# 第一章 深度学习数学基础

## 1.1高等数学之微积分

微积分分为微分学和积分学，两者互为逆运算。微分运算常常称为导数运算。

### 1.1.1重识微分

微分（differential）和导数（derivative）都与函数的变化率有关，但它们是两个相关但不完全相同的概念。让我们深入了解这两者的定义和区别。

#### 导数

导数描述了一个函数在某一点上的切线斜率。如果有一个函数，则其在点*x*处的导数通常表示为或。导数的定义是函数在该点的平均变化率的极限，公式如下：



#### 微分

微分描述了函数值的微小变化与自变量的微小变化之间的关系。对于函数，它的微分表示为和，其中是函数值的微小变化，而是自变量的微小变化。微分的定义基于导数，可以表示为：



所以，导数和微分都与函数的变化率有关，但它们的重点略有不同。导数关注的是函数在某点的切线斜率，而微分关注的是函数值的微小变化与自变量的微小变化之间的关系。简言之，导数是一个比率或斜率的概念，而微分描述了当自变量发生微小变化时，因变量如何变化。

下面笔者以导数为例子分别从几何、物理和代数角度解释导数的含义。

导数的几何解释是：该函数曲线在这一点上的切线斜率。

导数的物理解释是: 导数物理意义随不同物理量而不同,但都是该量的变化的快慢函数，即变化率。

导数的代数解释是：更精细的除法运算。

前两个解释的角度相信读者已经很熟悉了，那么怎么理解导数的代数是更精细的除法运算这一说法呢？

举一个物理例子：距离*s* = 25m 时间*t* = 5s 求平均速度*v*？

这个问题很好回答，正常的除法即可轻松处理（）。但是，如果速度不是均速，而且希望求得第5秒时的瞬时速度，怎么办？



如上解法，是一个很多的时段，用时刻走过的路程减去第五秒时走过的路程，再除以时段，解得的就是第五秒时的顺时速度。当无穷小时，就是导数的概念了，即​。

可以看出来导数是即时的变化率，放在路程和时间这个物理场景下，顺时速度就是路程的即时变化率。其求解的方法就是一个简单的除法而已！本质上还是除法运算。

### 1.1.2微分的解读

回忆一下微分的数学表达式：



导数含义的解读：

（1）导数揭示了函数在某点的切线斜率。

（2）导数揭示了函数在某点的变动规律。

在这里笔者更推崇第二种解读方法。其实可以把**乘到等号右边去会更形象，即：



举个例子来解读什么叫函数在某点的变动规律。假设，求处的导数。



即，变量变动一点点，将引起函数值相对于变量十倍的变化。这点很重要。

我们可以根据这个解读来推导一下微分的乘法法则和幂法则。也是举一个例子，假设函数，先回忆一下微分的乘法法则：



下面我们来推导一下这个法则怎么来的。首先，把函数放在求解矩形面积这个例子中，即是矩形面积、是宽、是高，此时当变量变动一点点时，根据微分的解读，其意义是矩形面积的变动率，如图1-1所示。

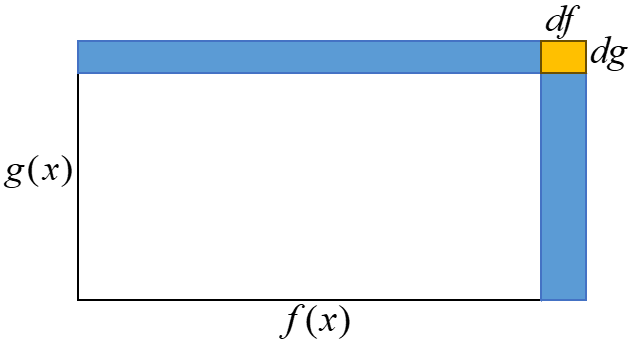


图1-1 微分推导

其中为面积的变动，即图1-1中蓝色和橙色的部分： ，由于是二阶无穷小，约等于0，可以约掉；再在等号左右分别除去就得到了微分的乘法法则。此时，为面积的变动，而为面积的变动率。

