# 微分

## 函數的微分



其中，稱為函數的增量，可以拆成兩部份組成：“線性主部”和“高級無窮小”。

其中，稱為“線性主部”，即“線性增長的主要部份”

其中，稱為“高階無窮小”。

那麼，把（線性主部）稱為函數的微分。

【疑問】：那是不是可以稱為導數呢？

## 微分的定義

如果在點的導數為，那么与自变量的改变量的乘积称为在点的微分。，就是的线性主部，即是微分。



其中，为无穷小量。

其中，表示“因为”。

其中，表示“可以推导出”



如果在点的极限是，那么就等于（其中为无穷小量），而不能写成。因为可以说在点的极限是，但不能说等于，必须要加上一个无穷小量。

其中，表示“可以互相推导”

# 概率论和统计

## 数学期望(mean)

在概率论和统计学中，数学期望(mean)（或均值，亦简称期望）是试验中每次可能结果的概率乘以其结果的总和，是最基本的数学特征之一。它反映随机变量平均取值的大小。

## 方差（variance)

方差（variance)是对随机变量或一组数据时离散程度的度量。**概率论**中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。**统计**中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。在许多实际问题中，研究方差即偏离程度有着重要意义。

### 历史

“方差”（variance）这一词语率先由罗纳德·费雪（Ronald Fisher）在其论文《The Correlation Between Relatives on the Supposition of Mendelian Inheritance》中提出。

### 方差的定义

方差在**统计描述**和**概率分布**中各有不同的定义，并有不同的公式。

在【**统计描述】**中，方差用来计算每一个变量（观察值）与总体均数之间的差异。为避免出现【离均差】总和为零，【离均差/的平方和】受样本数量的影响，统计学采用【平均/离均差/平方和】来描述变量的变异程度。【**总体方差**】计算公式：



其中，为【总体方差】，为【变量】，为【总体均值】， 为【总体例数】。

实际工作中，【总体均数】难以得到时，应用【样本统计量】代替【总体参数】，经校正后，【**样本方差**】计算公式：



其中，为【样本方差】，为【变量】，  为【样本均值】，为【样本例数】。

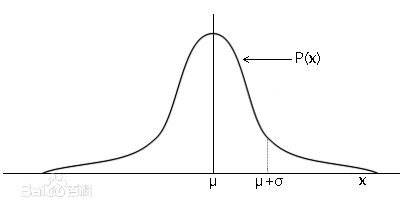
在【**概率分布**】中，设X是一个离散型随机变量，后面我看不懂了

## 标准差

标准差（Standard Deviation） ，中文环境中又常称【均方差】，是【离均差平方的算术平均数】的平方根，用表示。【标准差】是【方差】的算术平方根。【标准差】能反映一个数据集的离散程度。平均数相同的两组数据，【标准差】未必相同。

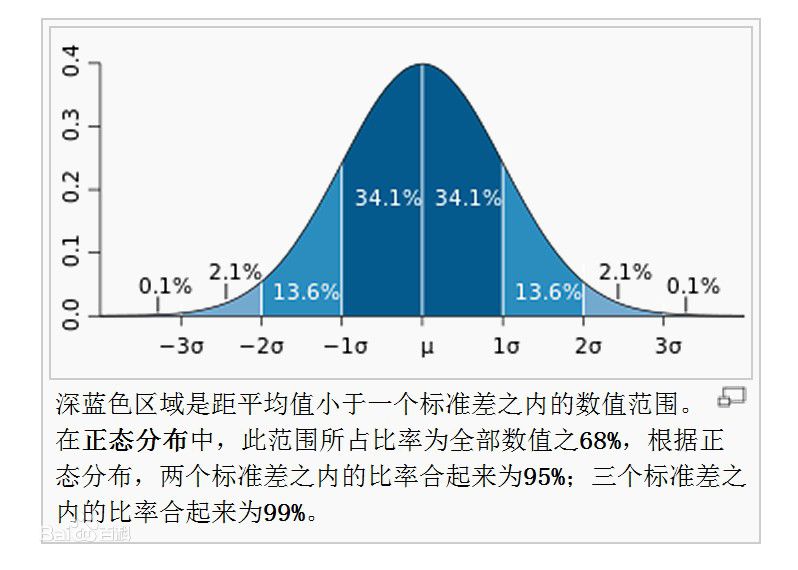
## 标准正态分布

【标准正态分布】（英语：standard normal distribution，），是一个在数学、物理及工程等领域都非常重要的概率分布，在【统计学】的许多方面有着重大的影响力。【期望值】，即曲线图象对称轴为Y轴，【标准差】条件下的正态分布，记为。



标准正态分布又称为u分布，是以0为【均数】、以1为【标准差】的正态分布，记为N（0，1）。

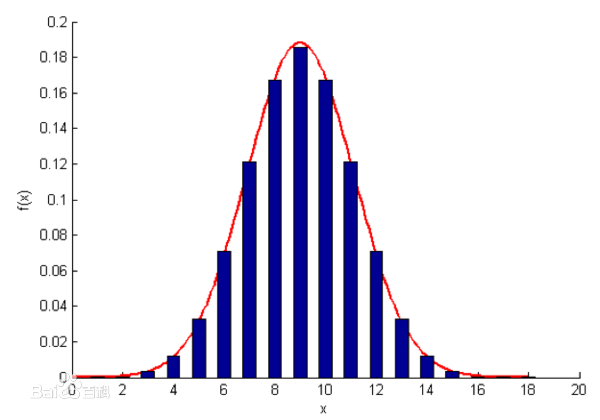
标准正态分布曲线下面积分布规律是：在-1.96～+1.96范围内曲线下的面积等于0.9500，在-2.58～+2.58范围内曲线下面积为0.9900。



在实际应用上，常考虑一组数据具有近似于【正态分布】的概率分布。若其假设正确，则约68.3%数值分布在距离【平均值】有【1个标准差之内的范围】，约95.4%数值分布在距离【平均值】有【2个标准差之内的范围】，以及约99.7%数值分布在距离【平均值】有【3个标准差之内的范围】。称为“68-95-99.7法则”或“经验法则”[1] 。

## 正态分布

正态分布（Normal distribution），也称“常态分布”，又名高斯分布（Gaussian distribution），最早由*A.棣莫弗*在求二项分布的渐近公式中得到。*C.F.高斯*在研究测量误差时从另一个角度导出了它。



【正态曲线】呈钟型，两头低，中间高，左右对称，因其曲线呈钟形，因此人们又经常称之为【钟形曲线】。

若随机变量X服从一个数学期望为、方差为的正态分布，记为。其【概率密度函数】为正态分布的【期望值】决定了其位置，其标准差决定了分布的幅度。当时的正态分布是【标准正态分布】。

### 定义

### 一维正态分布

若【随机变量】服从一个【位置参数】为、【尺度参数】为的【概率分布】，且其【概率密度函数】为



则这个【随机变量】就称为【正态随机变量】，【正态随机变量】服从的分布就称为【正态分布】，记作 ，读作服从 ，或服从【正态分布】。

### 标准正态分布

当时，【正态分布】就成为【标准正态分布】，其【概率密度函数】为



## 连续型随机变量

若【随机变量】的【分布函数】可表示成一个非负【可积函数】的积分，则称为【连续型随机变量】，称为的【概率密度函数】（分布密度函数）。

## 分布函数

分布函数（英文Cumulative Distribution Function, 简称CDF），是【概率统计】中重要的函数，正是通过它，可用【数学分析】的方法来研究【随机变量】。【分布函数】是【随机变量】最重要的概率特征，【分布函数】可以完整地描述【随机变量】的**统计规律**，并且决定【随机变量】的一切其他概率特征。

## 概率密度函数

在数学中，【连续型随机变量】的【概率密度函数】（probability density function是一个描述这个【随机变量】的输出值，在【某个确定的取值点】附近的可能性的函数。而【随机变量】的取值落在【某个区域之内的概率】则为【概率密度函数】在这个区域上的积分。

【lsieun解析】首先，有一个随机变量生成器X，它的作用是可以不断的产生随机数字。但是，随机生成器X的内部实现机制（即生成随机数的规律），我们并不知道，因此没有办法准确的判断下一个随机数字是多少。假如我们止步如此，可能就没有后续的内容的。但是，似乎有些数学家不甘心，他们认为“通过对历史数据的分析，可以预测未来”。这些数据家就想，既然无法准确判定下一个随机数是什么，但是可以查看已经生成的每一个随机数出现的频次，如果一个随机数出现的频次越高，说明它出现的概率比较高（由【频次】->【概率】转换）。因此，“准确预测下一个随机数是多少”的问题，就可以稍微变化一下“下一个随机数是【某个特定的数字】的概率是多少”的问题，还可以再进一步变化为“下一个随机数是【某个范围的数字】的概率是多少”的问题。

但是，本质上来说，随机是不存在的，只是人类本身无法把握其中的规律，而把这种无法预测的现象称之为“随机”。

## 最大似然估计

最大似然估计(maximum likelihood estimation, MLE)一种重要而普遍的求【估计量】的方法。【最大似然法】明确地使用【概率模型】，其目标是寻找能够以较高概率产生观察数据的系统发生树。最大似然法是一类完全基于统计的系统发生树重建方法的代表。

