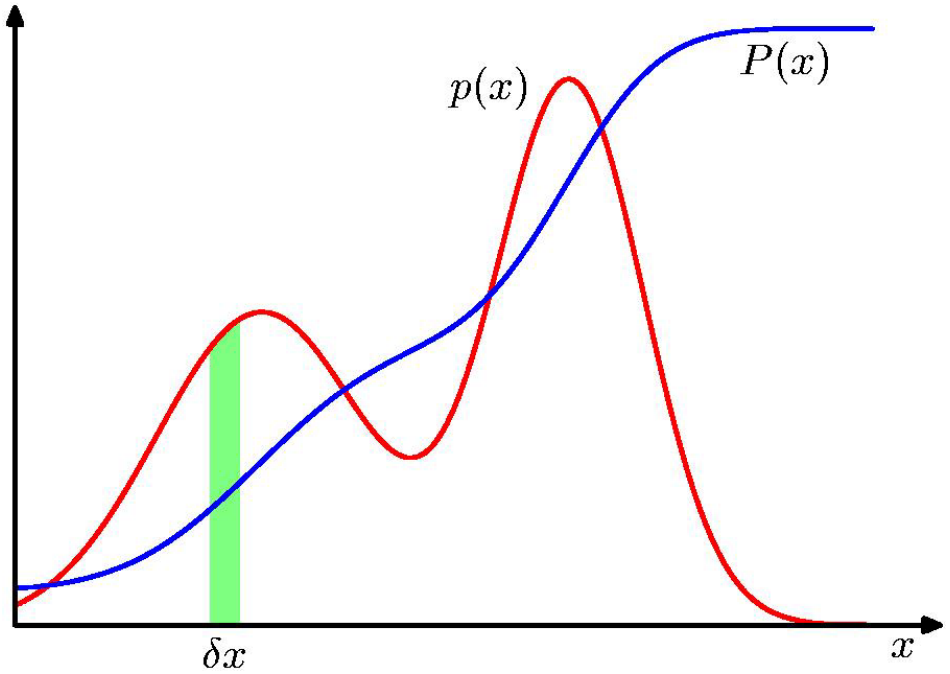
$$P(A,B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

- **全概率公式 (Law of Total Probability)**
  - A为任一事件，$B_1, B_2, \ldots, B_n$ 是样本空间Ω的一个划分，且B_i两两互斥，$P(B_i)>0, (i=1, 2, \ldots, n)$，那么：

$$P(A) = \sum_{i=1}^{n} P(A,B_i) = \sum_{i=1}^{n} P(A|B_i)\,P(B_i)$$

连续型随机变量

- **概率密度函数 (Probability Density Function)**
  - 是描述随机变量输出值在某确定取值点附近可能性的函数
  - 随机变量的取值落在某个区域之内的概率为概率密度函数在这个区域上的积分；当概率密度函数存在的时候，累积分布函数是概率密度函数的积分



$$p(x) \geq 0$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

$$P\big(x \in (a,b)\big) \geq \int_{a}^{b} p(x)dx$$

$$P(z) \geq \int_{-\infty}^{b} p(x)dx$$

期望

- **期望 (Expectation)**
  - 期望是一个随机变量所取值的概率平均
  - 离散变量

$$E[X] = \sum_{k=1}^{\infty} x_k p_k$$

  - 连续变量

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

方差

- **方差 (Variance)**
  - 随机变量的方差描述的该随机变量的值偏离其期望值的程度
  - 离散变量

$$Var(X) = E\left(\left(X - E(X)\right)^2\right)$$
$$= \sum_{k=1}^{\infty}\left(x_k - E(X)\right)^2 p_k = E\left(X^2\right) - E^2(X)$$

  - 连续变量

$$Var(X) = \int_{-\infty}^{+\infty}\left(x - E(x)\right)^2 f(x)dx$$

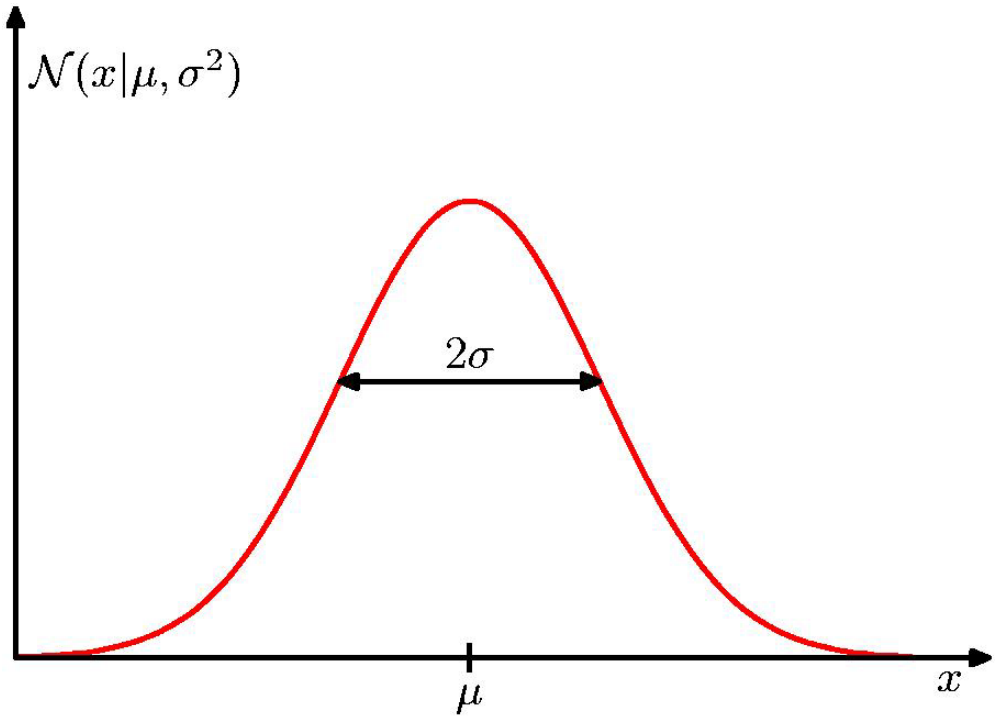
- **标准差 (Standard Deviation)**

$$\sigma(X) = \sqrt{Var(X)}$$

示例1：一维高斯分布

- 概率密度函数：

$$\mathcal{N}\left(x|\mu,\sigma^2\right)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



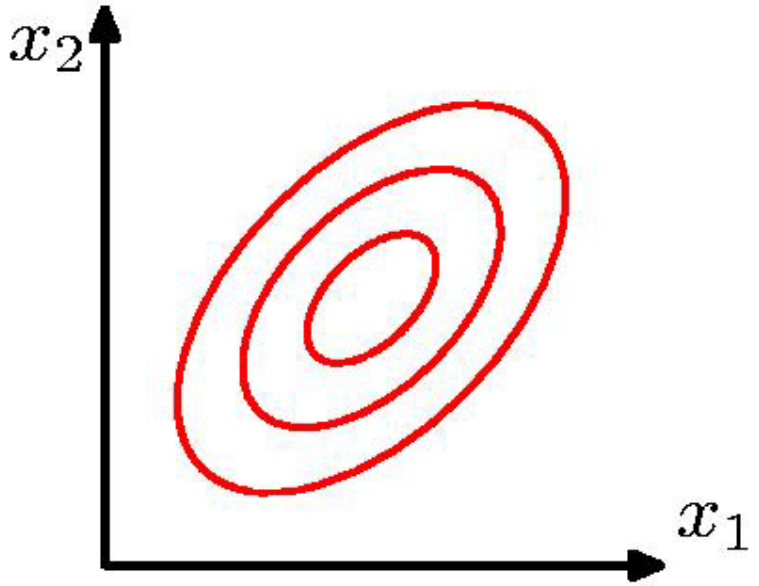
$$\mathcal{N}\left(x|\mu,\sigma^2\right)\geq 0$$

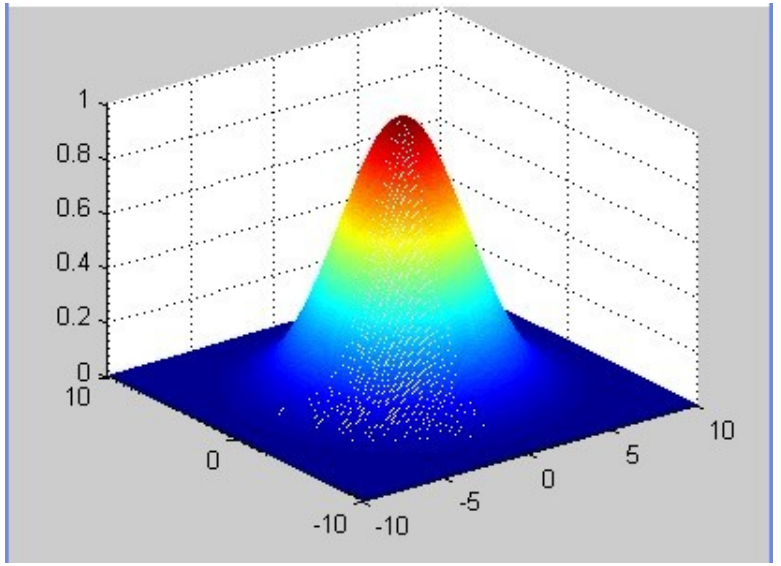
$$\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)dx=1$$

