数据科学的基石：统计学、机器学习、计算...

> 大多数人学习数据科学的重心放在编程上面，然而，要真正精通数据科学的话是不能够忽视数据科学背后的数据基础，理解背后的数学原理会帮助你更好地理解数据科学。今天的内容我们一起学习《数学基础：线代、概率论、微积分》。内容来自于笔者阅读各类数据科学相关书籍的读书摘录笔记，希望能够对数据分析行业从业者起到点滴帮助，由于笔者水平能力有限，整理的不妥之处请各位大佬批评指正！如涉版权问题请及时联系删除，谢谢！欢迎转发分享学习！——2019年12月8日

目录

各种算法和理论用到的数学知识

算法或理论	用到的数学知识点
贝叶斯分类器	随机变量，贝叶斯公式，随机变量独立性，正态分布，最大似然估计
决策树	概率，熵，Gini 系数
KNN 算法	距离函数
主成分分析	协方差矩阵，散布矩阵，拉格朗日乘数法，特征值与特征向量
流形学习	流形，最优化，测地线，测地距离，图，特征值与特征向量
线性判别分析	散度矩阵，逆矩阵，拉格朗日乘数法，特征值与特征向量
支持向量机	点到平面的距离，Slater 条件，强对偶，拉格朗日对偶，KKT 条件，凸优化，核函数，Mercer 条件
logistic	概率，随机变量，最大似然估计，梯度下降法，凸优化，牛顿法
随机森林	抽样，方差
AdaBoost 算法	概率，随机变量，极值定理，数学期望，牛顿法
隐马尔科夫模型	概率，离散型随机变量，条件概率，随机变量独立性，拉格朗日乘数法，最大似然估计
条件随机场	条件概率，数学期望，最大似然估计
高斯混合模型	正态分布，最大似然估计，Jensen 不等式
人工神经网络	梯度下降法，链式法则
卷积神经网络	梯度下降法，链式法则
循环神经网络	梯度下降法，链式法则
生成对抗网络	梯度下降法，链式法则，极值定理，Kullback-Leibler 散度，Jensen-Shannon 散度，测地距离，条件分布，互信息
K-means 算法	距离函数
贝叶斯网络	条件概率，贝叶斯公式，图
VC 维	Hoeffding 不等式

微积分

核心问题

极值问题与条件最优化。

核心技能

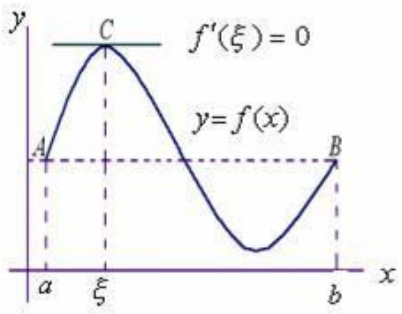
极限	极限是高等数学和初等数学的分水岭，也是微积分这座大厦的基石，是导数、微分、积分等概念的基础。虽然在机器学习里不直接用到极限的知识，但要理解导数和积分，它是必须的。
上确界与下确界	这一对概念对工科的微积分来说是陌生的，但在机器学习中会经常用到，不要看到论文或书里的sup和inf不知道什么意思。
导数	其重要性众所周知，求函数的极值需要它，分析函数的性质需要它。典型的如梯度下降法的推导，logistic函数导数的计算。熟练地计算函数的导数是基本功。
Lipschitz连续性	这一概念在工科教材中同样没有提及，但对分析算法的性质却很有用，在GAN，深度学习算法的稳定性、泛化性能分析中都有用武之地。
导数与函数的单调性	导数与函数的单调性。某些算法的推导，如神经网络的激活函数，AdaBoost算法，都需要研究函数的单调性。
导数与函数的极值	这个在机器学习中处于中心地位，大部分优化问题都是连续优化问题，因此可以通过求导数为0的点而求函数的极值，以实现最小化损失函数，最大化似然函数等目标。