函数从到的定积分。

表示【从到逐一增加】对连加求和（sigma:∑）。

表示【从到逐一增加】对连乘求积（pi:Π）。

## 解析解

解析解(**analytical solution**)就是【一些严格的公式】，给出任意的【自变量】就可以求出【其因变量】，也就是问题的解，他人可以利用【这些公式】计算各自的问题。【解析解】也被称为闭式解(closed-form expression)。

所谓的【解析解】是一种包含【分式】、【三角函数】、【指数】、【对数】甚至【无限级数】等基本函数的解的形式。用来求得【解析解】的方法称为**解析法**。比如一元二次方程：，其求解公式是，这就是【解析解】。

线性回归方程的解析解如下：



损失函数常用



## 数值解

数值解(**numerical solution**)是采用如【有限元方法】、 【数值逼近方法】、【插值方法】等计算方法得到的解。只能利用【数值】计算的【结果】， 而不能随意给出【自变量】并求出【计算值】。当无法藉由微积分技巧求得【解析解】时，这时便只能利用【数值分析】的方式来求得其【数值解】了。

在解组件特性相关的方程式时，大多数的时候都要去解偏微分或积分式，才能求得其正确的解。依照求解方法的不同，可以分成以下两类：【**解析解**】和【**数值解**】。

**【数值解】与【解析解】的区别**

【数值解】是在特定条件下通过近似计算得出来的一个数值，而【解析解】为该函数的解析式。

【数值解】就是用【数值方法】求出【解】，给出一系列对应的自变量和解。

【解析解】就是给出【解】的【具体函数形式】，从解的表达式中就可以算出任何对应值。

## 损失函数(Loss Function)

在统计学，统计决策理论和经济学中，【损失函数】是指一种将【一个事件】（在一个样本空间中的一个元素）映射到一个表达/与其事件相关的经济成本/或机会成本/的实数上的一种函数。更通俗地说，在统计学中【损失函数】是一种**衡量损失和错误**（这种损失与“错误地”估计有关，如费用或者设备的损失）**程度的函数**。

【线性回归】-【梯度下降】-【均方误差】









逻辑回归





信息增益（information gain,IG）



信息增益计算



基尼系数



信息熵

Information\_entropy







# 概率（Probability）的本质是什么？

https://www.zhihu.com/question/26895086/answer/224503078

概率（Probability）的本质是什么？

概率存在两种【解释基础】：

1. 物理世界本身存在的随机性（**客观概率**）。

2. 是我们由于信息不足而对事件发生可能性的度量（**主观概率**）。

由两种解释建立起了**传统数理统计学（频率论学派）**和**贝叶斯统计学**。

题主认为一个【基本概念】的定义终究是确定且唯一的，那么**两种概率的解释之间的矛盾和关联是什么**？

我还是从生活中的例子，从以下三个方面聊下概率：

1.什么是概率？

2.如何计算概率？

3.概率对我有啥用？

1 什么是概率？

我们经常会在生活中听到这句话：**选择比努力更重要**。

相信你也无数次听过这句话，但是有没有想过：**这句话背后的真实含义是什么呢**？

我们每天拥有固定的时间和精力，注意这里“固定”是指你的资源和时间都是有限的。在这个前提下，把它们投入到哪些方向上能够取得最佳的效果，这是我们每天都要思考的问题。

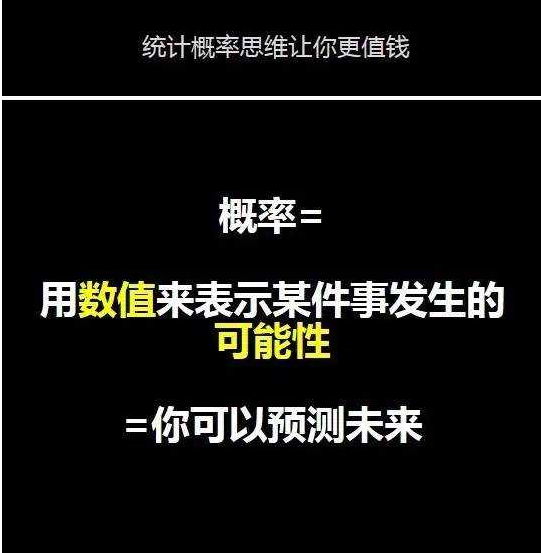
举个具体的例子，假如你刚毕业，已经拿到3家公司的offer，一家传统企业的职位，一家创业公司的职位，一家上市互联网公司的职位，你选哪个？因为你的时间和精力是有限的，所以你不能同时到这3个公司去上班赚钱，不然你就是孙悟空72变了。你只能在有限的资源和时间下，做一个最佳选择，这个选择代表你去哪家公司工作，未来3年内能让你赚钱能力提升的**可能性**最大。



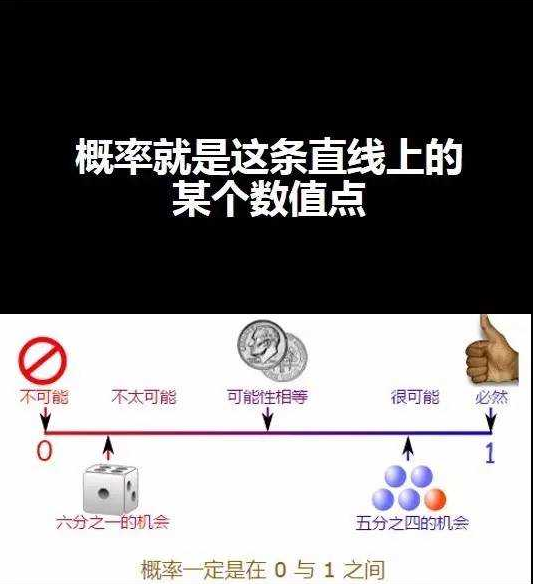
这里我们提到**可能性**，但是**只知道可能性是无法做出判断的**。

所以只知道可能性是不行的，这时候数学家就出来吼了一嗓子：**如果用某种办法计算出可能性，并用数值来表示这种可能性，不就解决了**。

这里用【数值】来衡量【可能性】就是下面图片里描述的概率。



**【概率】就是用【数值】来表示【某件事发生的可能性】**。 当你知道了概率这个数值，就代表你可以预测未来，因为你能通过概率来判断出哪种情况发生的可能性最大。



概率的值永远在0-1范围之间。

如果某件事不可能发生，则其概率为0，对应的就是这条直线上最左端的位置。

如果某件事肯定会发生，则其概率为1，对应的就是这条直线上最右端的位置，也就是那个点赞的大拇指。

大多数时候，你所面临的都是介于0和1之间的概率事件。

比如这条直线上更靠近左端的抛筛子，某一面数值朝上的概率是1/6。处于中间位置的是抛硬币，正面或者反面朝上的概率都是50%。靠近右端的从4个蓝色球，1个红色球里面选出4个蓝色球，正好都是蓝色球的概率是4/5。

解释概率的相关理论一般可以划分为两大传统：贝叶斯派和客观概率派。

（一）贝叶斯派（主观概率派）

贝叶斯派用【信念的强度】（degrees of partial belief）来定义【概率】。根据这个定义，【概率】并不是关于【物理系统】的，而是关于【物理系统】和【我们】之间的关系。

比如说，在经典力学的框架下，掷硬币这样的事件是完全决定性的（fully deterministic）：大概来说，硬币和其所在环境的组成的物理系统在某个时刻的状态是由其前一个时刻的状态决定的。如果我们知道这个系统的初始状态，知道组成这个系统每一个粒子最开始的速度和位置，原则上通过经典的动态方程，可以计算出这个系统在之后每一个时刻的状态。也就是说，硬币落地的朝向是完全由其初始状态和物理定律决定的；而如果知道硬币、掷硬币的手、周围空气的分布，硬币落下接触的地面等等每一个细节，原则上我们是可以准确预测出最后硬币是朝上还是朝下的。

但是，很明显，由于我们平时不知道这些细节，无法做出精准的预测，只能预测一个大概的结果，而这个结果就是通过概率的形式来表达的。

根据【贝叶斯派】，【概率】代表了**我们对于某个事件的信念**。如果我们相信这个事件一定会发生，概率则为1；如果我们相信这个事件一定不会发生，概率则为0；如果我们相信这个事件有可能发生，而测量关于它会发生这个信念的强度就是概率，介于0和1之间。

（二）客观概率派

相比【贝叶斯派】，【客观概率派】认为概率是关于【客观世界】的，关于【物理系统】的，**独立于人们对世界的信念**。

（2.1）原始派（Primitivism）

【原始派】宣称，【概率】是【单个物体】或者【整个系统】的一种【原始的属性】（primitive property），无法用非概率的语言来解释。比如在欧几里得几何学中，点就是一个原始概念，你无法解释点是什么。如果克鲁星人说不懂概率是什么的话，要么它们是在撒谎，要么对于它们而言没有任何可以理解概率的希望。为什么你会觉得我们可以用非概率的语言来解释概率是什么？