下面来推导一下这个导数的幂法则怎么来的，先来回忆一下幂法则：



还用刚刚的例子，如果此时和都等于，那么示例变成如图1-2所示。

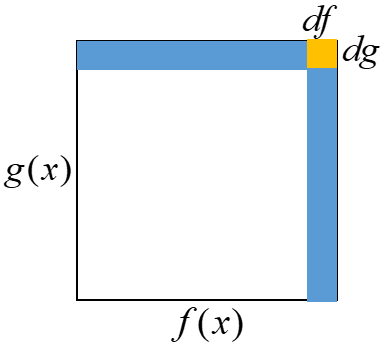


图1-2 导数幂法则推导

此时的公式推导如下：



同样的方法推广到三维空间，乘法法则和幂法则的推导也是适用的，如图1-3所示：

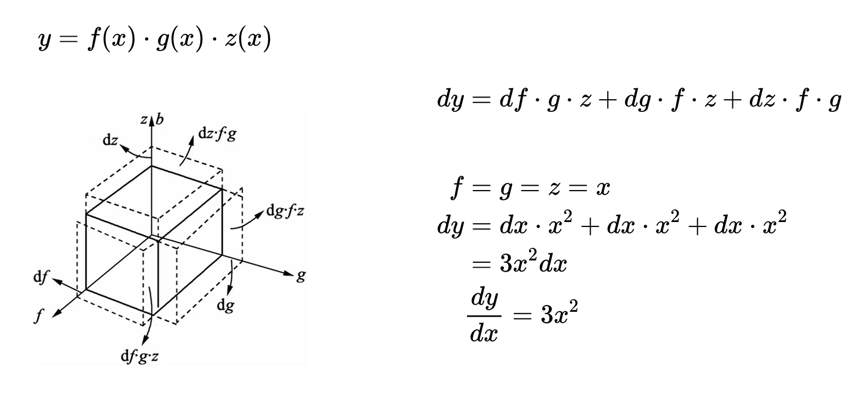
[](http://124.220.164.99:8090/upload/2022/11/image-1669327221374.png)

图1-3 高维空间的导数乘法法则和幂法则推导

此时的体积计算公式为，体积的变动为：



如果此时、 和都等于，那么体积的变动为：



体积的变动率为：

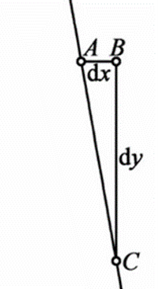
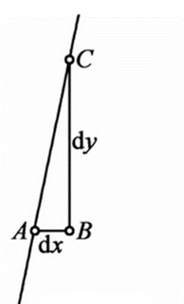


可以想象，继续推广到高维空间也是适用的，这里就不方便做可视化了。

### 1.1.3微分与函数的单调性和凹凸性

直接上定义：在内可导，如果，那么函数在内单调递增；如果, 那么函数在(a, b)内单调递减。

用微分的定义（微分解释了函数变动的规律）也容易解释单调性，如图1-4所示。

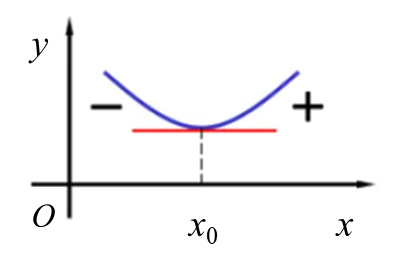
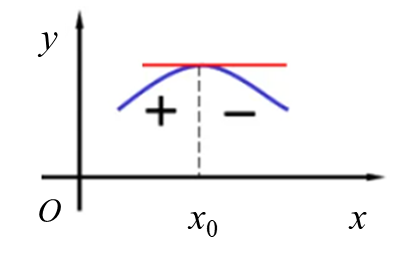


(a)的情况(b)的情况

图1-4 函数的单调性

当*x*的变动引起的函数变动是正增长时，函数单调递增。当*x*的变动引起的函数变动是负增长时，函数单调递减。

接下来介绍极值，如果函数在点处可导，且其导数等于0；当在的左领域，右邻域时,为的极大值；当在的左领域，右邻域时, 为的极小值，如图1-5所示。



(a)极大值的情况 (b)极小值的情况

图1-5 函数的极值

最后，理解一下函数的凹凸性，这与函数的二阶导数相关。如果函数**在内连续，且二阶可导，当在内，**，则函数为凹函数；当在内，**，则函数为凸函数，如图1-6所示。