示例2：多维高斯分布

- 概率密度函数：

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$





贝叶斯公式

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

后验概率：给定观测数据后，某事件发生的概率

先验概率：在没有观测数据之前，某事件发生的初始概率

似然概率：给定事件发生的情况下，观测数据出现的概率

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

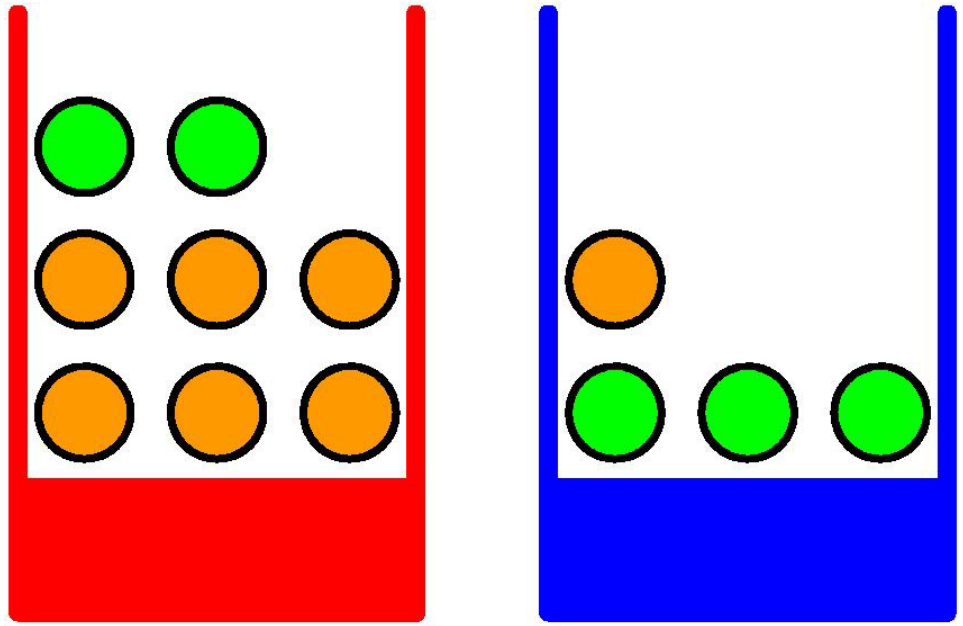
证据因子：观测数据的边际概率，即所有可能事件下观测数据的总概率

贝叶斯公式

- 贝叶斯公式 (Bayes' Theorem)
  - 选箱子事件变量B
  - 选水果事件变量F

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

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



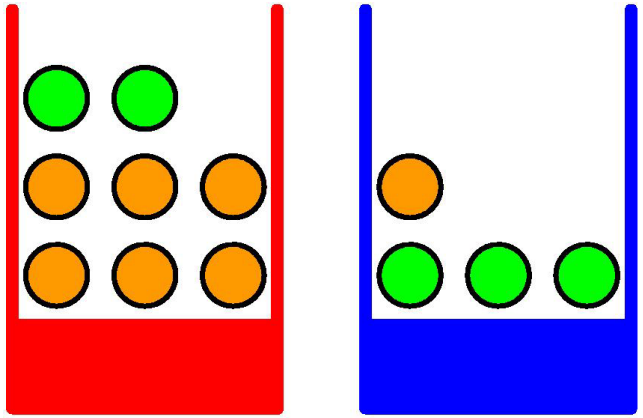
  - 1.取到苹果的概率?
  - 2.如果取到桔子，来自红箱子的概率?

贝叶斯公式

- 贝叶斯公式 (Bayes' Theorem)

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

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



- 1.取到苹果的概率?

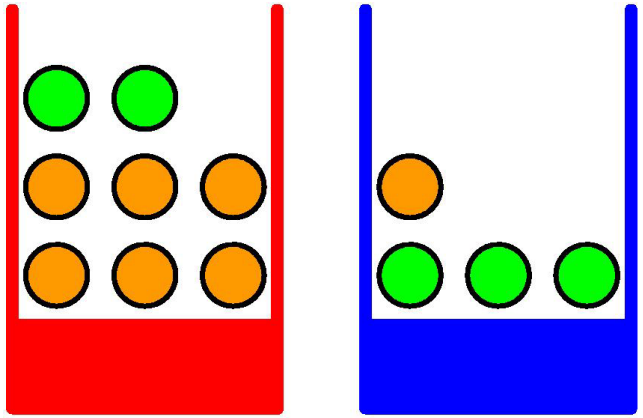
$$P(F = a|B = r) = 1/4$$
$$P(F = o|B = r) = 3/4$$
$$P(F = a|B = b) = 3/4$$
$$P(F = o|B = b) = 1/4$$

⇩ 全概率公式

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

贝叶斯公式

- 贝叶斯公式 (Bayes' Theorem)



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

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

➢ 2.如果取到桔子，来自红箱子的概率?

$$P(B = r|F = o) = \frac{P(F = o|B = r)P(B = r)}{P(F = o)}$$

贝叶斯公式

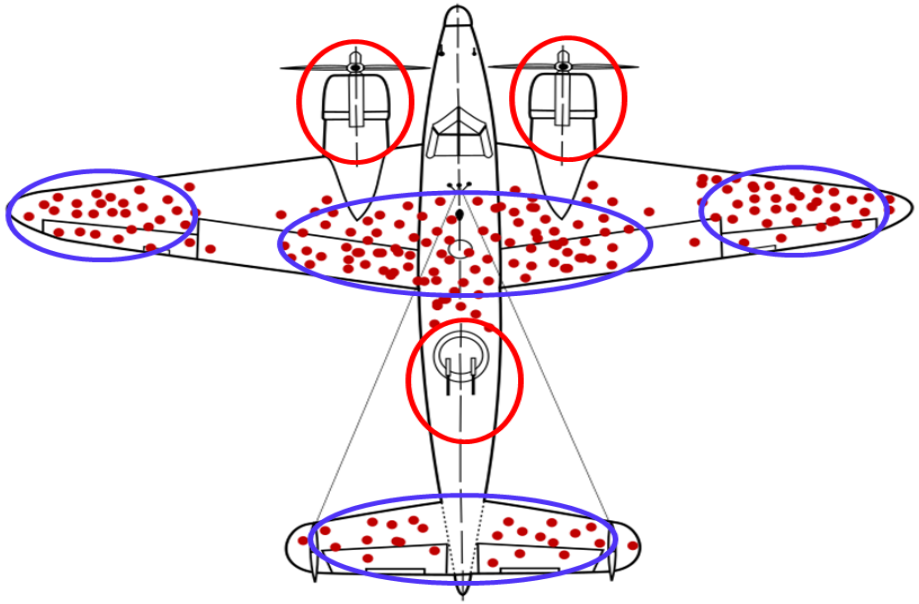
$$= 3/4 \times 4/10 \times 20/9$$

$$= 2/3$$

现实中的贝叶斯

- 贝叶斯公式与幸存者偏差
  - 二战期间，盟军邀请统计学家瓦尔德运用他在统计方面的专业知识给出关于“飞机应该如何加强防护，才能降低被炮火击落的几率”的建议
  - 应该加固弹孔多的机翼还是弹孔少的发动机?