导数与函数的凹凸性	导数与函数的凹凸性。在凸优化，Jensen不等式的证明中都有它的应用。
泰勒公式	又一个核心知识点。在优化算法中广泛使用，从梯度下降法，牛顿法，拟牛顿法，到AdaBoost算法，梯度提升算法，XGBoost的推导都离不开它。
不定积分	积分在机器学习中使用的相对较少，主要用于概率的计算中，它是定积分的基础。 定积分。包括广义积分，被用于概率论的计算中。机器学习中很大一类算法是概率型算法，如贝叶斯分类器，概率图模型，变分推断等。这些地方都涉及到对概率密度函数进行积分。
变上限积分	分布函数是典型的变上线积分函数，同样主要用于概率计算中。
牛顿-莱布尼兹公式	在机器学习中很少直接使用，但它是微积分中最重要的公式之一，为定积分的计算提供了依据。
常微分方程	在某些论文中会使用，但一般算法用不到。
偏导数	重要性不用多说，机器学习里绝大部分函数都是多元函数，要求其极值，偏导数是绕不开的。
梯度	决定了多元函数的单调性和极值，梯度下降法的推导离不开它。几乎所有连续优化算法都需要计算函数的梯度值，且以寻找梯度为0的点作为目标。
高阶偏导数	确定函数的极值离不开它，光有梯度值还无法确定函数的极值。
链式法则	同样使用广泛，各种神经网络的反向传播算法都依赖于链式法则。
Hessian矩阵	决定了函数的极值和凹凸性，对使用工科教材的同学可能是陌生的。
多元函数的极值判别法则	多元函数的极值判别法则。虽然不直接使用，但对理解最优化方法至关重要。
多元函数的凹凸性判别法则	证明一个问题是凸优化问题是离不开它的。
Jacobian矩阵	工科教材一般没有介绍这一概念，但和Hessian矩阵一样，并不难理解，使用它可以简化多元复合函数的求导公式，在反向传播算法中广泛使用。
向量与矩阵求导	常见的一次函数，二次函数的梯度，Hessian矩阵的计算公式要烂熟于心，推导并不复杂。
泰勒公式	理解梯度下降法，牛顿法的优化算法的基石。
多重积分	主要用于概率论中，计算随机向量的积分，如正态分布。
偏微分方程	在某些理论推导中可能会使用，如变分法中的欧拉-拉格朗日方程。

导数 / 偏导数梯度常用定理及简介

罗尔定理

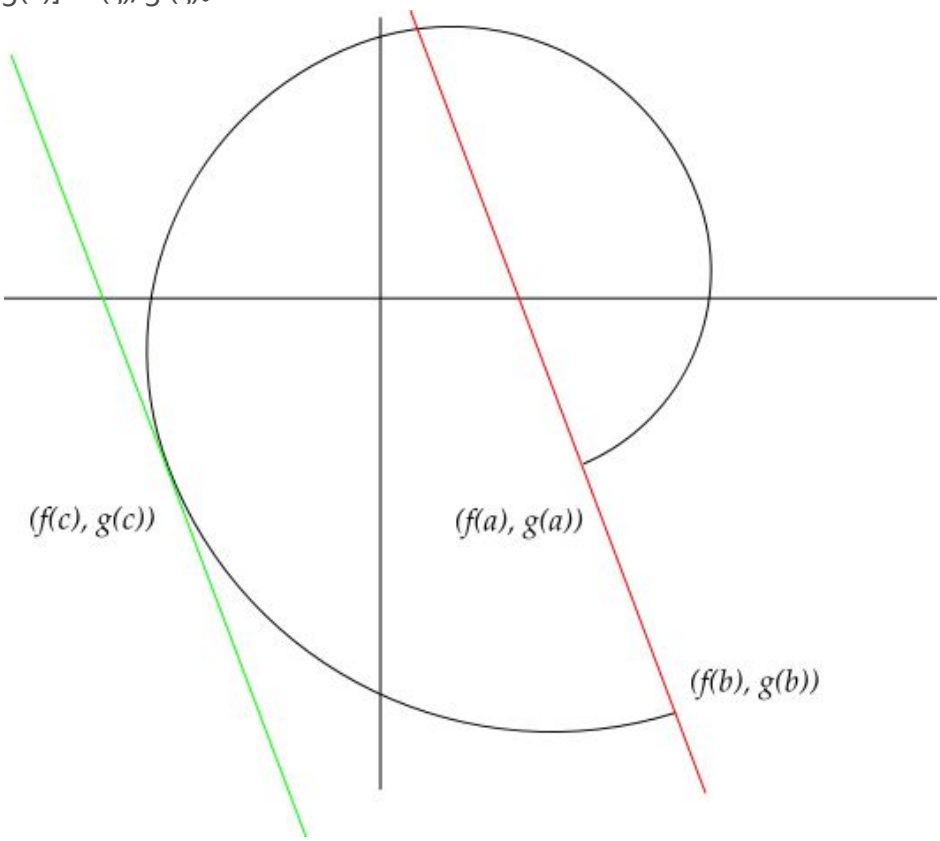
定理内容：如果 R 上的函数 f(x) 满足以下条件：（1）在闭区间 [a,b] 上连续，（2）在开区间 (a,b) 内可导，（3）f(a)=f(b)，则至少存在一个 $\xi \in (a,b)$，使得 $f'(\xi)=0$。