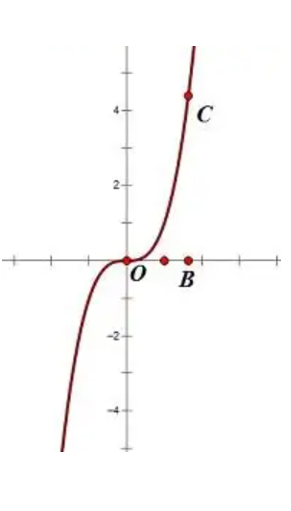
[](http://124.220.164.99:8090/upload/2022/11/image-1669332143686.png)

图1-6 函数*x*的三次方

图1-6中的函数在第一象限是凹函数，其一阶导始终大于0，说明函数单调递增；其二阶导也大于0，说明这个单调递增的速度越来越快，此时说明这个函数是一个凹函数。

同理，函数在第三象限是凸函数，其一阶导始终小于0，说明函数单调递减；其二阶导也小于0，说明这个单调递减的速度越来越快，此时说明这个函数是一个凸函数。

### 1.1.4微分的链式法则

在机器学习中，尤其是在深度学习和神经网络中，链式法则用于计算复合函数的导数，这在反向传播算法中尤为关键。具体来说，当我们训练一个深度神经网络时，我们需要计算损失函数相对于每个权重的梯度。由于神经网络的每一层都是复合的，链式法则使我们能够从输出层逐步回到输入层，计算这些梯度。

先来看一下微分链式法则的数学公式：



理解起来很简单，就像剥洋葱一样，我一层一层拨开你的心~链式法则一般用于复合函数的求导，先对外层函数求导，再乘上内层函数的导数。之前一直强调导数是函数的变动规律，那么链式法则就是变动的传导法则。举例子如下：



对求导的步骤如下：



解释一下， 的变动会引起的变动，进而引起的变动。链式法则就是变动的传导法则。

那么链式法则有什么用处呢？常见的作用有两个，一个是帮助推导一些常见初等函数的导数，具体见表1-1。

表1-1 常见初等函数的导数

|  |  |
| --- | --- |
| 函数类型 | 导数推导 |
| 常函数 | 若，则 |
| 幂函数 | 若，则 |
| 三角函数 | 若，则 |
| 若，则 |
| 指数函数 | 若，则 |
| 若，则 |
| 对数函数 | 若，则 |
| 若，则 |

第二个常见用处是可以对隐函数进行求导。先来解释什么叫做隐函数：形如的函数叫隐函数，将自变量和因变量放在同一个式子中，隐藏了二者之间的函数关系，因此称之为隐函数。

那么什么叫显函数呢？对应隐函数概念，显函数可以理解为自变量和因变量的函数关系明显的函数，形如。

对比来看是不是很清晰！最后，什么叫隐函数显示化？将隐函数变形成显函数的过程称为隐函数的显示化。圆的方程就是典型的隐函数，已知圆的方程为，点M的坐标为(4,3)，如图1-7所示。

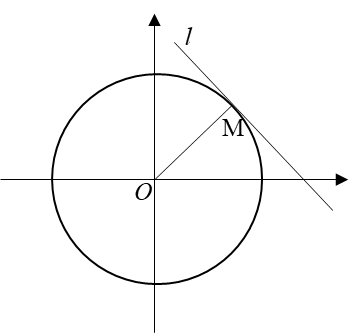


图1-7 求圆在(4,3)处的导数

求M点的切线，则，因为，所以。将圆的方程转化到的形式即称为隐函数的显示化，然后针对这个函数求导，再代入求出结果即可。

实际上，可以不进行隐函数的显示化，使用链式法则来直接求导。过程如下所示：



最后再带入求出结果即可。

### 1.1.5偏微分与全微分

在机器学习中，许多函数都是多变量的。我们需要知道每个输入变量的变化如何影响输出。偏微分正是用于这个目的的。例如，在线性回归中，我们可能要最小化多变量函数（即损失函数）。偏微分告诉我们每个权重的变化如何影响总体误差。