统计学家瓦尔德



■ 军方认为：“应该加强机翼的防护，因为这是最容易击中的位置”

■ 瓦尔德：“我们应该强化发动机的防护”

现实中的贝叶斯

- **贝叶斯公式与幸存者偏差**
  - **幸存者偏差是一种逻辑谬误，本质是样本偏差**

“大学生随处可见”

“我的年薪只有20万，太低了”

  - **真实情况是，中国的本科率只有约4%①，生活在大学里所以大学生随处可见**
  - **2021年，应届本科生月薪超1万元的只有4.3%②，只有互联网、金融等少数几个专业薪资水平较高。**

① 2020年数据，根据国家统计局数据计算：到2020年，本科毕业人数约5700万，总人口14.亿

②《2021年中国本科生就业报告》-麦克思

现实中的贝叶斯

- **幸存者偏差的贝叶斯视角**
  - 贝叶斯公式

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

  - 设A为飞机被击中的部位，B为飞机能否返航，1：返航，0：击落。则$P(A|B=1)$为返航飞机被击中部位的条件分布，有

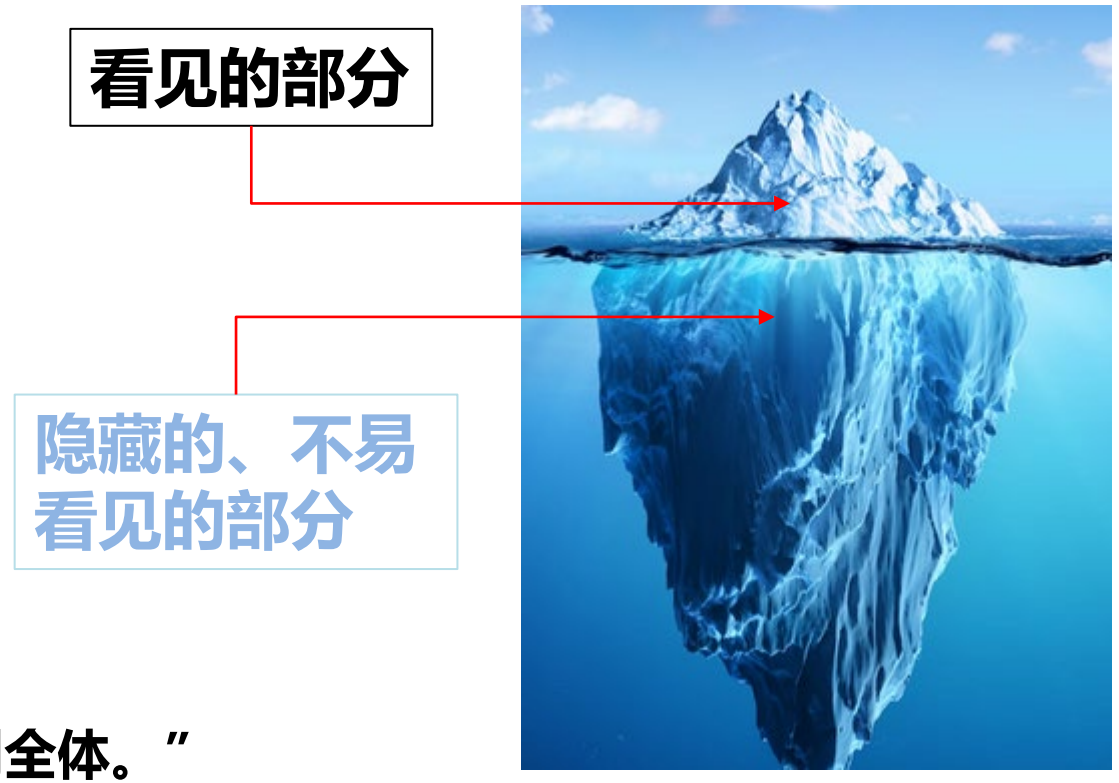
$$P(A|B=1) = \frac{P(B=1|A)P(A)}{P(B=1)}$$

$$P(A|B=1) \propto P(B=1|A)$$

  - 上式表明：被击中越多的部位返航概率越大，击中部位少的飞机返航概率小，所以要加强引擎

现实中的贝叶斯

- **克服认知偏差**
  - 全面、系统的看待问题
  - 不能忽略隐藏的、不易看见的部分
  - 用系统性思维，全面、综合的看待问题



“马克思主义者看问题，不但要看到部分，而且要看到全体。”

—毛泽东

2.4 贝叶斯决策理论

- 贝叶斯决策的概念
- 最小错误率贝叶斯决策
- 最小风险贝叶斯决策
- 朴素贝叶斯决策

贝叶斯决策

- 贝叶斯决策是统计决策理论中的一个基本方法，用于解决分类问题
  - 已知条件：
    - 1、属于一定数量类别的数据，类别为标签为： $\omega_i, i = 1, 2, \ldots, c$
    - 2、各类别 ω_i的类先验概率$P(\omega_i)$和类条件概率密度$P(x|\omega_i)$
  - 基本思想：根据贝叶斯公式计算后验概率，基于最大后验概率进行判决
  - 判决函数：最大化后验概率

$$x \in \omega_k \text{ 当且仅当} k = \arg\max_i \{P(\omega_i|x)\}, \text{ 其中} P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} P(x|\omega_j)P(\omega_j)}$$

贝叶斯决策

- 贝叶斯决策方法
  - 最小错误率贝叶斯决策：最小化分类错误的概率，基于最大后验概率原则，在分类中选择后验概率最大的类别。
  - 最小风险贝叶斯决策：在考虑错误率的基础上，还考虑不同错误的风险，通过最小化期望风险进行决策。
  - 朴素贝叶斯决策：在类先验概率和类条件概率密度未知，需要估计时，通过假设所有属性相互独立，实现降低样本集大小需求、降低复杂度的目的。

最小错误率贝叶斯决策

- **最小错误率贝叶斯决策目标**
  - **目标：最小化分类错误概率$P(e)$，即：** $\min P(e) = \int P(e|x)p(x)dx$
  - **因为$P(e|x) \geq 0, p(x) \geq 0$，所以决策目标也可表示为：** $\min_x P(e|x)$

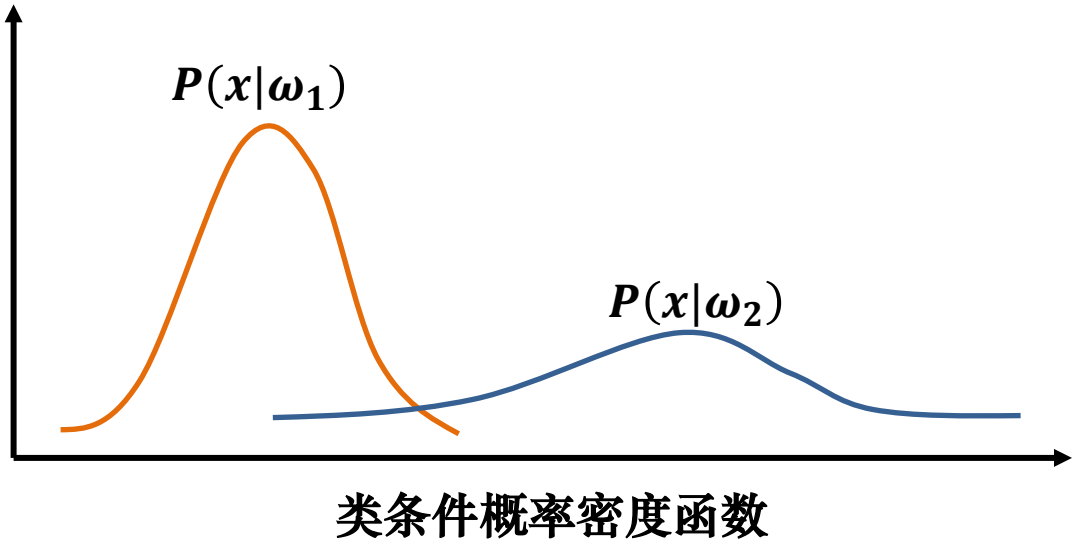
    最小错误率贝叶斯决策方法：

  - **对于样本x和类别ω_1, ω_2有：** $P(e|x) = \begin{cases} P(\omega_2|x), \text{若 } P(\omega_1|x) > P(\omega_2|x) \\ P(\omega_1|x), \text{若 } P(\omega_2|x) > P(\omega_1|x) \end{cases}$
  - **将样本x判别为后验概率更大的类别，对应的错误概率$P(e|x)$更小**
  - **对于每个样本x，最小错误率贝叶斯决策的结果为ω，它是后验概率最大的决策结果，即：**

$$P(\omega|x) = \max_{j=1,\ldots,c} P(\omega_j|x)$$

最小错误率贝叶斯决策

- 示例：航空发动机零件诊断
  - 假设在航空发动机中某个零件正常(ω_1)和异常(ω_2)两类的先验概率分别为：
    - 正常状态：$P(\omega_1) = 0.9$
    - 异常状态：$P(\omega_2) = 0.1$
  - 现有一待识别零件，其观察值为x，从类条件概率密度函数曲线上查得：
    - $P(x|\omega_1) = 0.2$
    - $P(x|\omega_2) = 0.4$
  - 试对该零件 x 进行分类



最小错误率贝叶斯决策

- **示例：航空发动机零件诊断**
  - **解：利用贝叶斯公式计算ω_1和ω_2的后验概率**

$$P(\omega_1|x)=\frac{P(x|\omega_1)P(\omega_1)}{\sum_{j=1}^{2}P(x|\omega_j)P(\omega_j)}=\frac{0.2\times0.9}{0.2\times0.9+0.4\times0.1}=0.818$$

$$P(\omega_2|x)=1-P(\omega_1|x)=0.182$$

根据贝叶斯决策规则，有

$$P(\omega_1|x)=0.818>P(\omega_2|x)=0.182$$

因此，该零件为正常零件

最小错误率贝叶斯决策

- 最小错误率贝叶斯决策的风险
  - 不同的决策具有不同的风险或损失
  - 以航空发动机零件诊断为例：
    - 故障零件判断为正常：极大增加安全隐患，后果严重
    - 正常零件判断为故障：进一步进行检修，损失较小