几何意义：若连续曲线y=f(x) 在区间 [a,b] 上所对应的弧段 AB，除端点外处处具有不垂直于 x 轴的切线，且在弧的两个端点 A,B 处的纵坐标相等，则在弧 AB 上至少有一点 C，使曲线在C点处的切线平行于 x 轴。

柯西中值定理

如果函数 f(x), g(x) 满足：（1）在闭区间 [a,b] 上连续，（2）在开区间 (a,b) 内可导，（3）对任一 x∈(a,b) 有 g'(x)≠0，则存在 ξ∈(a,b)，使得 [f(b)-f(a)]/[g(b) g(a)]=f'(ξ)/g'(ξ)。

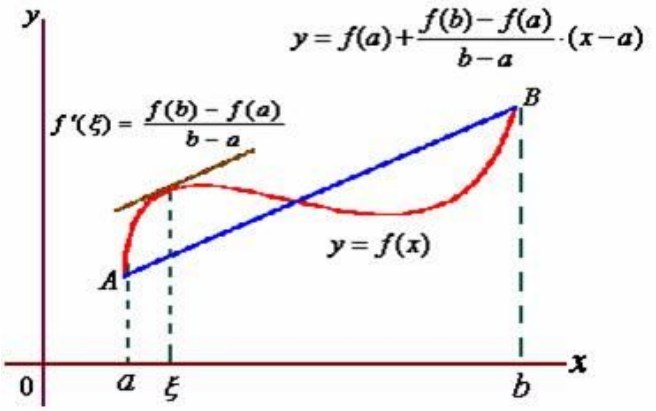


几何意义：若令 u=f(x), v=g(x），这个形式可理解为参数方程，而 [f(a)-f(b)]/[g(a)-g(b)] 则是连接参数曲线的端点斜率，f'(ξ)/g'(ξ) 表示曲线上某点处的切线斜率，在定理的条件下，可理解为：用参数方程表示的曲线上至少有一点，它的切线平行于两端点所在的弦。

拉格朗日中值定理

拉格朗日中值定理是罗尔中值定理的推广，同时也是柯西中值定理的特殊情形，是泰勒公式的弱形式（一阶展开），它反映了可导函数在闭区间上的整体的平均变化率与区间内某点的局部变化率的关系。

定理内容：如果函数 f(x) 满足：（1）在 (a,b) 内可导，（2）[a,b] 上连续，则必有一个 ξ∈(a,b)，使得 f'(ξ)*(b-a)=f(b)-f(a)。



几何意义：若连续曲线 y=f(x) 在 A(a,f(a)), B(b,f(b)) 两点间的每一点处都有不垂直于x轴的切线，则曲线在 A,B 间至少存在一个点 P(c,f(c))，使得该曲线在P点的切线与割线AB平行。

