数学基础

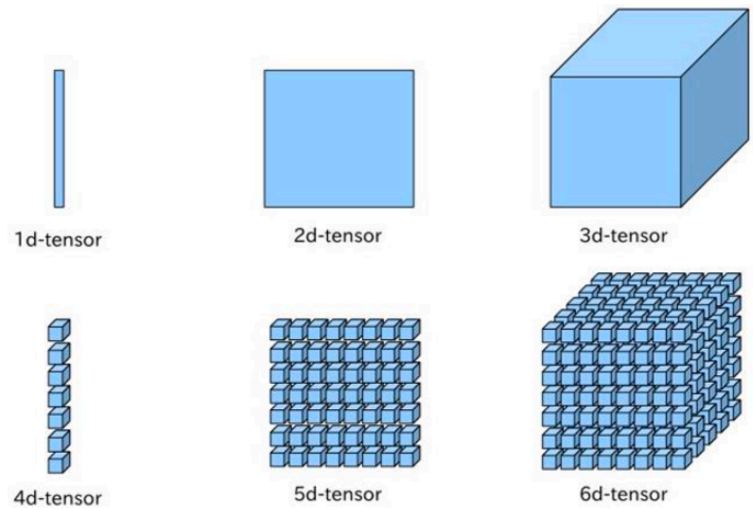
1. 张量、矩阵运算、矩阵的基础知识、矩阵分解
2. 概率统计、常见的（多变量）分布
3. 信息论、熵、互信息、相对熵、交叉熵
4. 最优化估计方法、最小二乘、线性模型

矩阵论

矩阵基本知识

矩阵：是一个二维数组，其中的每一个元素一般由两个索引来确定一般用大写变量表示，m行n列的实数矩阵，记做$A \in R_{m\times n}$.

张量(Tensor)：是矢量概念的推广，可用来表示在一些矢量、标量和其他张量之间的线性关系的多线性函数。标量是0阶张量，矢量是一阶张量，矩阵是二阶张量，三维及以上数组一般称为张量。



矩阵的秩(Rank)：矩阵列向量中的极大线性无关组的数目，记作矩阵的列秩，同样可以定义行秩。行秩=列秩=矩阵的秩，通常记作rank(A)。

矩阵的逆

- 若矩阵A为方阵，当 $rank(A_{n\times n}) < n$时，称A为奇异矩阵或不可逆矩阵；
- 若矩阵A为方阵，当 $rank(A_{n\times n}) = n$时，称A为非奇异矩阵或可逆矩阵

其逆矩阵 A^{-1} 满足以下条件，则称 A^{-1} 为矩阵A的逆矩阵：

$$AA^{-1} = A^{-1}A = I_n$$

其中 I_n 是 $n\times n$ 的单位阵。

矩阵的广义逆矩阵

- 如果矩阵不为方阵或者是奇异矩阵，不存在逆矩阵，但是可以计算其广义逆矩阵或者伪逆矩阵；
- 对于矩阵A，如果存在矩阵 B 使得 $ABA = A$，则称 B 为 A 的广义逆矩阵。

矩阵分解

机器学习中常见的矩阵分解有特征分解和奇异值分解。

先提一下矩阵的特征值和特征向量的定义

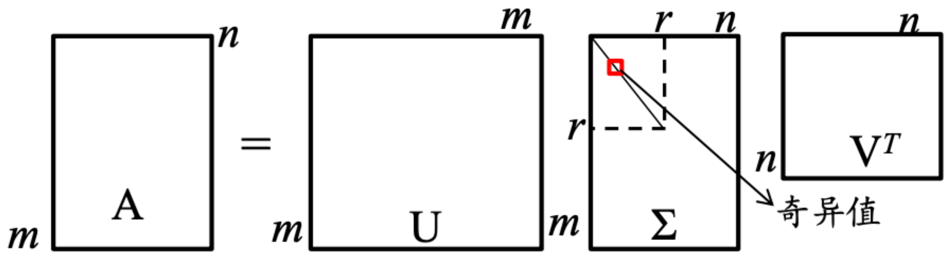
- 若矩阵 A 为方阵，则存在非零向量 x 和常数 λ 满足 $Ax=\lambda x$，则称 λ 为矩阵 A 的一个特征值，x 为矩阵 A 关于 λ 的特征向量。
- $A_{n\times n}$ 的矩阵具有 n 个特征值，$\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$ 其对应的n个特征向量为 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$， $\cdots$， $\boldsymbol{u}_n$
- 矩阵的迹(trace)和行列式(determinant)的值分别为

$$\operatorname{tr}(\mathrm{A})=\sum_{i=1}^{n}\lambda_i \quad |\,\mathrm{A}|=\prod_{i=1}^{n}\lambda_i$$

矩阵特征分解：$A_{n\times n}$ 的矩阵具有 n 个不同的特征值，那么矩阵A可以分解为 $A=U\Sigma U^T$.