除零件诊断外，在医疗疾病诊断、金融投资风险判断等领域，不同决策同样会带来不同程度的风险

  - 最小错误率贝叶斯决策以错误率最小为准则，未考虑决策的风险

最小风险贝叶斯决策

- **损失函数**：
  - 对于特定的x，分类为ω_j，对应决策α_i的损失记为：$\lambda(\alpha_i, \omega_j)$
- **条件期望损失**：
  - $R(\alpha_i|x) = E[\lambda(\alpha_i, \omega_j)] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|x), i = 1, 2, \ldots, a$
- **期望风险**：
  - 对所有可能的x，采取决策α_x所造成的条件期望损失之和 $R(\alpha) = \int R(\alpha(x)|x)p(x)dx$
- 基于已知样本计算的期望风险称为经验风险，在实际决策中，常用经验风险来估计期望风险。经验风险只能反映在当前已知数据下的表现，当未知数据和已知数据偏差大时，存在过拟合风险。

最小风险贝叶斯决策

- **最小风险贝叶斯决策目标**
  - 目标：最小化期望风险，即：$\min R(\alpha) = \int R(\alpha(x)|x)p(x)dx$
  - 若对每个决策都使其条件期望损失最小，则对所有x做出决策时，期望风险也最小
- **最小风险贝叶斯决策方法：**
  - 对于样本x，最小风险贝叶斯决策的判决结果是ω，设结果ω对应的决策为α_k，它们满足：

$$R(\alpha_k|x) = \min_{i=1,2,\dots,a} R(\alpha_i|x)$$

  - 若损失函数R为0-1损失，即：$\lambda(\alpha_i, \omega_j) = \begin{cases} 0, i = j \\ 1, i \neq j \end{cases}, i, j = 1, 2, \dots, c$，则最小风险贝叶斯决策等价于最小错误率贝叶斯决策，即：

$$\omega_k = \arg\min_i R(\alpha_i|x) = \arg\min_i \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|x) = \arg\min_i \sum_{\substack{j=1 \\ j\neq i}}^{c} P(\omega_j|x)$$

最小风险贝叶斯决策

- **示例：航空发动机零件诊断**
  - 假设在航空发动机中某个零件正常(ω_1)和异常(ω_2)两类的先验概率分别为
    - 正常状态：$P(\omega_1)=0.9$；异常状态：$P(\omega_1)=0.1$
  - 现有一待识别零件，其观察值为x，从类条件概率密度曲线上查得
    - $P(x|\omega_1)=0.2$；$P(x|\omega_2)=0.4$
  - 决策与损失表：

损失 / 分类 / 决策	ω_1	ω_2
α_1	0	6
α_2	1	0

  - 试对该零件 x 进行分类

最小风险贝叶斯决策

- **示例：航空发动机零件诊断**
  - **解：**
    - **损失函数：** $\lambda(\alpha_1,\omega_1)=0$，$\lambda(\alpha_1,\omega_2)=6$，$\lambda(\alpha_2,\omega_1)=1$，$\lambda(\alpha_2,\omega_2)=0$
    - **后验概率：** $P(\omega_1|x)=0.818$，$P(\omega_2|x)=0.182$
    - **计算条件风险：** $R(\alpha_1|x)=\sum_{j=1}^{2}\lambda(\alpha_1,\omega_j)P(\omega_j|x)=\lambda(\alpha_1,\omega_2)P(\omega_2|x)=1.092$，

      $R(\alpha_2|x)=\lambda(\alpha_2,\omega_1)P(\omega_1|x)=0.818$
    - **决策：** 由于$R(\alpha_1|x)>R(\alpha_2|x)$，因此样本$x$的决策结果为异常零件

两种贝叶斯决策步骤对比

- **决策步骤：**
  - 1、已知先验概率$P(\omega_i)$，类条件概率$P(x|\omega_i), i=1,2,\dots,c$，待分类输入数据$x$
  - 2、根据贝叶斯公式计算后验概率：$P(\omega_i|x)=\frac{P(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{c}P(x|\omega_j)P(\omega_j)}$

  - 3、得到所有类别后验概率 $P(\omega_j|x)$
  - 4、决策：$P(\omega|x)=\max_{j=1,\dots,c} P(\omega_j|x)$

  - 3、利用后验概率与损失函数，计算条件风险 $R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j|x), i=1,2,\dots,a$
  - 4、决策：$R(\alpha_k|x)=\min_{i=1,2,\dots,a} R(\alpha_i|x)$

最小错误率贝叶斯决策　　最小风险贝叶斯决策

朴素贝叶斯决策

- 最小错误率/最小风险贝叶斯决策的问题：实际情况无法预先获取类先验概率和类条件概率，需要通过现有样本预估
  - 实际能获取的样本往往较少，且具有多个维度的属性
  - 类条件概率 $P(x|\omega_i)$是所有属性上的联合概率，难以从有限的训练样本直接估计
- 解决方法：朴素贝叶斯决策
  - 引入属性条件独立性假设：对于已知类别，假设所有属性相互独立；即假设各属性独立地对分类结果发生影响，即

$$P(x|\omega) = P(x_1 x_2, \dots, x_i, \dots, x_d|\omega) = \prod_{i=1}^{d} P(x_i|\omega)$$

  - 引入属性独立性假设后，不需要大量数据来估计各属性的联合概率分布，从而达到降低样本集大小需求的目的

朴素贝叶斯决策

- **朴素贝叶斯决策思想**
  - 基于最大化后验概率进行判决，但在使用贝叶斯公式求后验概率时，类条件概率转化为各属性条件概率的乘积，即

$$P(\omega|x) = \frac{P(\omega)P(x|\omega)}{P(x)} = \frac{P(\omega)}{P(x)}\prod_{i=1}^{d} P(x_i|\omega)\text{，}\ x_i\text{为样本的各属性}$$

- **朴素贝叶斯决策方法**
  - 训练：基于训练样本数据来估计类先验概率 $P(\omega_j)$，并为每个属性x_i估计条件概率$P(x_i|\omega_j)$，也就是它们在训练数据上的频率。
  - 决策：$\omega_k = \arg\max_j P(\omega_j)\prod_{i=1}^{d} P(x_i|\omega_j)$

朴素贝叶斯决策

- **示例：无人机飞行条件判断**
  - **训练样本**

天气	温度	湿度	风速	是否适宜放飞
晴天	高	高	弱	否
晴天	高	高	强	否
阴天	高	高	弱	是
雨天	低	高	弱	是
雨天	低	低	弱	是
雨天	低	低	强	否
阴天	低	低	强	是
晴天	中	高	弱	否
晴天	低	低	弱	是
雨天	中	低	弱	是

  - **测试数据：“晴天，高温，高湿度，弱风速”**

朴素贝叶斯决策

- **示例：无人机飞行条件判断**
  - **1、计算类先验概率：$P(\text{适宜}=\text{是})=\frac{6}{10}=0.6$，$P(\text{适宜}=\text{否})=0.4$**
  - **2、对每个属性计算类条件概率：**
    - $P(\text{天气}=\text{晴}|\text{适宜}=\text{是})=\frac{2}{6}=0.33$，$P(\text{温度}=\text{高}|\text{适宜}=\text{是})=0.5$，$P(\text{湿度}=\text{高}|\text{适宜}=\text{是})=0.5$，$P(\text{风速}=\text{弱}|\text{适宜}=\text{是})=0.67$
    - $P(\text{天气}=\text{晴}|\text{适宜}=\text{否})=0.75$，$P(\text{温度}=\text{高}|\text{适宜}=\text{否})=0.5$，$P(\text{湿度}=\text{高}|\text{适宜}=\text{否})=0.75$，$P(\text{风速}=\text{弱}|\text{适宜}=\text{否})=0.5$
  - **3、计算后验概率：**
    - $P(\text{适宜=是}|\text{晴天, 高温, 高湿度, 弱风速})=0.33\times0.5\times0.5\times0.67\times0.6=0.033$
    - $P(\text{适宜=否}|\text{晴天, 高温, 高湿度, 弱风速})=0.75\times0.5\times0.75\times0.5\times0.4=0.056$，概率更大
  - **4、决策：晴天，高温，高湿度，弱风速条件不适宜放飞无人机**