偏微分适用于自变量有两个及两个以上的函数，本质上是一种降维后的运算。举个例子：

求解在处，关于的偏导，求法是先将带入函数后进行关于的求导，再将带入求得最后的结果，如下所示。



函数 由于有两个自变量，所以其函数存在于三维空间，如图1-8所示。

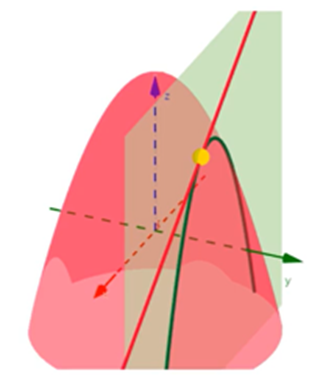


图1-8 *f*(*x, y*)关于*x*的偏微分

求解关于的偏微分时，固定住， 相当于图中的矩形平面，然后再关于进行求导，本质上是一种降维后的求导运算。

至于全微分，则是偏微分的求和，公式如下所示：



其含义可类比于微分的含义，即自变量的微小变动引起的函数变动等于微小变动引起的变化加上微小变动引起的变化。

### 1.1.6梯度与方向导数

梯度是机器学习中的核心概念，尤其是在优化中。它为我们提供了一个方向，告诉我们如何调整参数以最小化损失函数。尤其在梯度下降算法中，我们使用梯度的负方向来更新模型的权重，以逐步减少误差。

梯度是一个向量，它的各个分量是多变量函数的各个偏微分。对于函数，其梯度在点处定义为：



如果函数是三维空间中的一个标量场，例如，那么其梯度是：



梯度的方向是函数增长最快的方向，其大小（或长度）表示该方向上的变化率。

在介绍偏微分时，我们可以求或方向上的导数。那么，可不可以求任意方向上的微分呢？这就是方向导数，方向导数考虑的是函数在某一指定方向上的导数。这个方向由一个单位向量**给出。对于函数**，在方向**上的方向导数定义为：



其中，是方向向量，是函数在点处的梯度。实际上，方向导数意味着即时变化量，梯度意味着函数增长最快的方向，方向向量意味着当前选择的求导方向。接下来逐一解释这些概念。

方向导数意味着即时变化量：方向导数描述了多变量函数在给定方向上的变化率。你可以将其视为在一个特定方向上，当你从一个点移动一个微小的距离时，函数值的“即时”变化。例如，考虑地形图：方向导数可以告诉你，如果你朝某个特定方向走，地形的斜率或高度变化是多少。

梯度意味着函数增长最快的方向：梯度是一个向量，它的方向指向函数增长最快的方向，而其大小表示该方向上的变化率。继续使用地形图的例子：梯度会指向你应该走的方向，以便最快地上山。它的长度（或大小）告诉你这个方向上的斜率或坡度有多陡峭。

方向向量意味着当前选择的求导方向：方向向量是一个单位向量，它定义了你想要计算方向导数的方向。在地形图上，这就像你决定朝哪个方向走，以确定那个方向上的斜率。这个方向不一定是最陡峭的方向（由梯度给出），但它是你选择的方向。

综上所述，当我们考虑一个多变量函数的方向导数时，我们实际上是在问：“如果我朝这个特定方向（由方向向量给出）移动，我的函数值会怎么变化？” 梯度为我们提供了一个参考，告诉我们哪个方向上函数的变化最大。

### 1.1.7泰勒公式与麦克劳林公式

泰勒公式允许用多项式来近似复杂的函数，这在算法中有时用于简化计算。 例如，在高斯过程回归和一些其他贝叶斯方法中，泰勒展开用于线性化关于后验的计算。

泰勒公式的本质是用简单的多项式来近似拟合复杂的函数，我们一点一点来引出泰勒公式。

先回忆一下微分：



若存在，在附近有，令，将带入式(1-20)整理得到：



这就是泰勒公式思想的起源，即函数可能是一个很复杂的函数，甚至复杂到写不出函数公式，但是我们可以用该函数中某点的函数值和导数进行近似，那么这种近似是否准确呢？如图1-9所示。