原始派一般和倾向派（propensity）被划分为同一个观点。倾向派认为作为原始属性代表了物理系统具有某种倾向（or disposition, tendency）。比如盐在水中会有溶解的倾向；硬币被抛后有朝上或者朝下的倾向。波普（对，可证伪的那个波普）就是一个倾向派。

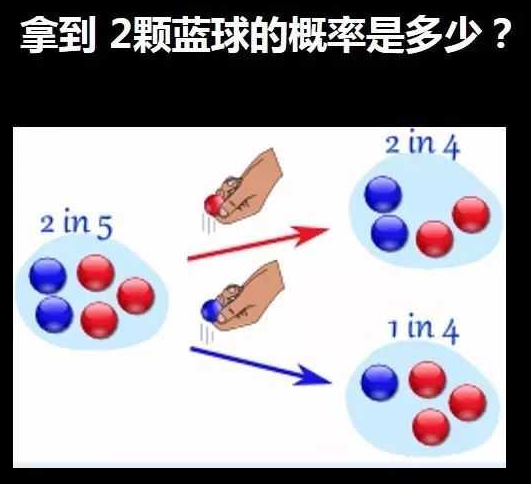
（2.2）频率派（Frequentism）

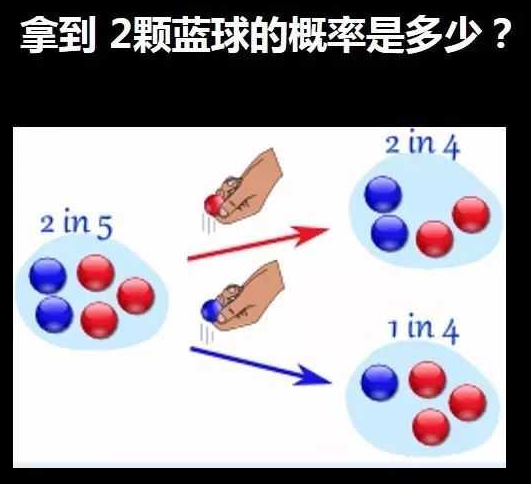
如名字所示，【频率派】直接将【概率】和【频率】化作等号。 频率派的问题其实在开头和克鲁星人的对话中已经有所提及了。概括来说就是，也许用频率来解读概率看起来符合直觉，但事实上频率和概率并不完全相等。

我们能做到的最好的证明是大数定则（the Law of Large Numbers），但大数定则并没有从真正意义上解决问题。

# 什么是条件概率？

【相关事件的概率】也叫【条件概率】，什么是【条件概率】呢？





在计算概率之前，我们需要弄清楚，第1次拿球和第2次拿球是相关事件还是独立事情。

1）第1次随机拿一颗，拿到蓝色的概率是多少？可能性是 五分之二

2）但拿掉一颗之后情形便不同了，所以拿第二个的时候：

如果第一次拿的是红的（对应图中的红色箭头），剩下的球里面是2颗篮球，2颗红球。所以第二次拿到蓝球的可能性是四分之二。

如果第一次拿到是蓝的（对应图中的蓝色箭头），剩下的球里面是1颗篮球，3颗红球，所以第二次拿到蓝球的可能性是四分之一。

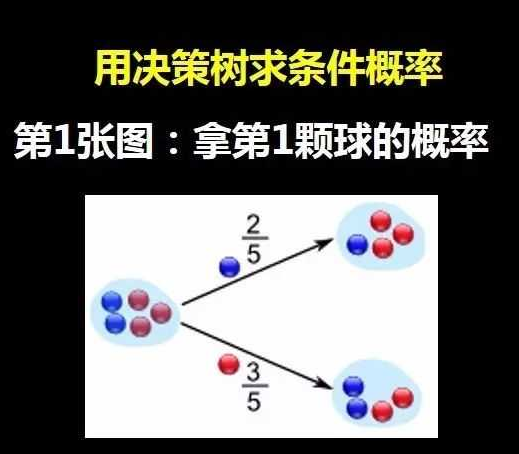
你看，这是一个相关事件，因为第1次拿球的结果，会影响第2次拿球的概率，他们是互相影响的。

**相关事件的概率**也叫“【条件概率】”。【条件概率】是指【事件A】在【另外一个事件B】已经发生条件下的发生概率。

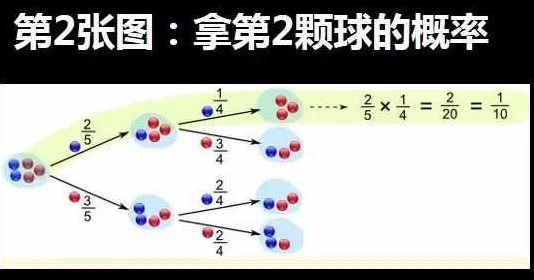
如何用决策树表示条件概率

我们通常用决策树来辅助计算。下图我们用决策树来表示刚才的例子。

我们先看第1张图：拿第1颗球的可能情况：有 2/5 的概率会拿到蓝球，3/5 的概率会拿到红球



我们再来看第2张图：拿第二颗球时的情形。



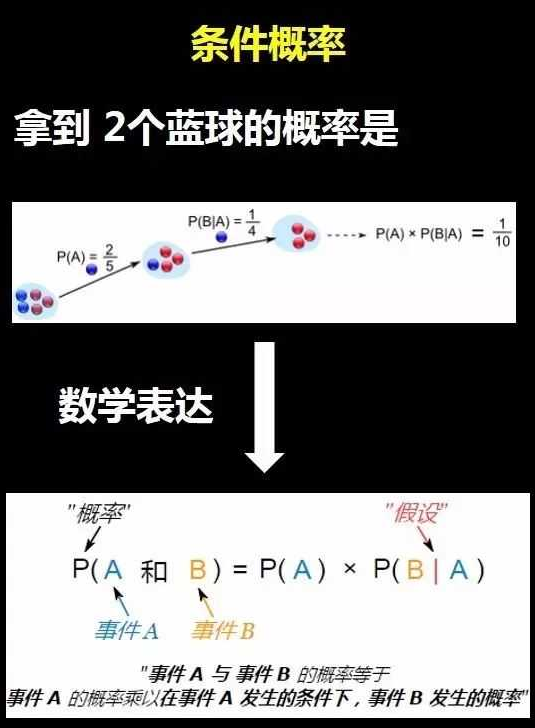
若先拿的是蓝色，第二颗是蓝色的概率是 1/4，第二颗是红色的概率是 3/4。

若先拿的是红色，第二颗是蓝色的概率是 2/4，第二颗是红色的概率是 2/4。

现在我们可以尝试解答像这样的问题了：“拿到2颗蓝球的概率是多少？"

我们把第2张图里第1次拿到篮球的概率2/5，乘以第2次拿到蓝球的概率1/4相乘就可以了。

好了，我们通过【决策树】已经计算出了【条件概率】，下面图片我们进一步看条件概率在数学上的表示就立马明白了。



P(A) 的意思是 “事件 A 的概率”。

在以上的例子，事件 A 是 “第一次拿到蓝球的概率”是 2/5，所以这里P(A) = 2/5

事件 B 是 “第二次拿到蓝求的概率”是1/4，这里用p(B|A)来表示。

这里的竖杆"**|**"来"在【事件 A 】发生的条件下，【事件 B 】发生的概率"。换句话说，事件 A 已经发生了，现在事件 B 发生的可能性是多少。

P(B|A) 也叫在A 发生的情况下 B 发生的 "条件概率"。

【条件概率】是指【事件A】在另外一个【事件B】已经发生条件下的发生概率。【条件概率】表示为：P（A|B），读作“在B条件下A的概率”。若只有两个事件A，B，那么，。

【联合概率】表示两个事件共同发生的概率。A与B的联合概率表示为 P(AB) 或者P(A,B),或者P（A∩B）。

# 贝叶斯公式



公式描述：公式中，事件Bi的概率为P(Bi)，事件Bi已发生条件下事件A的概率为P(A│Bi)，事件A发生条件下事件Bi的概率为P(Bi│A)。

怎么简单理解贝叶斯公式？

<https://www.zhihu.com/question/51448623>

作者：李现民

链接：https://www.zhihu.com/question/51448623/answer/147298455

来源：知乎

著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

【贝叶斯公式】与【两个概率】有关系，一个是【先验概率】（基础概率），一个是【现象概率】（观察到的条件）

例子：某城市发生了一起汽车撞人逃跑事件，该城市只有两种颜色的车，蓝色15%，绿色85%，事发时有一个人在现场看见了，他指证是蓝车。但是根据专家在现场分析,当时那种条件能看正确的可能性是80%。那么,肇事的车是蓝车的概率到底是多少？

【先验概率】：该城市只有两种颜色的车，蓝色15%，绿色85%

【现象概率】：根据专家在现场分析,当时那种条件能看正确的可能性是80%

令B是城市里车为蓝色的事件，G为车子是绿色的事件，E为观察到车子为蓝色的事件。

则由已知条件可以得出P(B)=0.15，P(G)=P(~B)=0.85，至于P(E)我们一会儿再说。

好了，现在，**如果没有证人**看到肇事者车的话，那么我们只能盲猜，因此肇事者的车子为蓝色的概率只能是整个城市里面车为蓝色的概率，也就是先验概率P(B)=0.15，因为这时我们还没有其他证据介入，只能做个粗略的估算。