贝叶斯决策理论小结

- 基本思想：已知类条件概率$P(x|\omega_i)$和类先验概率$P(\omega_i)$，计算后验概率$P(\omega_i|x)$进行决策
- 最小错误率贝叶斯决策：最小化分类错误的概率，等价于最大化后验概率
- 最小风险贝叶斯决策：在最小错误率贝叶斯决策的基础上，考虑不同错误的损失，通过最小化期望风险进行决策
- 朴素贝叶斯决策：在类条件概率和类先验概率未知，需要较多样本进行估计时，朴素贝叶斯决策假设所有属性相互独立，在有限样本时更好估计类条件概率$P(x|\omega_i)$和类先验概率$P(\omega_i)$

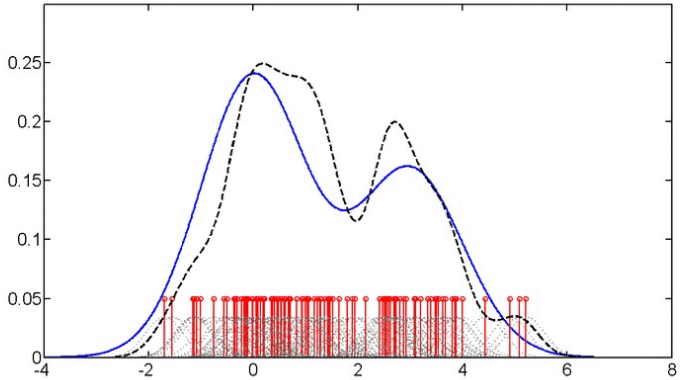
■ 当样本充足时，可以通过贝叶斯估计类条件概率密度函数和类先验概率密度函数

2.5 参数化概率密度估计方法

- 概率密度估计的概念
- 参数化概率密度估计方法的概念
- 极大似然估计方法
- 贝叶斯估计方法
- 估计量的性质与评价标准

概率密度估计

- 概率密度估计的任务：
  - 根据样本数据估计类条件概率密度函数 $P(x|\omega_i)$和类先验概率$P(\omega_i)$
- 为什么需要估计概率密度
  - 概率密度估计可以建模原始数据分布，帮助我们更精细地了解数据特性。进而帮助我们识别数据中的异常值、甚至用于生成新数据



- 概率密度估计的方法：
  - 参数化方法：已知概率密度函数的形式，其中几个参数未知
  - 非参数化方法：概率密度函数的形式未知

https://en.wikipedia.org/wiki/Density_estimation

参数化概率密度估计

- 参数化概率密度估计的任务
  - 已知：已知概率密度函数的形式，只是其中一个或几个参数未知
  - 目标：依据样本估计这些未知参数的值
- 典型方法
  - 极大似然估计：把待估计参数看做是确定的量，只是其取值未知。最佳估计就是使产生已观测到样本的概率最大的那个值
  - 贝叶斯估计：把待估计参数看做是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，从而利用样本信息修正参数的初始估计值的过程

极大似然估计

- 极大似然估计的假设条件
  - $P(x|\omega_i)$具有某种确定的解析函数形式，只有部分参数θ未知；
  - 参数θ通常为向量，如一维正态分布$N(\mu, \sigma^2)$中的μ、σ
  - 参数θ是确定的未知量，不是随机量
  - 各类样本集x_i，$i=1,2,\ldots,c$ 满足独立同分布条件$(i.i.d.)$，即x_i均为从密度为$P(x|\omega_i)$的总体中独立抽取出来的
  - 各类样本只包含本类分布的信息；因此，$P(x|\omega_i)$可记为$P(x|\omega_i;\theta_i)$或$P(x;\theta_i)$
- 基于上述假设，各类条件概率密度可根据各类样本分别估计

极大似然估计

- **似然函数**
  - 针对一类已知样本$X=\{x_i, i=1,2,\dots,N\}$，定义参数θ下观测到样本集X的联合分布概率密度，称为相对于样本集X的θ的似然函数

$$l(\theta)=P(x_1,x_2,\dots,x_n;\theta)=\prod_{i=1}^{n}P(x_i;\theta)$$

- **基本思想**
  - 在θ可能的取值范围内选择使似然函数达到最大的参数值作为参数θ的估计值
  - 形式化描述为：求$\hat{\theta}$，使得$l(\hat{\theta})=\max_{\theta} l(\theta)$
  - 如果参数$\theta=\hat{\theta}$时，$l(\theta)$最大，则$\hat{\theta}$是最可能的参数估计值。它是样本集的函数，记作：$\hat{\theta}=d(x_1,x_2,\dots,x_N)=d(X)$，称为极大似然估计量
  - 为便于分析求解，实际运用中往往采用对数似然函数：$H(\theta)=\ln l(\theta)$

极大似然估计

- **求解**
  - 若似然函数连续可微，最大似然函数估计量就是方程$\frac{dl(\theta)}{d\theta}=0$或$\frac{dH(\theta)}{d\theta}=0$的解
  - 若未知参数不止一个，即$\theta=[\theta_1,\theta_2,\dots,\theta_s]^T$，则需联立以下$s$个方程组求解

$$\frac{dH(\theta)}{d\theta_i}=0, i=1,2,\dots,s$$

- **极大似然函数的求解性质**
  - 若似然函数连续可导，存在最大值且必要条件方程有唯一解，则解就是极大似然估计量
  - 如果似然函数有多个解，则使似然函数值最大者为极大似然估计量
  - 若似然函数单调，可根据极大似然思想，将似然函数最大值点作为参数的极大似然估计值

极大似然估计

- **极大似然估计的求解步骤**
  - 1、确定需要估计的概率分布$P(x|\theta)$，其中$\theta = [\theta_1, \theta_2, \dots, \theta_s]^T$
  - 2、构造似然函数$l(\theta) = P(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} P(x_i; \theta)$
  - 3、求对数似然函数$H(\theta) = \ln l(\theta)$
  - 4、令$\frac{dH(\theta)}{d\theta_i} = 0, i = 1, 2, \dots, s$，联立求解

极大似然估计

- **示例：单变量正态分布**
  - 已知
    - 参数：$\boldsymbol{\theta = [\theta_1, \theta_2]}$，$\boldsymbol{\theta_1 = \mu, \theta_2 = \sigma^2}$
    - 概率密度函数：$\boldsymbol{P(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]}$
    - 样本集：$\boldsymbol{X = \{x_1, x_2, \ldots, x_N\}}$
  - 目标：估计参数$\boldsymbol{[\theta_1, \theta_2]}$

极大似然估计示例：单变量正态分布

- **解：**
  - 1、求似然函数：$l(\theta) = P(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} P(x_i; \theta)$
  - 2、求对数似然函数：$H(\theta) = \ln l(\theta) = \sum_{i=1}^{N} \ln P(x_i; \theta)$
  - 3、构造方程组：$\begin{cases} \frac{\partial H}{\partial \mu} = 0 \\ \frac{\partial H}{\partial \sigma^2} = 0 \end{cases}$，$\begin{cases} \frac{1}{\sigma^2}\left[\sum_{i=1}^{N} x_i - N\mu\right] = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i - \mu)^2 = 0 \end{cases}$
  - 4、联立求解：$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$，$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$

贝叶斯估计

- **贝叶斯估计的基本思想**
  - 把待估计参数 θ 看作是具有先验分布$P(\theta)$的随机变量，其取值与样本集 X 有关，贝叶斯估计利用样本集 X 将 θ 的先验概率分布修正为后验概率分布
  - 贝叶斯决策用于分类，计算离散形式的后验概率值，而贝叶斯估计则用于回归，计算连续形式的后验概率密度函数
- **贝叶斯估计损失函数**
  - 把 θ 估计为 $\widehat{\theta}$ 所造成的损失，记为$\lambda(\widehat{\theta},\theta)$
  - 不同于离散形式贝叶斯决策的损失表，由于参数化概率密度估计为连续值估计，因此常采用损失函数，常用平方误差损失函数$\lambda\left(\widehat{\theta},\theta\right)=(\theta-\widehat{\theta})^2$

贝叶斯估计

- 贝叶斯估计相关概念
  - **条件期望损失：** $R(\hat{\theta}|x)=\int_{\Theta}\lambda(\hat{\theta},\theta)P(\theta|x)d\theta$，其中，$x\in E^d$，$\theta\in\Theta$，$E^d$为样本集，$\Theta$为待估参数集
  - **期望风险：** $R=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)P(x;\theta)d\theta dx$

    $$=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)P(\theta|x)P(x)d\theta dx=\int_{E^d}R(\hat{\theta}|x)P(x)dx$$
  - **贝叶斯估计量：** 使条件期望损失最小的估计量$\hat{\theta}$，即

    $$\hat{\theta}=\operatorname{argmin}\left(R(\hat{\theta}|x)\right)=\operatorname{argmin}(\int_{\Theta}\lambda(\hat{\theta},\theta)P(\theta|x)d\theta)$$
  - **定理2.5.1：** 若采用平方误差损失函数，则θ的贝叶斯估计量是在给定样本集X时θ的条件期望，即$\hat{\theta}=E(\theta|x)=\int_{\Theta}\theta P(\theta|x)d\theta$