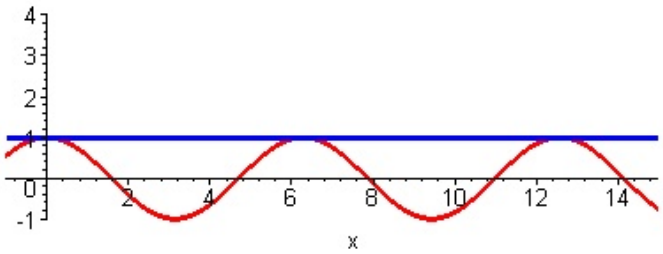
物理意义：对于直线运动，在任意一个运动过程中至少存在一个位置（或一个时刻）的瞬时速度等于这个过程中的平均速度。

泰勒展开

泰勒公式是一个用函数在某点的信息描述其附近取值的公式。如果函数足够平滑的话，在已知函数在某一点的各阶导数值的情况之下，泰勒公式可以用这些导数值做系数构建一个多项式来近似函数在这一点的邻域中的值。泰勒公式还给出了这个多项式和实际的函数值之间的偏差。

$$f(x)=\frac{f(x_0)}{0!}+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\ldots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$$ 泰勒公式

用一张动图模拟这个过程：



夹逼准则

定理内容：若函数 $F(x)$ 和 $G(x)$ 在 x_0 的邻域连续， $x\to x_0$ 时极限都为 A ，即 $\lim\limits_{x\to x_0}F(x)=\lim\limits_{x\to x_0}G(x)=A$ ，且在该 x_0 的邻域一直满足 $F(x)\le f(x)\le G(x)$ 。则当 $x\to x_0$ 时也有 $\lim\limits_{x\to x_0}F(x)\le\lim\limits_{x\to x_0}f(x)\le\lim\limits_{x\to x_0}G(x)$ ，也就是 $A\le\lim\limits_{x\to x_0}f(x)\le A$ ，

所以 $\lim\limits_{x\to x_0}f(x)=A$.

简单地说：函数 A>B，函数 B>C，函数 A 的极限是 X，函数 C 的极限也是 X ，那么函数 B 的极限就一定是 X，这就是夹逼定理。
洛必达法则
定理内容：设
（1）当 $x \to \infty$ 时，函数 $f(x)$ 及 $F(x)$ 都趋于零；
（2）当 $|x| > N$ 时 $f'(x)$ 与 $F'(x)$ 都存在，且 $F'(x) \neq 0$ ；
（3） $\lim_{x \to \infty} \frac{f'(x)}{F'(x)}$ 存在（或为无穷大），
那么 $\lim_{x \to \infty} \frac{f(x)}{F(x)} = \lim_{x \to \infty} \frac{f'(x)}{F'(x)}$
洛必达法则是在一定条件下，通过分子分母分别求导再求极限来确定未定式值的方法。这种方法主要是在一定条件下通过分子分母分别求导再求极限来确定未定式的值。
在运用洛必达法则之前，首先要完成两项任务：一是分子分母的极限是否都等于零（或者无穷大）；二是分子分母在限定的区域内是否分别可导。如果这两个条件都满足，接着求导并判断求导之后的极限是否存在：如果存在，直接得到答案；如果不存在，则说明此种未定式不可用洛必达法则来解决；如果不确定，即结果仍然为未定式，再在验证的基础上继续使用洛必达法则。

线性代数

核心问题

求多元方程组的解。

核心技能

向量及其运算	机器学习算法的输入很多时候是向量，如样本的特征向量
矩阵及其运算	与向量一样，是线性代数的核心概念，各种运算，常用矩阵，必须烂熟于心
行列式	直接使用的少，在概率论，某些模型的推导中偶尔使用
线性方程组	直接使用的少，但这是线性代数的核心内容
特征值与特征向量	在机器学习中被广泛使用，很多问题最后归结于求解矩阵的特征值和特征向量
广义特征值	工科线性代数教材一般不提及此概念，但在流形学习，谱聚类等算法中经常用到它
Rayleigh商	工科教材一般不提及它
矩阵的谱范数与条件数	工科教材一般不提及它
二次型	很多目标函数是二次函数，因此二次型的地位不言而喻
Cholesky分解	某些算法的推导中会用到它，工科教材一般不提及它
特征值分解	对机器学习非常重要，很多问题最后归结于特征值分解，如主成分分析，线性判别分析等
奇异值分解	在机器学习中广泛使用，从正态贝叶斯分类器，到主题模型等，都有它的影子