接下来，当当当当，有证人了。证人说他看到了车子，并且说是蓝色的，注意，这分两种情况，…………重要的事情说两遍：**贝叶斯里面现象(新的证据)部分总是分两种情况出现的**：

一是车子的确是蓝色的，并且证人也正确的分辨出车是蓝色的来了，概率为 P(E,B)=P(B)xP(E|B)=0.15x0.8=0.12，

二是车子根本就是绿色的，只是证人看成蓝色的了，概率为P(E,~B)=P(~B)xP(E|~B)=P(~B)x（1 - P(~E|~B))=0.85x(1-0.8)=0.17，所以P(E)=P(E,B)+P(E,~B)=0.12+0.17=0.29

然后，我们要求解的其实是在有证人的条件下车子为蓝色的概率，也就是P(B|E)=P(E,B)/P(E)=0.12/0.29=0.41

你看，P(B|E)根本就是P(B)的加强版本，【条件概率】跟【先验概率】描述的根本就是**同一件事**。那么当当当当，又一个结论来了：**当有新的证据出现时，P(B|E)会替代原来P(B)的角色**。换句话说，现在警察找到了一个新的证人，他也觉得这辆肇事车是蓝色的，这时在新一轮的贝叶斯概率计算中，基础概率P(B)=0.41，而不是原先的0.15，大家可以算一下，新的P(B|E)=0.73，换句话说，当有两个人看见肇事车辆为蓝色的时候，对比只有一个人看到肇事车辆为蓝色的时候，该车实际为蓝色的概率大大增加。

https://github.com/lixianmin/cloud/blob/master/writer/R/bayes.md

The diachronic interpretation

在很多书中使用字母A、B表示事件，使用P(A|B)表示条件概率，这相对太抽象。我们使用另外一套字母体系：H和E(D)，其中H= hypothesis，E= evidence（或D=data）。这样Bayes的推理过程可以表述为：**通过不断的收集【证据E】来强化对【假设事件H】的信心**。

这种表述方法称为diachronic interpretation，其中diachronic是“随时间变化”的意思。在Bayes理论中，就是指**每当我们收集到一个新的证据之后，都可以加入到原有Bayes系统中用于调整对原有事件的看法（可能是增删改 + - x），因此事件H的概率会不断调整**。