定理2.5.1证明过程，可参考《模式识别》-边肇祺张学工P52

贝叶斯估计

- 贝叶斯估计步骤
  - 1、确定参数 θ 所遵从的先验分布： $P(\theta)$
  - 2、求样本集的联合分布： $P(X|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$
  - 3、求 θ 的后验概率分布： $P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{\Theta} P(X|\theta)P(\theta)d\theta}$
  - 4、求 θ 的贝叶斯估计量（定理2.5.1）： $\hat{\theta} = \int_{\Theta} \theta P(\theta|x)d\theta$

贝叶斯估计

- **示例：单变量正态分布**
  - **已知**
    - **参数：$\boldsymbol{\theta = [\theta_1, \theta_2]}$，$\boldsymbol{\theta_1 = \mu, \theta_2 = \sigma^2}$，其中$\boldsymbol{\theta_2}$已知**
    - **概率密度函数：$\boldsymbol{P(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]}$**
    - **样本集：$\boldsymbol{X = \{x_1, x_2, \ldots, x_N\}}$**
  - **目标：估计均值$\boldsymbol{\theta_1}$**

贝叶斯估计示例：单变量正态分布

- **解：（仅保留关键步骤）**
  - 1、确定概率密度函数形式：$P(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$
  - 2、设估计量μ遵从以下分布：$\mu\sim\mathcal{N}(\mu_0,\sigma_0^2)$
  - 3、根据观测样本x求得μ的后验分布：$\mu|x\sim\mathcal{N}(\frac{\mu_0\tau_0+n\bar{x}\tau}{\tau_0+n\tau},\frac{1}{\tau_0+n\tau})$，其中$\tau=\frac{1}{\sigma^2}$，$\tau_0=\frac{1}{\sigma_0^2}$
  - 4、μ的贝叶斯估计量为

$$E(\mu|x)=\frac{\mu_0\tau_0+n\bar{x}\tau}{\tau_0+n\tau}=w\mu_0+(1-w)\bar{x}$$，其中n为样本数，$w=\frac{\tau_0}{\tau_0+n\tau}$

  - 5、当$n=0$时，估计量$\hat{\mu}=\mu_0$，当$n\to\infty$时，估计量$\hat{\mu}=\bar{x}$

具体可参考《模式识别》-边肇祺张学工P56-57

估计量的性质与评价标准

估计量的性质

无偏性	渐进无偏性	有效性	一致性
$E\left[\hat{\theta}(x_1, x_2, \ldots, x_N)\right] = \theta$	$E\left[\hat{\theta}(x_N)\right] \overset{N\to\infty}{\Longrightarrow} \theta$	对估计$\widehat{\boldsymbol{\theta}}_1$，$\widehat{\boldsymbol{\theta}}_2$，若方差$\boldsymbol{\sigma^2(\widehat{\theta}_1) < \sigma^2(\widehat{\theta}_2)}$，则$\widehat{\boldsymbol{\theta}}_1$估计更有效	当样本数无穷多时，每一次估计都在概率意义上任意接近真实值，即：$\forall \varepsilon > 0,\ \lim_{N\to\infty} P\left(\left\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\right\| > \varepsilon\right) = 0$

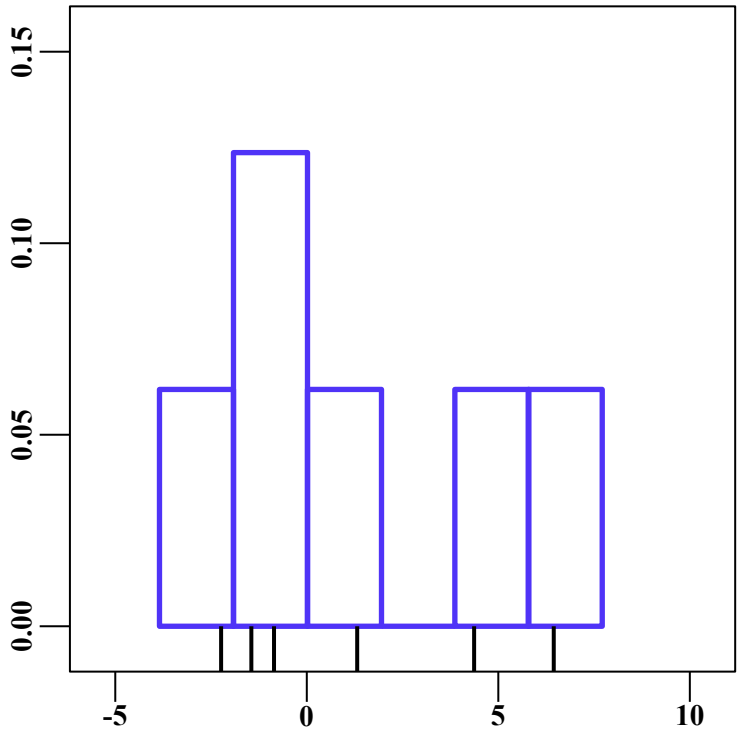
- 结合无偏性和有效性，要求估计量能够在多次估计中，以较小的方差平均地表示真实值
- 极大似然估计是无偏的，在样本充足时有效性更好，一致性更强；
- 贝叶斯估计可能有偏，在样本量有限且有合理的先验信息时更有效。

2.6 非参数概率密度估计方法

- 什么是非参数估计?
- Parzen窗算法
- k近邻算法

什么是非参数估计?

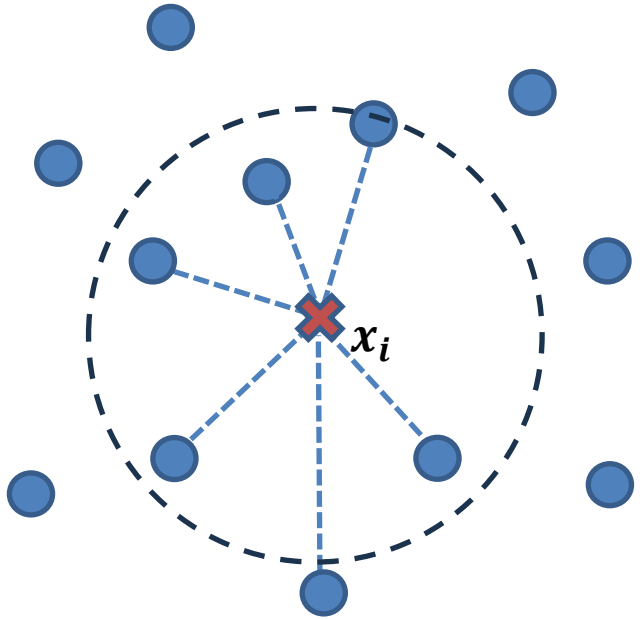
- 非参数估计
  - 概率密度函数的形式非已知的模型，直接依赖于数据本身来进行推断和估计



- 常见的非参数估计方法
  - Parzen窗法
  - k近邻法

非参数估计方法

- 非参数估计方法
  - 基本思路：要估计x_i点的密度，可把相关样本在该点的“贡献”相加近似作为其概率密度，进而可以以此方法估计每个点的概率密度
  - 具体流程
    - 1.确定计算x_i点处概率密度的相关贡献点
    - 2.确定贡献点对x_i点处的贡献
    - 3.重复1-2计算所有点处的概率密度



非参数估计方法

- **非参数概率密度估计**
  - 假设N为样本总数，以x_i为中心的区域R（足够小，体积为V）内的k个点对估计x_i的概率密度$p(x)$有贡献，则R中落入k个样本的概率为：

$$P_R = k/N = \int_R p(x)dx = \hat{p}(x)V$$

  - 估计得到的概率密度$\hat{p}(x_i)$为：　$\hat{p}(x_i) = k/NV$
  - 当满足以下条件时，我们的估计概率$\hat{p}(x_i)$收敛于$p(x_i)$：
    - 贡献点的区域大小越小越好　$\lim_{N\to\infty} V_N = 0$
    - 贡献点越多越好　$\lim_{N\to\infty} k_N = \infty$
    - 贡献点与总样本数比例越小越好　$\lim_{N\to\infty} \frac{k_N}{N} = 0$