乘积、内积、秩

已知矩阵 A 和矩阵 B，求 A 和 B 的乘积 C=AB。

矩阵 A 大小为 mxn，矩阵 B 大小为 nxp。

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{bmatrix}$$

常规方法：矩阵 C 中每一个元素 Cij = A 的第i行 乘以（点乘）B 的第 j 列。

设有 n 维向量

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

令 $[x, y] = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$ ，称 $[x, y]$ 为向量 x 与 y 的内积。

在线代中秩的定义：

一个矩阵 A 的列秩是 A 的线性无关的列的极大数目。类似地，行秩是 A 的线性无关的行的极大数目。矩阵的列秩和行秩总是相等的，因此它们可以简单地称作矩阵 A 的秩，通常表示为 rank(A)。

矩阵列空间、行空间的维度相等。任意一个矩阵都可以经过一些列的初等变换为阶梯形矩阵，而且阶梯形矩阵的秩等于其中非零行的个数。

所以矩阵秩的计算方法：

用初等变换把矩阵华为阶梯形，则该阶梯形矩阵中的非零行数就是所求矩阵的秩。

例3 求矩阵 $B = \begin{pmatrix} 2 & -1 & 0 & 3 & -2 \\ 0 & 3 & 1 & -2 & 5 \\ 0 & 0 & 0 & 4 & -3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ 的秩.

解 ∵ B是一个行阶梯形矩阵，其非零行有3行，

∴ B 的所有4阶子式全为零.

而 $\begin{vmatrix} 2 & -1 & 3 \\ 0 & 3 & -2 \\ 0 & 0 & 4 \end{vmatrix} \neq 0$， ∴ $r(B) = 3$.

高斯消元法

高斯消元法（Gaussian Elimination），是线性代数中的一个算法，可用来为线性方程组求解，求出矩阵的秩，以及求出可逆方阵的逆矩阵。当用于一个矩阵时，高斯消元法会产生出一个“行梯阵式”。

值得提一下的是，虽然该方法以数学家卡尔·高斯命名，但最早出现于中国古籍《九章算术》，成书于约公元前150年。

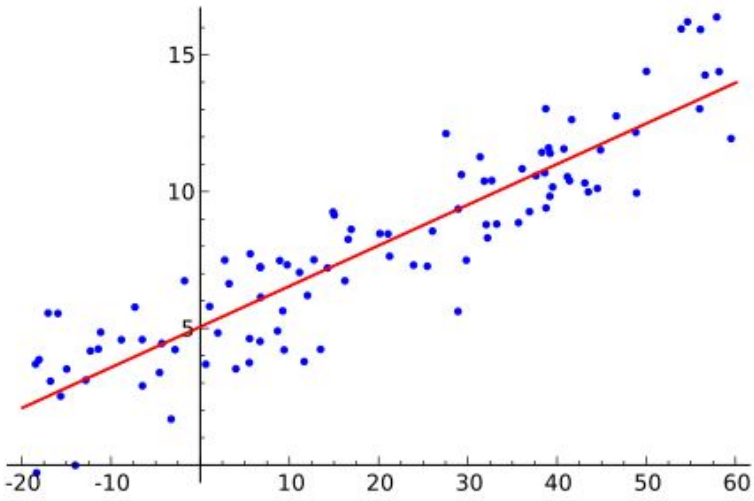
复杂度：高斯消元法的算法复杂度是 $O(n^3)$ ；这就是说，如果系数矩阵的是 n × n，那么高斯消元法所需要的计算量大约与 n^3 成比例。

矩阵求逆

求矩阵的逆矩阵的常用方法有两种：

伴随矩阵法初等变换法最小二乘法

最小二乘法是对过度确定系统，即其中存在比未知数更多的方程组，以回归分析求得近似解的标准方法。在这整个解决方案中，最小二乘法演算为每一方程式的结果中，将残差平方和的总和最小化。