# sigmoid function





















# 第01讲、导数与方向导数

极限包括：数列极限、函数极限

## 1.0、函数极限统一定义

时刻，从此时刻以后，恒有。

其中，表示“存在”，表示“任意的”。

### 极限存在准则

1、夹逼定理

（1）在的某个去心邻域内有

（2）且或者且，那么或存在，且等于。

2、单调有界准则

如果数列满足条件：单调数列

单调增加

单调减少

单调有界数列必有极限。

## 1.1、导数定义

## 1.2、基本求导公式与求导法则









### 1.2.2、函数的和、差、积、商的求导法则

（1）

（2）

（3） 

示例：



### 1.2.3、复合函数的求导法则

设，而，则复合函数的导数为或者。

利用上述公式及法则，【初函数】求导问题可完全解决。

## 1.3、高阶导数与泰勒公式

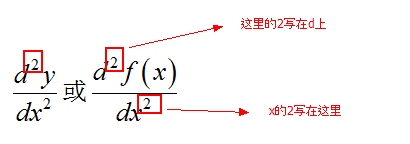
### 1.3.1、高阶导数

**二阶导数**



二阶导数记法：、、或

n阶导数记法：、、或



### 1.3.2、泰勒公式



其中，表示余项

当时，就是【麦克劳林公式】。

### 1.3.3、麦克劳林公式



其中，是高阶无穷小，称为【皮亚诺余项】。

常用函数的麦克劳林公式









### 1.3.4、泰勒公式应用

1、近似计算

2、函数展开

## 1.4、导数的应用

### 1.4.1、判定单调性

如果在内，那么函数在上单调增加；

如果在内，那么函数在上单调减少。

### 1.4.2、求极值、最值

### 1.4.3、凹凸性差别

如果在上连续，在内具有二阶导数，若在内

（1），则在上是凹的

（2），则在上是凸的

## 1.5、多元函数导数

### 1.5.1、偏导数

设函数

如果存在

则称此极限为函数在点处对的偏导数，记为、、或。

对于的偏导数，为 记为、、或。

偏导数的概念可以推广到二元以上的函数。

#### 高阶偏导数

高阶偏导数可以推广到二元以上的函数。

### 1.5.2、与偏导数相关的几个概念

#### 1.5.2.1、梯度

三元函数在点处的梯度为

。

上式是由各个偏导数组成的一个向量。梯度，可以表示为。

举例：，那么，最后；当，则。

梯度的特点：

（1）方向：变化率最大的方向

（2）模：的最大变化率之值。

可以推广到三元以上的函数