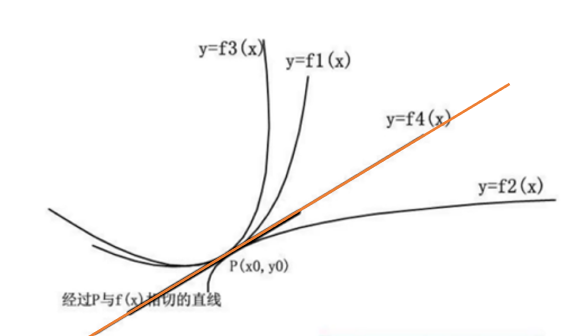
[](http://124.220.164.99:8090/upload/2022/12/image-1669928570591.png)

图1-9 一阶导

从图1-9中看出来一阶导在函数值附近的拟合效果还不错，但是在远离函数值的拟合效果很差，而且很多函数都可以拟合这个一阶导，这说明一阶导的拟合并不准确，那如果再加上二阶导呢？

一阶导决定函数增减性，二阶导决定函数凹凸性，当一阶导和二阶导同正时，函数只能被限制在图1-10的阴影部分，明显可以看出这个约束比只有一阶导时强很多。

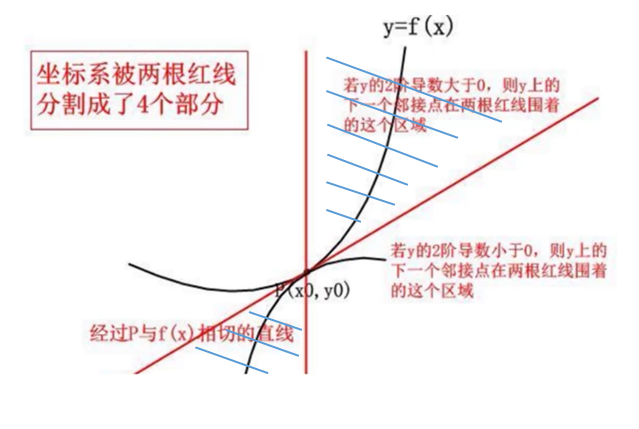
[](http://124.220.164.99:8090/upload/2022/12/image-1669929098471.png)