其中 $\Sigma=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ 0 & 0 & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{bmatrix} \quad \mathrm{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n] \quad \|\boldsymbol{u}_i\|_2=1\,.$

奇异值分解：对于任意矩阵 $A_{m\times n}$，存在正交矩阵 $U_{m\times m}$ 和 $V_{n\times n}$，使其满足 $A=U\Sigma V^T \quad U^TU=V^TV=I$，则称上式为矩阵 A 的特征分解。



概率统计

随机变量

随机变量(Random variable)是随机事件的数量表现，随机事件数量化的好处是可以用数学分析的方法来研究随机现象。

随机变量可以是离散的或者连续的，离散随机变量是指拥有有限个或者可列无限多个状态的随机变量，连续随机变量是指变量值不可随机列举出来的随机变量，一般取实数值。

随机变量通常用概率分布来指定它的每个状态的可能性。

举例：

1. 投掷一枚硬币为正面是离散型随机事件X，发生概率P(X=1)=0.5
2. 每次射箭距离靶心的距离X可以认为连续型随机变量，距离靶心小 于1cm的概率P(X<1cm)

常见的概率分布

伯努利分布

- 伯努利试验：只可能有两种结果的单次随机实验
- 又称0-1分布，单个二值型离散随机变量的分布
- 其概率分布：$P*(*X=1)=p, P(X=0)=1-p.$

二项分布

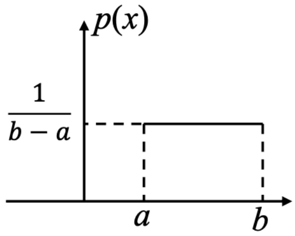
- 二项分布即重复*n*次伯努利试验，各试验之间都相互独立
- 如果每次试验时，事件发生的概率为*p*，不发生的概率为1-p，则*n*次重复独立试验中事件发生*k*次的概率为

$$P(X=k)=C_n^k p^k (1-p)^{n-k}$$

均匀分布

均匀分布，又称矩形分布，在给定长度间隔[*a*,*b*]内的分布概率是等可能的，均匀分布由参数*a*，*b*定义，概率密度函数为：

$$p(x)=\frac{1}{b-a}, \quad a<x<b$$



高斯分布

高斯分布，又称正态分布(normal)，是实数中最常用的分布，由均值*μ*和标准差*σ*决定其分布，概率密度函数为：

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

指数分布

常用来表示独立随机事件发生的时间间隔，参数为λ>0的指数分布概率密度函数为：$p(x)=\lambda e^{-\lambda x} \quad x \geq 0.$ 指数分布重要特征是无记忆性。

多变量概率分布

条件概率(Conditional probability)：事件X在事件Y发生的条件下发生的概率，P(X|Y)

联合概率(Joint probability)：表示两个事件X和Y共同发生的概率，P(X,Y)

条件概率和联合概率的性质：$P(Y|X)=\frac{P(Y,X)}{P(X)} \quad P(X)>0.$

推广到 n 个事件，条件概率的链式法则：

$$
\begin{aligned}
P\left(X_{1}, X_{2}, \ldots, X_{n}\right) &= P\left(X_{1} \mid X_{2}, \ldots, X_{n}\right) P\left(X_{2} \mid X_{3}, X_{4}, \ldots, X_{n}\right) \ldots P\left(X_{n-1} \mid X_{n}\right) P\left(X_{n}\right) \\
&= P\left(X_{n}\right) \prod_{i=1}^{n-1} P\left(X_{i} \mid X_{i+1}, \ldots, X_{n}\right)
\end{aligned}
$$

先验概率(Prior probability)：根据以往经验和分析得到的概率，在事件发生前已知，它往往作为“由因求果”问题中的“因”出现。

后验概率(Posterior probability)：指得到“结果”的信息后重新修正的概率，是“执果寻因”问题中 的“因”，后验概率是基于新的信息，修正后来的先验概率所获得 的更接近实际情况的概率估计。

举例说明：一口袋里有3只红球、2只白球，采用不放回方式摸取，求:
(1) 第一次摸到红球(记作A)的概率;
(2) 第二次摸到红球(记作B)的概率;
(3) 已知第二次摸到了红球，求第一次摸到的是红球的概率?

解：(1) $P(A=1)=3/5$，这就是先验概率;
(2) $P(B=1)=P(A=1)P(B=1|A=1)+P(A=0)P(B=1|A=0)=\frac{3}{5}\frac{2}{4}+\frac{2}{5}\frac{3}{4}=\frac{3}{5}$
(3) $P(A=1|B=1)=\frac{P(A=1)P(B=1|A=1)}{P(B=1)}=\frac{1}{2}$，这就是后验概率。

全概率公式：设事件$\{A_i\}$是样本空间 Ω 的一个划分，且$P(A_i)>0(i=1,2,\ldots,n)$，那么：
$P(B)=\sum_{i=1}^{n}P(A_i)P(B|A_i)$.

贝叶斯公式：全概率公式给我们提供了计算后验概率的途径，即贝叶斯公式

$$
P(\mathrm{~A}_{i} \mid \mathrm{B})=\frac{P(\mathrm{~B} \mid \mathrm{A}_{i}) P(\mathrm{~A}_{i})}{P(\mathrm{~B})}=\frac{P(\mathrm{~B} \mid \mathrm{A}_{i}) P(\mathrm{~A}_{i})}{\sum_{j=1}^{n} P(\mathrm{~A}_{j}) P(\mathrm{~B} \mid \mathrm{A}_{j})}
$$

常用统计量

方差：用来衡量随机变量与数学期望之间的偏离程度。统计中的方差则为样本方差，是各个样本数据分别与其平均数之差 的平方和的平均数，计算过程为：

$$
\operatorname{Var}(X)=E\left\{[x-E(x)]^{2}\right\}=E\left(x^{2}\right)-[E(x)]^{2}
$$

协方差：衡量两个随机变量X和Y直接的总体误差，计算过程为：

$$
\operatorname{Cov}(X, Y)=E\{[x-E(x)][y-E(y)]\}=E(xy)-E(x)E(y)
$$

信息论

熵

信息熵，可以看作是样本集合纯度一种指标，也可以认为是样本集合包含的平均信息量。

联合熵

条件熵

互信息

相对熵

交叉熵

最优化估计

最小二乘估计

最小二乘估计又称最小平方法，是一种数学优化方法。它通过最小化误差的平方和寻找数据的最佳函数匹配。最小二乘法经常应用于回归问题，可以方便地求得未知参数，比如曲线拟合、最小化能量或者最大化熵等问题。