非参数估计方法

- **Parzen法**
  - 使区域体积序列V_N以N的某个函数的关系不断缩小
  - 同时限制k_N和k_N/N

有限的N, V选择很敏感

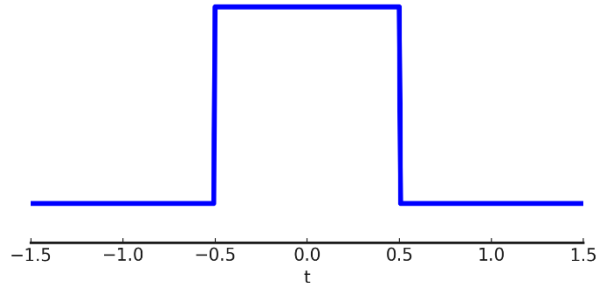
↓

动态变化V的取值

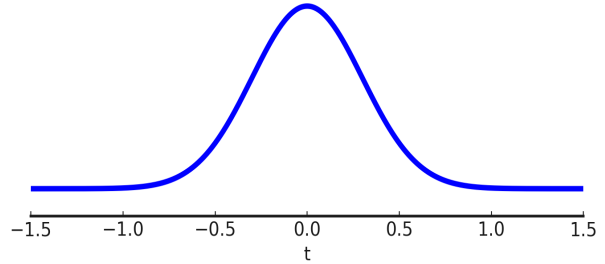
- **k近邻法**
  - 使落入区域样本数k_N为N的某个函数
  - 选择不同的V_N使区域包含x的k_N个近邻

非参数估计方法——Parzen窗法

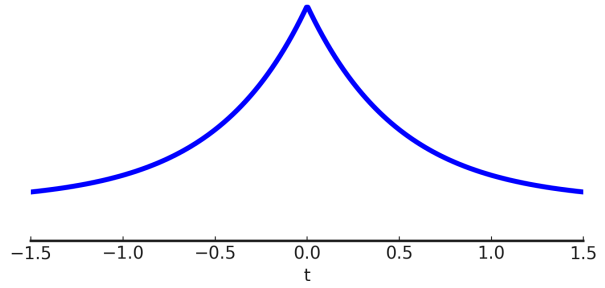
- Parzen窗法使用窗函数对贡献点进行选择
- 窗函数
  - 形式：$k(x, x_i)$，反映x_i对$p(x)$的贡献同时进行区域选择
  - 条件：$k(x, x_i) \geq 0, \int k(x, x_i)dx = 1$
  - 窗函数选择
    - 方窗函数、正态窗函数、指数窗函数等
  - 窗宽选择
    - 原则：样本数多则选小些；样本数少则选大些



方窗函数



正态窗函数

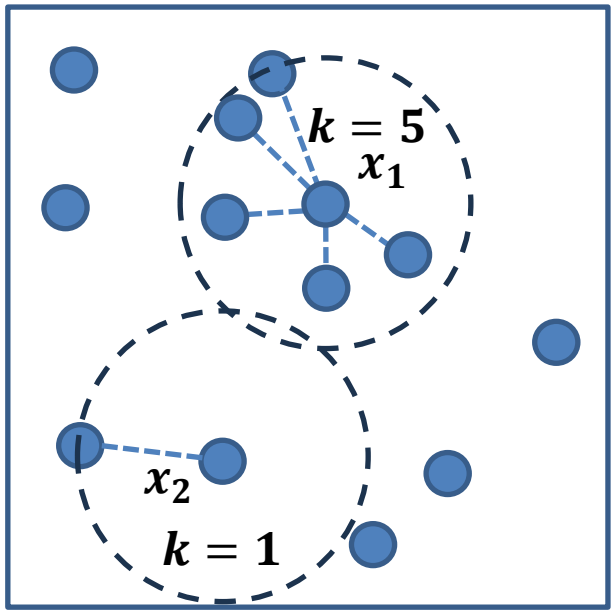


指数窗函数

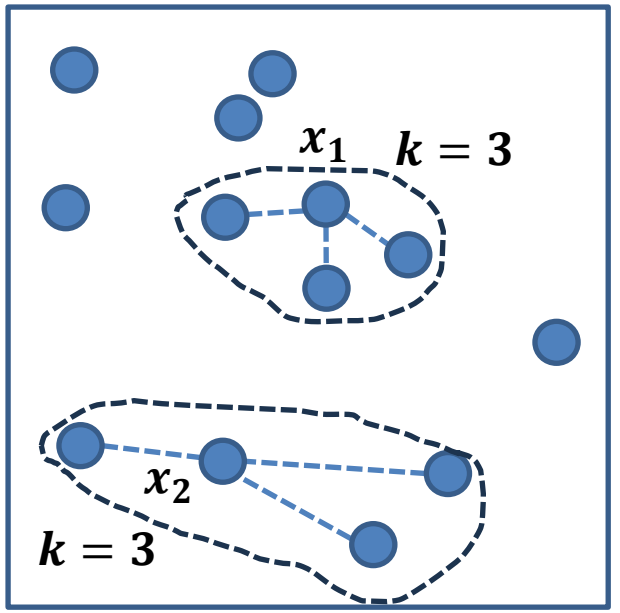
非参数估计方法——k近邻法

- k近邻法使用k近邻算法对贡献点进行选择
  - 选择样本x_i一定范围内确定个数的k个样本后根据k/NV计算概率密度
  - k近邻法更适用于样本分布不均匀的数据

Parzen窗法



k近邻法



k近邻算法

- **k近邻算法流程：**
  - 选择一个正整数k，表示需要考虑的邻近样本的数量
  - 对于待预测的样本，计算其与训练集中所有样本之间的距离
    - 常用的距离度量包括欧氏距离、曼哈顿距离、余弦相似度等
  - 选择k个最近邻居：根据计算得到的距离，选择距离待预测样本最近的k个邻居
- **k值选择：**
  - k值决定了决策的局部性，k值越大，模型越平滑，越小则越敏感

小结

- Parzen窗法与k近邻法均使用广泛
- 在样本分布不均匀时k近邻法比Parzen窗法表现更好
- 在高维空间中k近邻法比Parzen窗法更易应用，且可通过技术手段缓解维度问题
- k近邻法和Parzen窗法在边界附近都可能会遇到估计偏差

2.7* 矩阵理论基础（拓展）

- 矩阵的定义
- 矩阵的运算

矩阵的定义

- **定义：** 由$m\times n$个数a_{ij} $(i=1,2,\cdots, m\ ; j=1,2,\cdots, n)$ 排成的m行n列的数表

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

称为m行n列矩阵，简称$m\times n$矩阵

- 简记为: $A=A_{m\times n}=(a_{ij})_{m\times n}$
- 这$m\times n$个数称为矩阵A的元素，a_{ij}称为矩阵A的第i行第j列元素

矩阵的定义

● 几种特殊矩阵

(1) 行数与列数都等于n的矩阵A, 称为n阶方阵. 也可记作A_n,

例如: $\begin{pmatrix} 13 & 6 & 5 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$ 是一个3 阶方阵.

(2) 只有一行的矩阵 $A=(a_1,a_2,\cdots,a_n)$,称为行矩阵(或行向量)

(3) 只有一列的矩阵 $B=\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$, 称为列矩阵(或列向量).

矩阵的定义

- **几种特殊矩阵**

(4) 元素全为零的矩阵称为零矩阵, 记作O。注意：不同阶数的零矩阵是不相等的

例如 $\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \neq (0 \quad 0 \quad 0 \quad 0).$

(5) 形如 $\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ 的方阵, 称为单位矩阵，

其中主对角线上元素都是1，其他元素都是0。记作:E_n 或 E

矩阵的定义

- **几种特殊矩阵**

(6) 形如 $\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$ 的方阵, 称为对角矩阵(或对角阵),

其中$\lambda_1, \lambda_2, \cdots, \lambda_n$不全为零.记作$A=\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$

(7) 设$A=(a_{ij})$为 n 阶方阵, 对任意 i, j, 如果$a_{ij}=a_{ji}$都成立, 则称A为对称矩阵.

例如: $A=\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 4 \\ 3 & 4 & 6 \end{pmatrix}$ 为对称矩阵.