#### 1.5.2.2、Hesse矩阵

设，则其Hesse矩阵



#### 1.5.2.3、方向导数

**定义**：若函数在点处沿方向（方向角为）存在下列**极限**：



其中

则称该极限为函数在点处沿方向的【方向导数】，记作。

**定义**：若函数在点处**可微**，则函数在该点**沿任意方向**的方向导数存在，且有



其中， 为的方向角。

### 1.5.3、应用

#### 1.5.3.1、极值

求函数极值的一般步骤：

第一步：解方程组求出实数解，得驻点。

第二步：对于每一个驻点求出二阶导数的值A、B、C。。

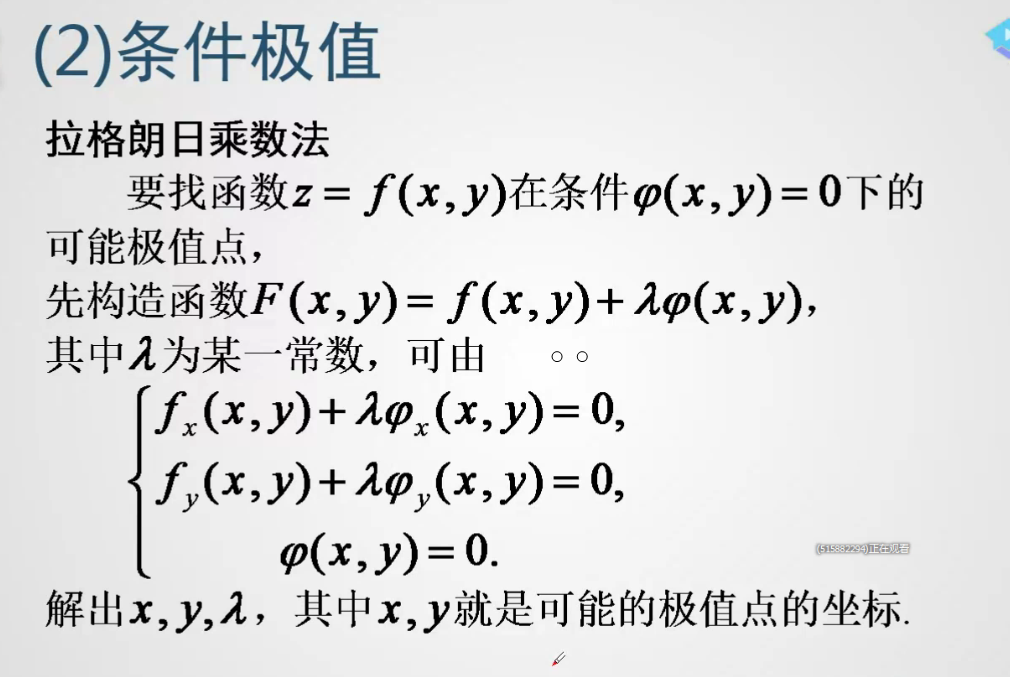
第三步：定出的符号，再判定是否是极值。

#### 1.5.3.2、条件极值

拉格朗日乘数法

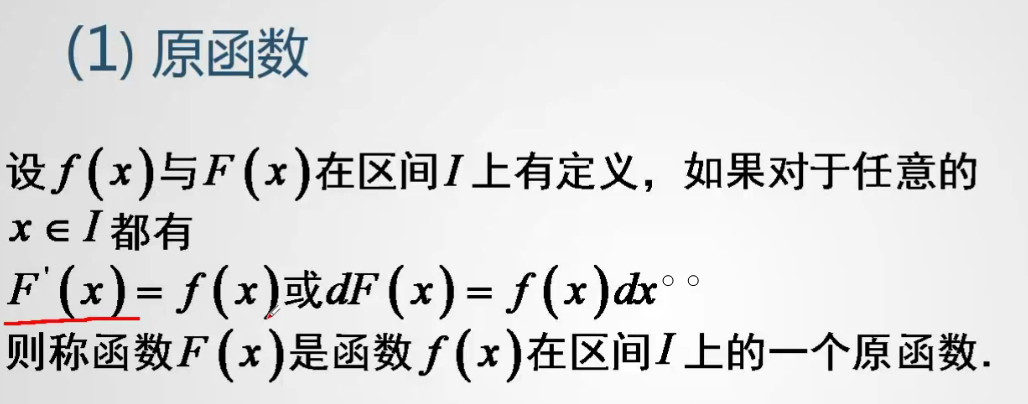
要找函数在条件下的可能极值点，先构造函数

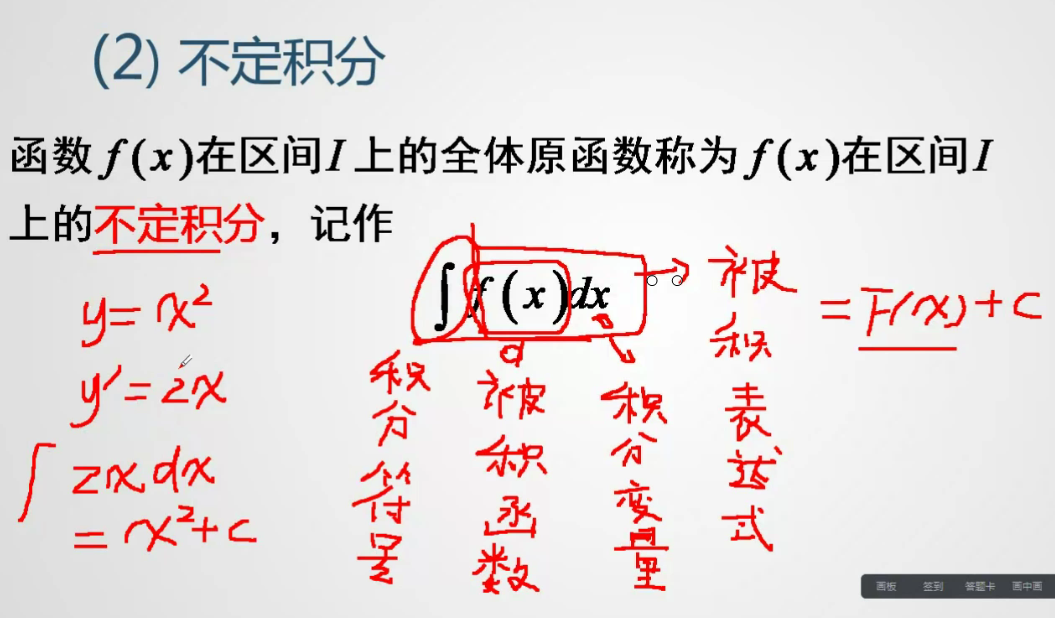
其中，为某一常数，可由

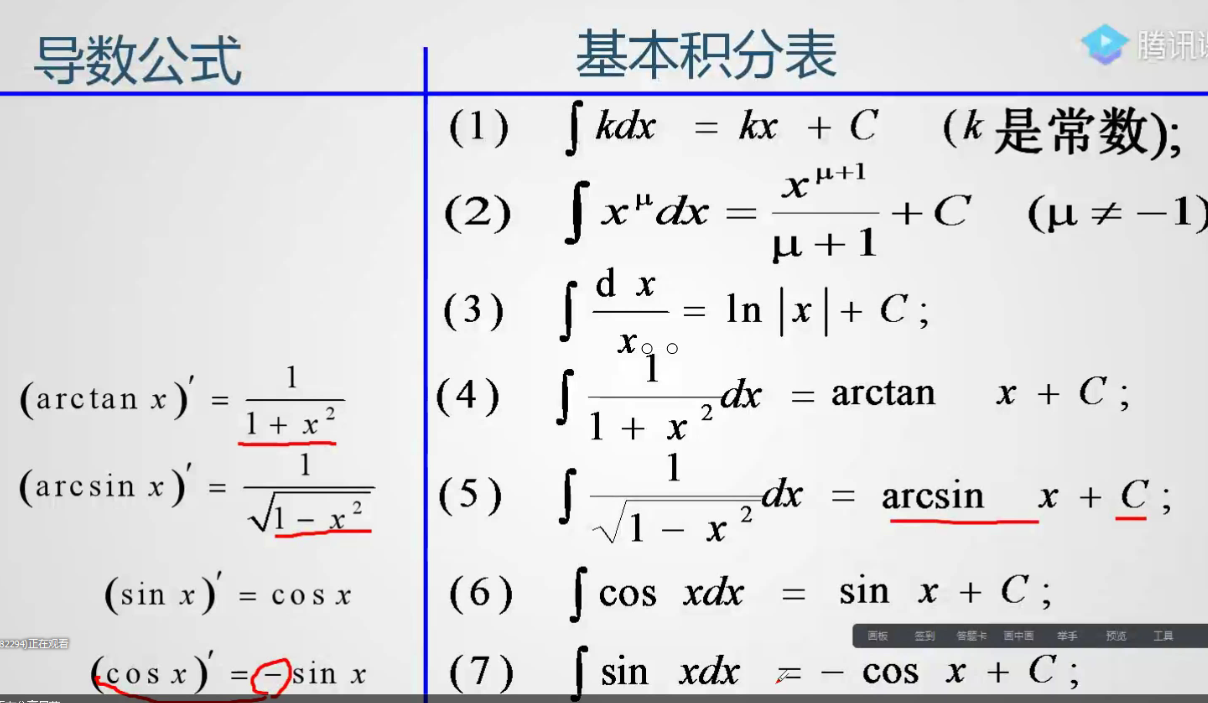


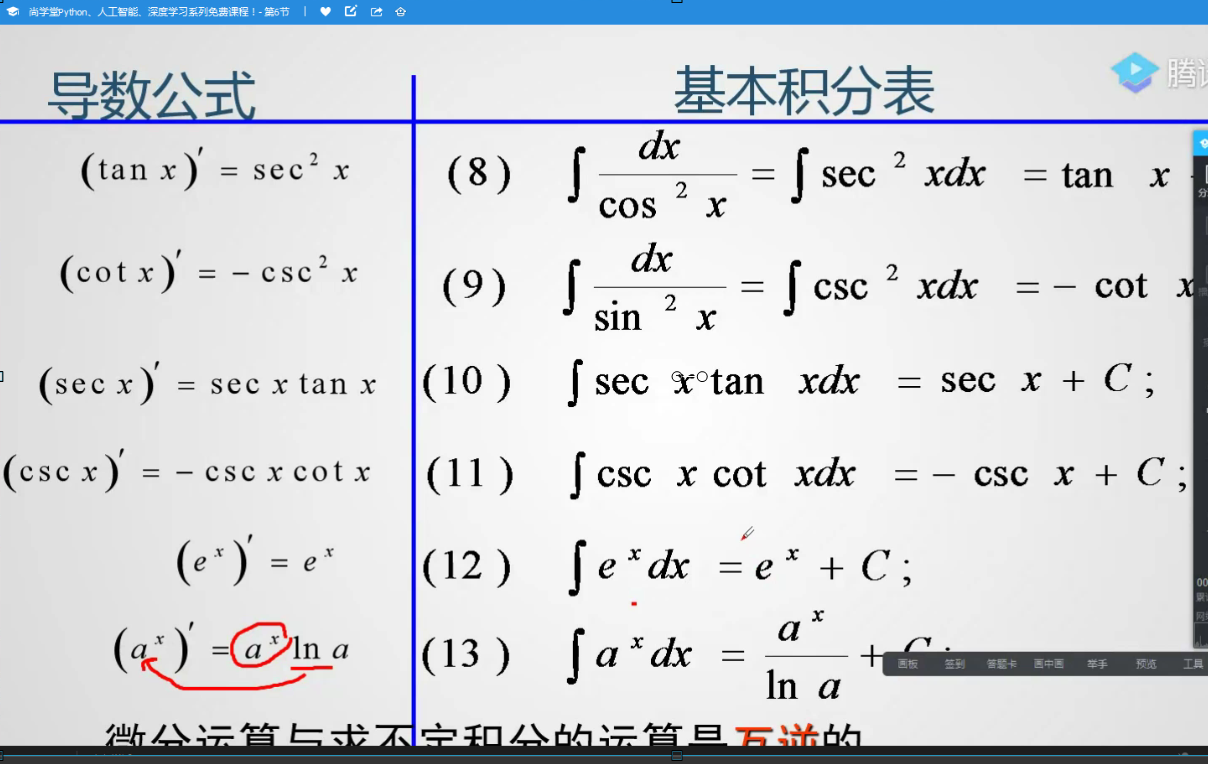
# 第02讲、积分

## 2.1、不定积分

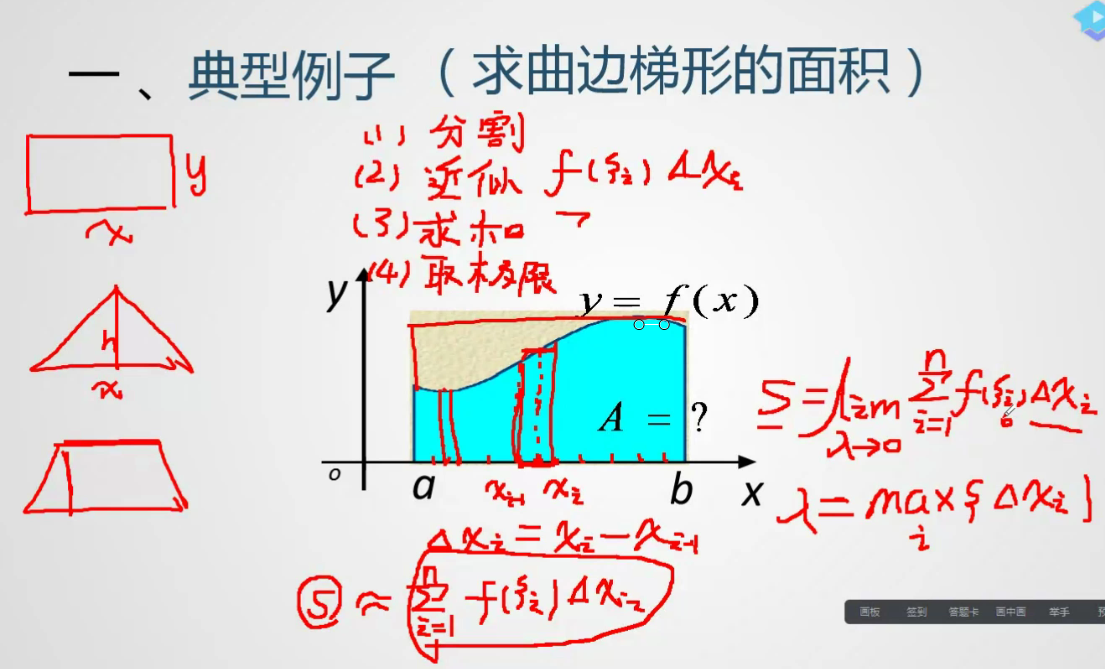


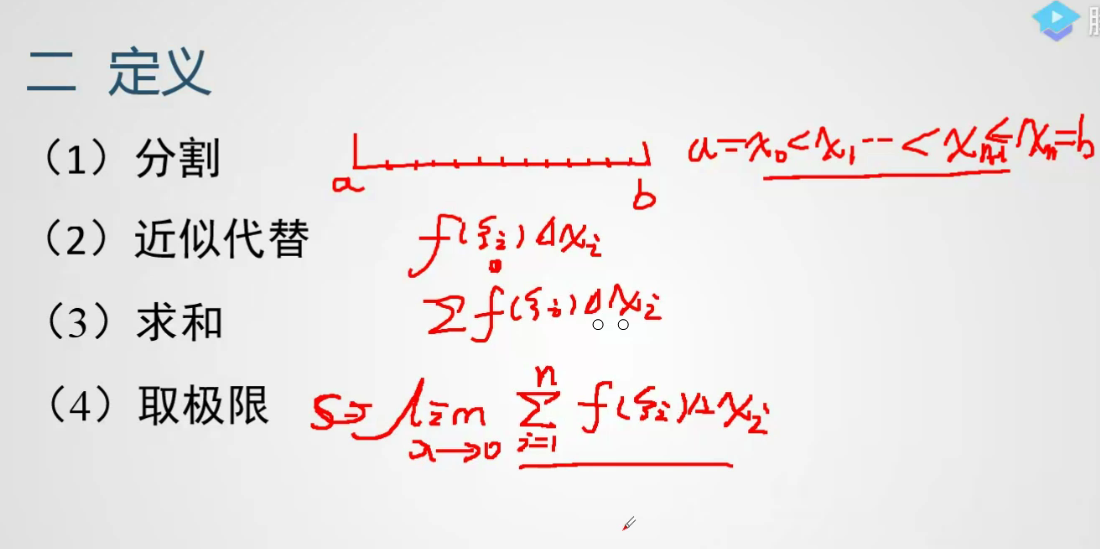




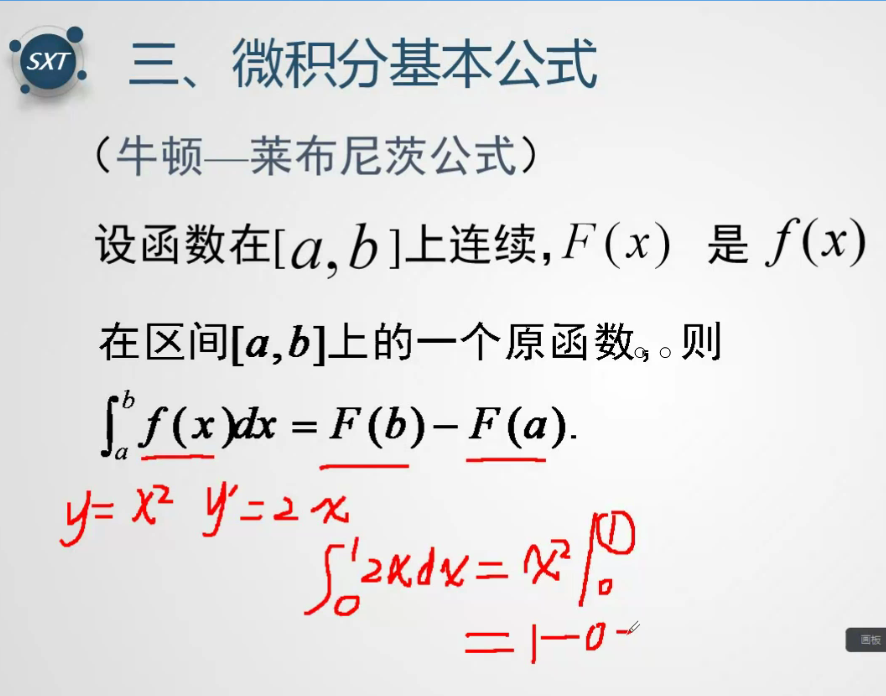


## 2.2、定积分

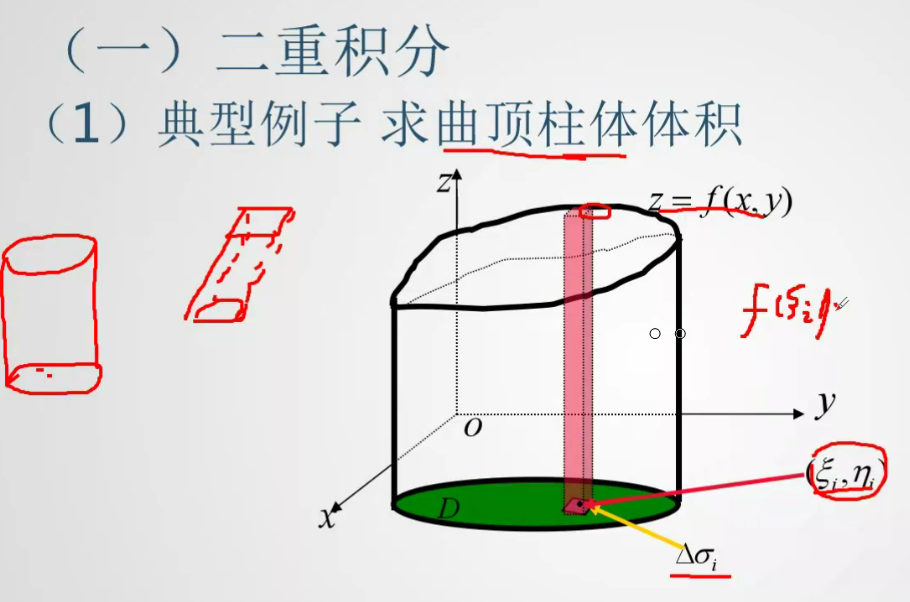


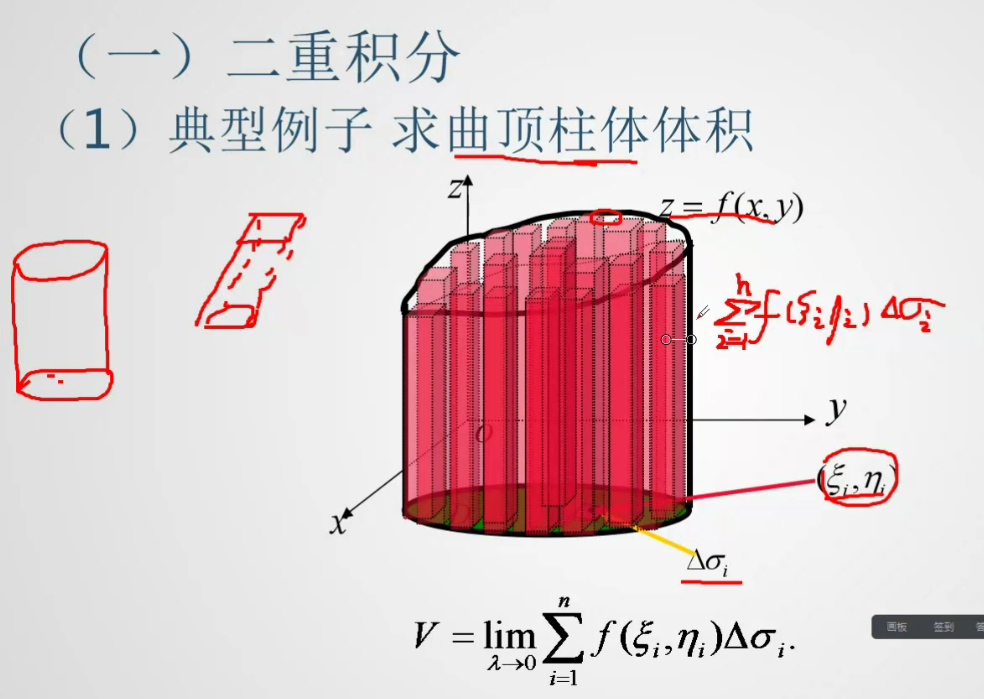


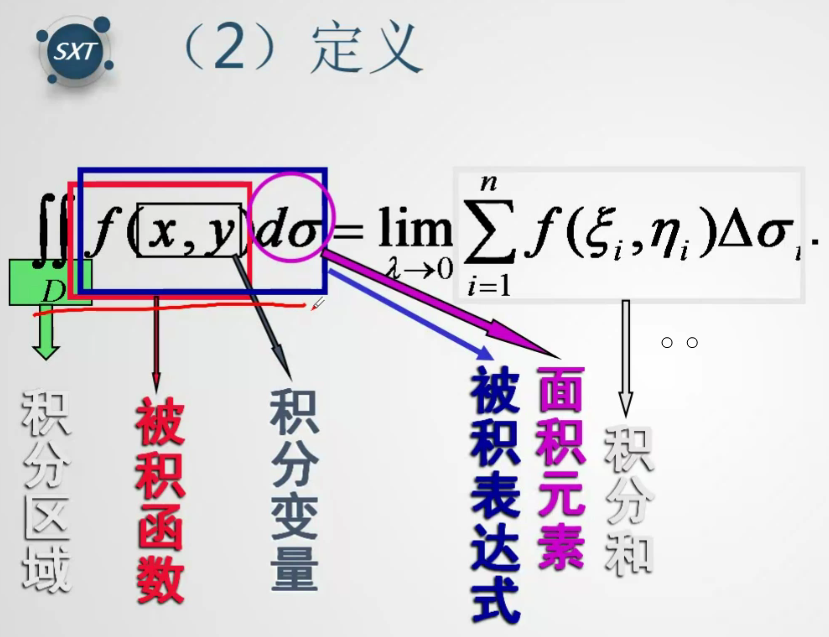


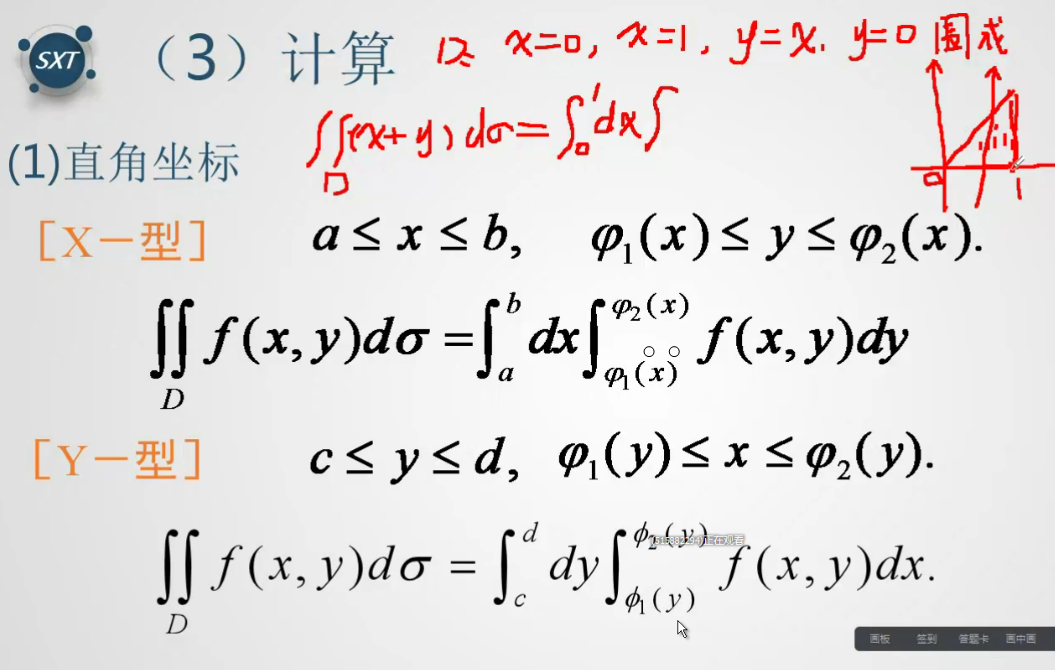


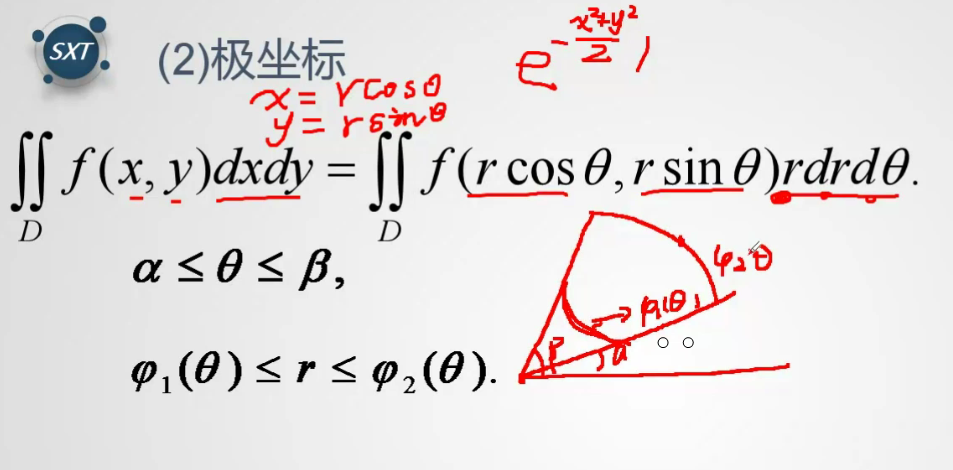
## 2.3、重积分













# 逻辑回归

阈值为0.5



误差函数



Log loss



求偏导



# Soft-max

Softmax是进行多分类的算法。