回归分析模型

主要思想：选择未知参数，使得理论值与观测值之差的平方和达到最小：

$$H = \sum_{0}^{m} (y - y_i)^2$$

最重要的应用是在曲线拟合上。最小平方所涵义的最佳拟合，即残差（残差为：观测值与模型提供的拟合值之间的差距）平方总和的最小化。

概率论

核心问题

发现数字的隐藏规律，完成分类。

核心技能

随机事件与概率	这是理解随机变量的基础，也是概率论中最基本的知识
条件概率与独立性	条件概率非常重要，在机器学习中，只要有概率模型的地方，通常离不开它
条件独立	在概率论图模型中广泛使用，一定要理解它
全概率公式	基础公式，地位不用多说
贝叶斯公式	在机器学习的概率型算法中处于灵魂地位，几乎所有生成模型都要用到它
离散型随机变量与连续型随机变量	重要性不用多说，概率质量函数，概率密度函数，分布函数，一定要熟练掌握
数学期望	非常重要，好多地方都有它的影子
方差与标准差	非常重要，刻画概率分布的重要指标
Jensen不等式	在很多推导和证明中都要用它，如EM算法，变分推断
常用的概率分布，	包括均匀分布，正态分布，伯努利分布，二项分布，多项分布，t

	分布等，在各种机器学习算法中广泛使用
随机向量	多元的随机变量，在实际中更有用
协方差	经常使用的一个概念，如主成分分析，多元正态分布中
参数估计	包括最大似然估计，最大后验概率估计，贝叶斯估计，核密度估计，一定要弄清楚它们是怎么回事
随机算法	包括采样算法，遗传算法，蒙特卡洛算法，在机器学习中也经常使用
信息论中的一些概念	包括熵，交叉熵，KL散度，JS散度，互信息，信息增益，一定要深刻理解这些概念如果你不理解KL散度，那怎么理解变分推断和VAE?

最大似然估计

给定一个概率分布 D ，已知其概率密度函数（连续分布）或概率质量函数（离散分布）为 f_D ，以及一个分布参数 θ ，我们可以从这个分布中抽出一个具有 n 个值的采样 $X_1, X_2, \ldots, X_n$,利用 f_D 计算出其似然函数：

$lik(\theta|x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n).$

若 D 是离散分布， f_θ 即是在参数为 θ 时观测到这一采样的概率。若其是连续分布， f_θ 则为 $X_1, X_2, \ldots, X_n$ 联合分布的概率密度函数在观测值处的取值。一旦我们获得 $X_1, X_2, \ldots, X_n$, 我们就能求得一个关于 θ 的估计。最大似然估计会寻找关于 θ 的最可能值（即，在所有可能的 θ 取值中，寻找一个值使这个采样的“可能性”最大化）。从数学上来说，我们可以在 θ 的所有可能取值中寻找一个值使得似然函数取到最大值。这个使可能性最大的 $\tilde{\theta}$值即成为 θ 的最大似然估计。