注: 行列式与矩阵的区别: 一个是算式（数值），一个是数表；一个行列数相同，一个行列数可不同；对 n 阶方阵可求它的行列式. 记为: $|A|$

矩阵的运算-加法

- **定义:** 设有两个$m\times n$ 矩阵$A=(a_{ij})$与 $B=(b_{ij})$，那么矩阵A与B的**和**记作$A+B$，规定为

$$A+B=\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix}$$

注意: 仅当两个矩阵是同型矩阵时，才能进行加法运算

- **性质:** (1) $A+B=B+A$　　(2) $(A+B)+C=A+(B+C)$

(3) $A+(-A)=0$　　$A-B=A+(-B)$

矩阵的运算-数乘

- **定义:**数λ与矩阵A的乘积记作λA或$A\lambda$，规定为

$$\lambda A = A\lambda = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda a_{m1} & \lambda a_{m1} & \cdots & \lambda a_{mn} \end{pmatrix}$$

- **性质：**（设A、B都是$m\times n$ 矩阵, λ, μ 为数）：

$$(1)\ (\lambda\mu)A = \lambda(\mu A)$$

$$(2)\ (\lambda+\mu)A = \lambda A + \mu A$$

$$(3)\ \lambda(A+B) = \lambda A + \lambda B$$

矩阵相加与矩阵数乘合起来统称为**矩阵的线性运算**

矩阵的运算-乘法

- **定义:** 设$A=(a_{ij})$是一个 $m\times s$ 矩阵，$B=(b_{ij})$是一个$s\times n$ 矩阵，定义矩阵A与矩阵B的乘积 $C=(c_{ij})$是一个$m\times n$ 矩阵（$i=1,2,\cdots,m;\ j=1,2,\cdots,n$），其中

$$c_{ij}=a_{i1}b_{1j}+a_{i2}b_{2j}+\cdots+a_{is}b_{sj}=\sum_{k=1}^{s}a_{ik}b_{kj}$$

并把此乘积记作$C=AB$。记号AB常读作A左乘B或B右乘A

注意: 只有当第一个矩阵的列数等于第二个矩阵的行数时，两个矩阵才能相乘

矩阵的运算-乘法

- **性质:**
  - 结合律： $A(BC)=(AB)C$
  - 分配率： $A(B+C)=AB+AC,\qquad (B+C)A=BA+CA$
  - $\lambda(AB)=(\lambda A)B=A(\lambda B)\quad AE=EA=A$
- **注意：** (1)矩阵乘法不满足交换律，即: $AB\neq BA$

$$A=\begin{pmatrix}1&1\\0&2\end{pmatrix},B=\begin{pmatrix}0&0\\2&1\end{pmatrix},\text{但}AB=\begin{pmatrix}2&1\\4&2\end{pmatrix}\neq BA=\begin{pmatrix}0&0\\2&4\end{pmatrix}$$

(2)若$AB=0$；不能推出$A=0$或$B=0$

$$A=\begin{pmatrix}1&0\\0&0\end{pmatrix},B=\begin{pmatrix}0&0\\0&1\end{pmatrix},AB=\begin{pmatrix}0&0\\0&0\end{pmatrix},\quad \text{但}A\neq 0,B\neq 0.$$

(3) 不满足消去率，即若$AB=AC$且$A!=0$，不能推出$B=C$

$$A=\begin{pmatrix}1&0\\0&0\end{pmatrix},B=\begin{pmatrix}0&0\\0&1\end{pmatrix},C=\begin{pmatrix}0&0\\0&0\end{pmatrix}.\quad AB=AC,\text{但}B\neq C$$

矩阵的运算-幂

- **定义:**若A是n 阶方阵, 则A^k为A的k次幂，即

$$A^{k+1} = A^k A = \underbrace{AA\cdots A}_{k} A$$

且满足幂运算律: $A^kA^m=A^{k+m}$, $(A^m)^k=A^{mk}$, 其中k, m为正整数

- **注意:** 由于矩阵乘法不满足交换律，则:

$(1)(AB)^k \neq A^k B^k$

$(2)A^2 - B^2 \neq (A+B)(A-B)$

$(3)(A+B)^2 \neq A^2 + 2AB + B^2$

$(4)(A-B)^2 \neq A^2 - 2AB + B^2$

矩阵的运算-转置

- **定义:**把矩阵A的行换成同序数的列得到一个新矩阵，叫做A的转置矩阵，记作A^{T}

$$A=\begin{pmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{pmatrix},\qquad A^{T}=\begin{pmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{pmatrix}$$

- **性质** (假设运算都是可行的) :

(1) $(A^{T})^{T}=A$;　　(2) $(A+B)^{T}=A^{T}+B^{T}$;

(3) $(\lambda A)^{T}=\lambda A^{T}$;　　(4) $(AB)^{T}=B^{T}A^{T}$;

矩阵的运算-共轭矩阵

- **定义**: 当 $A=(a_{ij})$ 为复矩阵时, 用 $\overline{a_{ij}}$ 表示 a_{ij} 的共轭复数, 记 $\overline{A}=(\overline{a_{ij}})$ ，称 $\overline{A}$ 为 A 的**共轭矩阵**

- **性质**(设 A, B 为复矩阵, λ 为复数, 且运算都是可行的):

$$(1)\quad \overline{A+B}=\overline{A}+\overline{B}$$

$$(2)\quad \overline{\lambda A}=\overline{\lambda}\,\overline{A}$$

$$(3)\quad \overline{AB}=\overline{A}\,\overline{B}$$

$$(4)\quad \overline{(A^T)}=(\overline{A})^T$$

矩阵的运算-逆矩阵

- **定义:**对于n阶矩阵A，如果有一个n阶矩阵B，使$AB = BA = E$，则说矩阵A是可逆的，并把矩阵B称为A的逆矩阵，简称逆阵。记作：$A^{-1} = B$

  唯一性：若A是可逆矩阵，则A的逆矩阵是唯一的

- **性质**

(1) 若矩阵A可逆，则A^{-1}亦可逆，且 $(A^{-1})^{-1} = A$

(2) 若矩阵A可逆，且$\lambda \neq 0$，则 λA 亦可逆，且

(3) 若A, B为同阶可逆方阵，则AB亦可逆， $(AB)^{-1} = B^{-1}A^{-1}$

(4) 若矩阵A可逆，则A^T亦可逆，且$(A^T)^{-1} = (A^{-1})^T$.

(5) 若矩阵A可逆， 则有$|A^{-1}| = |A|^{-1}$

矩阵的运算-秩

- **定义-k阶子式:**在$m\times n$矩阵A中任取 k 行 k 列($k\le m, k\le n$)，位于这 k 行 k 列交叉处的k^2个元素，不改变它们在A中所处的位置次序而得到的 k 阶行列式，被称为矩阵A的k阶子式

- **定义:**设在矩阵A中有一个不等于0的r阶子式D，且所有$r+1$阶子式(如果存在的话)全等于0，那么D称为矩阵A的一个最高阶非零子式，数r称为**矩阵A的秩，**记作 $R(A)$

矩阵的运算-秩

- 规定：零矩阵的秩等于0
- 说明：$m\times n$矩阵A的秩$R(A)$是A中不等于零的子式的最高阶数
- 可逆矩阵的秩等于阶数。故又称可逆(非奇异)矩阵为满秩矩阵，奇异矩阵又称为降秩矩阵

性质

1: $0 \le R(A_{m\times n}) \le \min\{m, n\}$

2: $R(A^T) = R(A)$

3: 若$A \sim B$, 则$R(A) = R(B)$

4: 若P, Q可逆, 则$R(PAQ) = R(A)$

5: $\max\{R(A), R(B)\} \le R(A\,|\,B) \le R(A) + R(B)$

6: $R(A+B) \le R(A) + R(B)$

7: $R(AB) \le \min\{R(A), R(B)\}$

8: 若$A_{m\times n}B_{n\times l} = O$, 则$R(A)+R(B) \le n$

矩阵的运算-迹

- **定义:** 如果一个矩阵A是 $n \times n$ 的方阵，则该矩阵的迹(trace) 为

$$\mathbf{tr}(\mathrm{A})=\sum_{i=1}^{n} a_{ii}$$

即：等于所有主对角线元素之和，一个实数的迹是它本身

- **性质:**
  - $\mathbf{tr}(\mathrm{A}^{\mathrm{T}}) = \mathbf{tr}(\mathrm{A})$
  - $\mathbf{tr}(\mathrm{AB}) = \mathbf{tr}(\mathrm{AB})$
  - $\mathbf{tr}(\mathrm{ABC}) = \mathbf{tr}(\mathrm{CAB}) = \mathbf{tr}(\mathrm{BCA})$

矩阵的运算-求导

- **定义:**针对函数$\mathbf{y}=\Psi(\mathbf{x})$

其中 $\boldsymbol{y}\in R^{m}\times 1$，$\boldsymbol{x}\in R^{n}\times 1$，则向量$\boldsymbol{y}$关于$\boldsymbol{x}$的导数可以表示为:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}=\begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

该矩阵也称为雅可比(Jacobian)矩阵(m×n)；

- 如果x 是一个标量，则雅可比矩阵是一个m×1的矩阵
- 如果y是一个标量，则雅可比矩阵是一个1×n的矩阵

矩阵的运算-求导

- 如果$y \in R^{m} \times 1$，$x \in R^{n} \times 1$，$A \in R^{m \times n}$， $y=Ax$ ，则 $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$
- 如果 x 是关于 z 的函数，$y=Ax$，则 $\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}$
- 如果：$\alpha=y^{T}Ax$ 则：$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^{T}\mathbf{A}$ $\frac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^{T}\mathbf{A}^{T}$
- 如果：$\alpha=x^{T}Ax$, $x \in R^{n} \times 1$, $A \in R^{n \times n}$ 则： $\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^{T}(\mathbf{A}+\mathbf{A}^{T})$
- 设 $\alpha=y^{T}x$, 其中 x 和 y 是关于 z 的函数，则

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{x}^{T}\frac{\partial \mathbf{y}}{\partial \mathbf{z}} + \mathbf{y}^{T}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}$$