## 问题描述

现在已知有【K个类别】，Softmax要做的就是【判断一个样本】究竟属于【哪一个具体的类别k中去】。

注意：

（1）【K个类别】是已知的，Softmax是将一个样本分到K类别中某一个类别中去，因此不需要再去求解K。

（2）要将【一个样本】划分到【某一个类别】当中，不是凭空相像的划分，而是依据一定的“量化标准”，这个“量化标准”就是【某一个类别k的分数】。

（3）大K表示所有类别，小k表示某一种类别。

## 求解过程

（1）求得【该样本】在【每一个类别k】中的【分数】，即【量化指标】。

（2）将【每个类别k的分数（或量化指标）】转换成【每个类别k的概率】

（3）挑选【最大概率的类别k】作为【分类的结果】。

### 类别k的分数计算

类别k的分数如下：



### 类别k的概率计算

Softmax函数，计算出来的结果是“概率”，即某一个类别k的概率。



将读为“p head”。

其中，代表第k个类别的概率，这是一个归一化的概率，即。

注意：这里的（sigmoid函数）与**逻辑回归**的（sigmoid函数）不一样。

最终分类的结果：那个类别的概率大，就选择哪个类别。

|  |
| --- |
| Softmax中“Soft”的理解  如果根据“概率”来进行划分，就称为“Soft”；如果根据“分数”来划分，就称为“Hard”。 |

### 预测分类

预测分类的方法：



|  |
| --- |
| argmax是一种函数，函数，的意思就是参数满足为的最大值；换句话说就是 argmax(f(x))是使得取得最大值所对应的变量。arg即argument，此处意为“自变量”。 |

## 交叉熵损失函数

交叉熵损失函数，要使得损失函数最小。



其中，表示某个类别的真实结果，应该是0或1；表示预测属于k类别的概率。K表示一共有K个类别；m表示一共有m个样本，表示求一个平均值。

如果K是2，其实就是逻辑回归



由于是2分类，那么和分别表示“正例”和“负例”，并且。因此上式，可以简写成如下形式：



13:42

求导得梯度



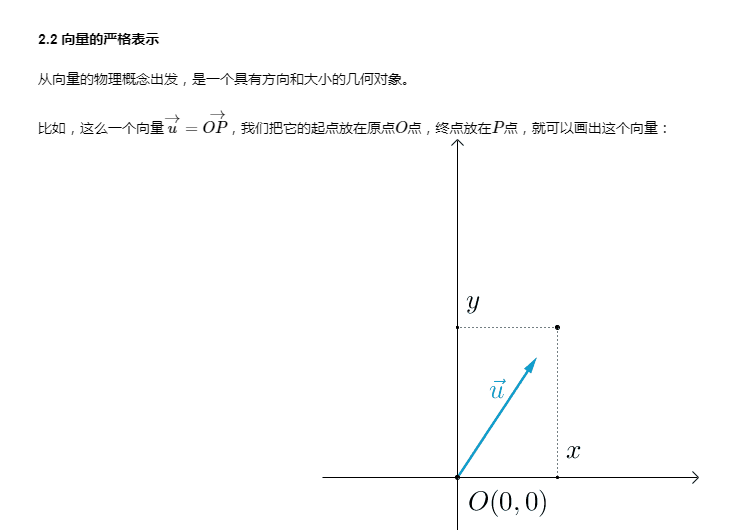
## 逻辑回归与Softmax的区别

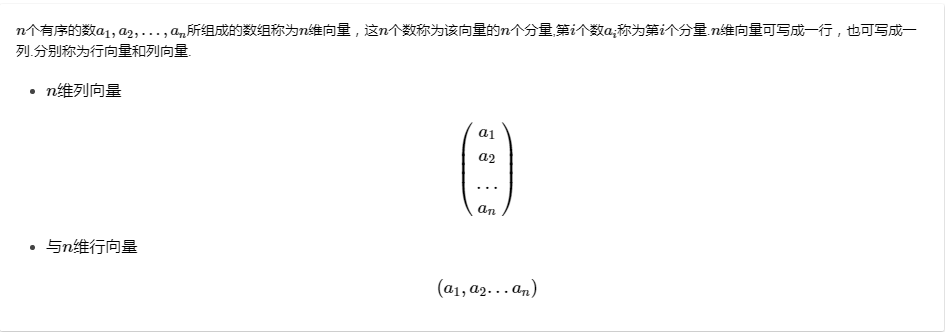
（1）从本质上来，【逻辑回归】是用来做二分类，但是也可以做多分类；是因为可以将一个多分类问题，拆分成多个二分类问题；【Softmax】天生就是用来做多分类的。