⚠注意：1）这里的似然函数是指 $x_1, x_2, \ldots, x_n$ 不变时，关于 θ 的一个函数。

2）最大似然估计不一定存在，也不一定唯一。

贝叶斯模型

首先复习一下贝叶斯定理：贝叶斯定理是关于随机事件 A 和 B 的条件概率的一则定理。

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

其中 P(A|B) 是在 B 发生的情况下 A 发生的可能性。

在贝叶斯定理中，每个名次都有约定俗成的名称：

P(A|B) 是已知 B 发生后 A 的条件概率，也由于得自 B 的取值而被称作 A 的后验概率。

P(A) 是 A 的先验概率，之所以称为“先验“是因为它不考虑任何 B 方面的因素。

P(B|A) 是已知 A 发生后 B 的条件概率，也由于得自 A 的取值而被称作 B 的后验概率。

P(B) 是 B 的先验概率。

按这些术语，贝叶斯定理也可以表述为：

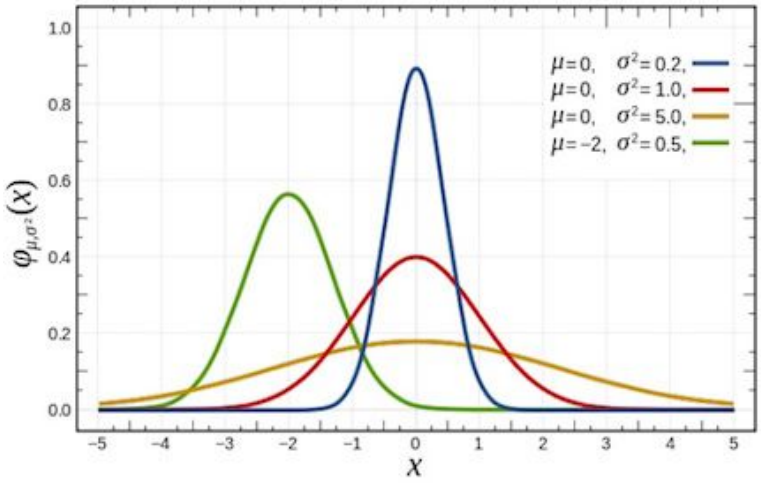
后验概率 = （相似度 * 先验概率）/ 标准化常量

也就是说，后验概率与先验概率和相似度的乘积成正比。

关于朴素贝叶斯算法的具体应用，看到一篇文章讲得很详细，点击这里传送~

高斯分布

高斯分布（Gaussian Distribution）， 也叫自然分布或正态分布。



若随机变量 X 服从一个数学期望为 μ 、标准方差为 σ^2 的高斯分布，记为：

$X \sim N(\mu, \sigma^2)$

则其概率密度函数为：

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

高斯分布的期望值 μ 决定了其位置，其标准差 σ 决定了分布的幅度。我们通常提到的标准正态分布是 $\mu = 0, \sigma = 1$ 的正态分布。

关于多元高斯分布在机器学习中的应用，具体可以参考这篇文章：多元高斯分布（Multivariate Gaussian Distribution）

显著性检验

显著性检验就是事先对总体（随机变量）的参数或总体分布形式做出一个假设，然后利用样本信息来判断这个假设（原假设）是否合理，即判断总体的真实情况与原假设是否显著地有差异。或者说，显著性检验要判断样本与我们对总体所做的假设之间的差异是纯属机会变异，还是由我们所做的假设与总体真实情况之间不一致所引起的。

P 值即概率，反映某一事件发生的可能性大小。统计学根据显著性检验方法所得到的 P 值，一般以 P < 0.05 为显著， P<0.01 为非常显著，其含义是样本间的差异由抽样误差所致的概率小于0.05 或0.01。

时间序列

时间序列（time series）是一组按照时间发生先后顺序进行排列的数据点序列。通常一组时间序列的时间间隔为一恒定值（如1秒，5分钟，12小时，7天，1年），因此时间序列可以作为离散时间数据进行分析处理。

BTC 价格走势

EM算法

EM算法，即最大希望算法（Expectation-maximization algorithm）。在统计计算中，EM算法是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐性变量。最大期望算法经常用在机器学习和计算机视觉的数据聚类（Data Clustering）领域。

EM算法经过两个步骤交替进行计算，第一步是计算期望（E），利用对隐藏变量的现有估计值，计算其最大似然估计值；第二步是最大化（M），最大化在E步上求得的最大似然值来计算参数的值。M步上找到的参数估计值被用于下一个E步计算中，这个过程不断交替进行。

蒙特卡洛

蒙特卡罗是一类随机方法的统称。这类方法的思想可以参考一个例子，用蒙特卡洛法求圆周率：

已知：一个半径为R的圆，它有一个边长为2R的外切正方形。

圆面积：pi*R^2，正方形面积：2R*2R=4R^2

在正方形内随机取一个点，要求每次取的点在正方形内任意一个点位置的概率都是平均分布的，那么这个点在圆内的概率大概为：pi*R^2/4R^2=pi/4

取若干个这样的点，利用平面上两点间的距离公式，计算这个点到圆心的距离，从而判断是否在圆内。

当我们统计过的点的个数足够多时，得到的概率值就会接近 pi/4，从而得到圆周率的值。

蒙特卡洛是依靠足够多次数的随机模拟，来得到近似结果的算法，说白了就是通过频率来估计概率。