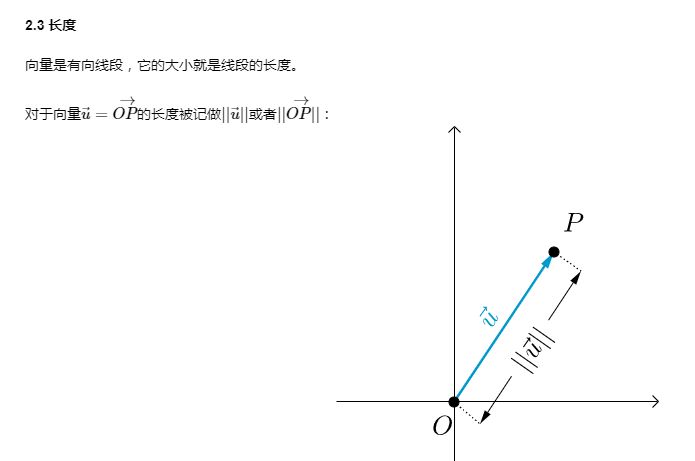
（2）从求解（theta）的角度来说，【逻辑回归】在求解的时候，每个类别的是相对独立的；而【Softmax】在求解的时候，每个类别的是相互依赖的。

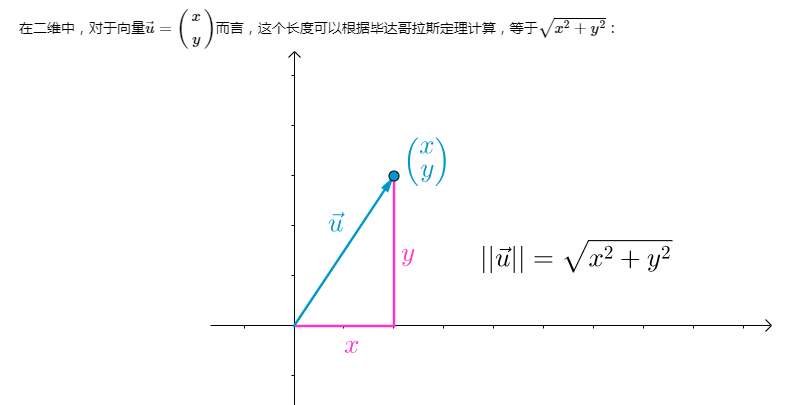
# 马同学线性代数基础

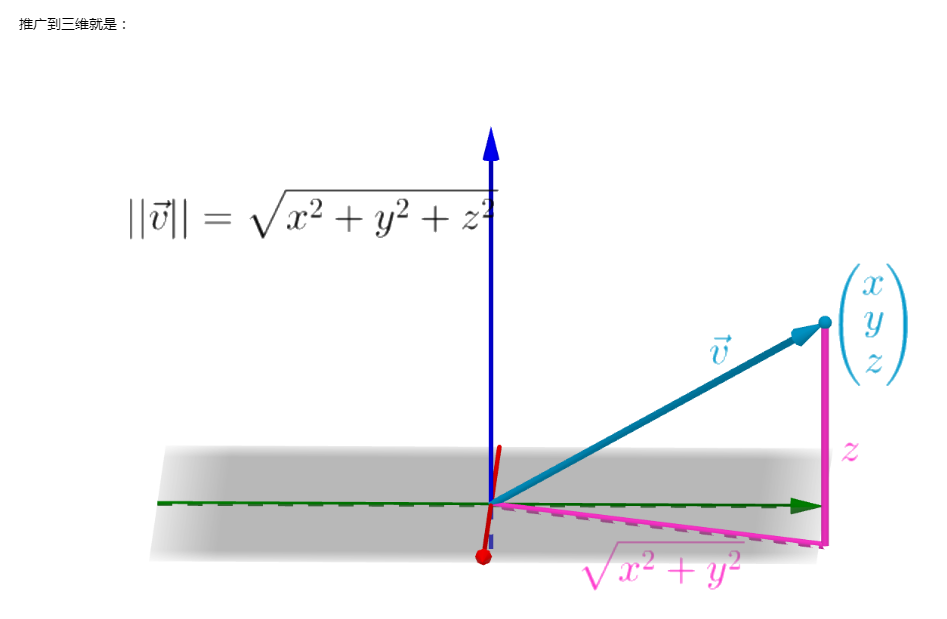
## 01、向量

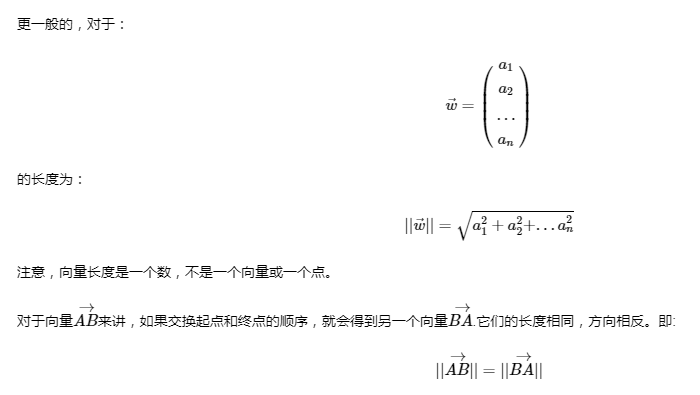




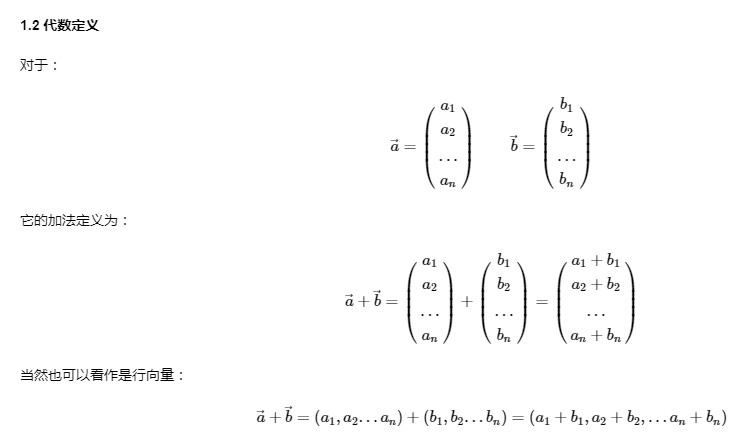


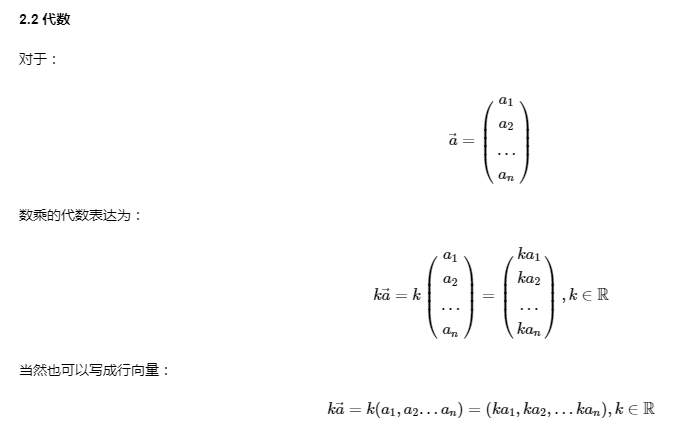


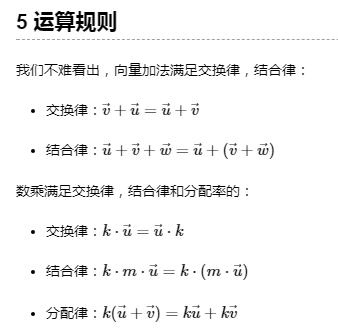




## 02、向量基本操作（数乘和加法）







## 03、线性表示与线性相关