图1-10 一阶导&二阶导

推导下去，导数阶数越多对函数的约束能力越强，越能拟合出一个确定的函数。所以泰勒公式可以写成：



刚刚介绍了导数阶数，下面想想，这个多项式有什么用？举个例子：

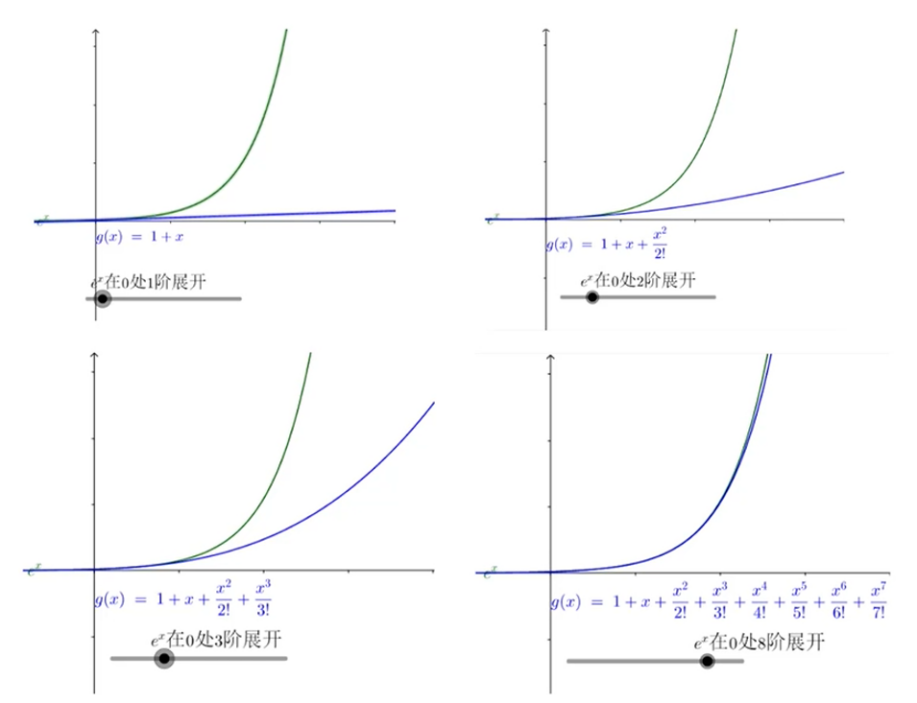
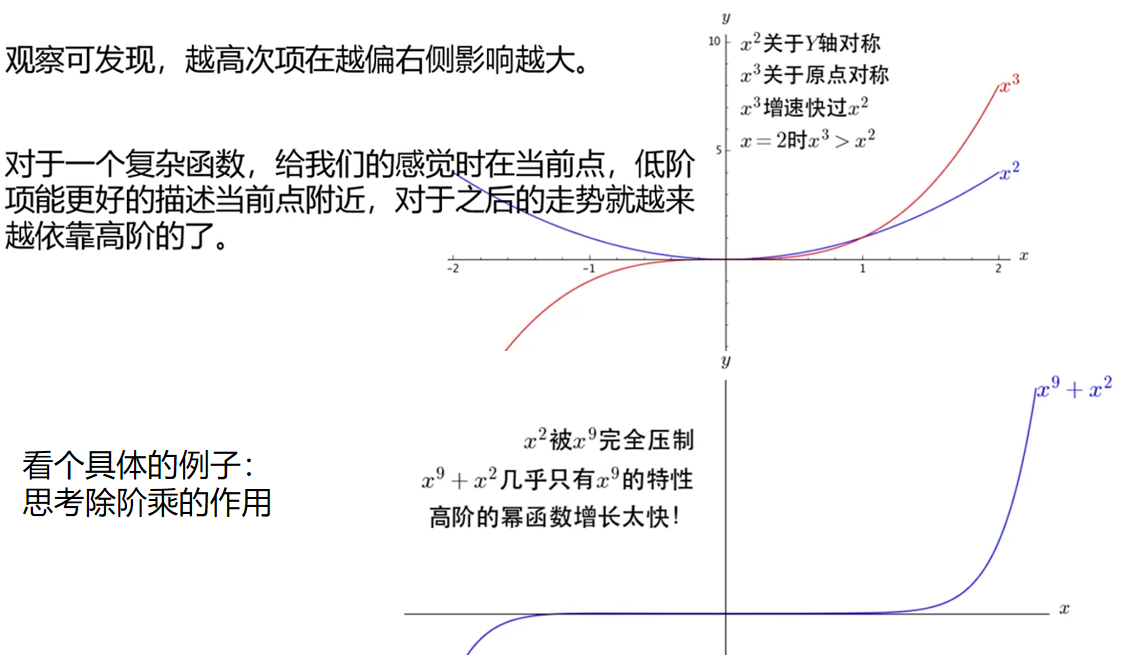
[](http://124.220.164.99:8090/upload/2022/12/image-1669929467911.png)

图1-11 多项式

我们可以明显看出多项式的项数越多，在某点远处的拟合效果越好，这就是为什么泰勒公式中有多项式的原因。最后解释一些阶乘的作用，如图1-12所示。

观察可发现，越高次项越偏右侧影响越大。对于一个复杂函数，

[](http://124.220.164.99:8090/upload/2022/12/image-1669929787398.png)

至于麦克劳林公式则是泰勒公式的简单版，即时的泰勒公式：



最后，再提一下泰勒公式的本质：当一个复杂函数太复杂不可求时，可以用该函数某点的值和该点的多阶导数进行拟合。

### 1.1.8拉格朗日乘子法

拉格朗日乘子法是一种优雅的方法，用于解决在约束条件下寻找函数的极值点的问题。例如，著名的机器学习算法支持向量机（SVM）就使用拉格朗日乘子法来最大化两个类别之间的边界，同时满足某些约束。

让我们更详细地讨论这个方法的步骤和原理。

考虑一个要最大化或最小化的目标函数，其中是一个向量。同时，有一个约束。

当约束满足时，目标函数的梯度和约束的梯度必须是平行的。这是因为在任何其他方向上，都可以在不违法约束的情况下进一步增加或减少。这意味着存在一个常数，使得：



其中即为拉格朗日乘子。

步骤：

（1）构建拉格朗日函数：将目标函数和约束结合成一个新的函数。



（2）计算偏导数：对于每一个变量和拉格朗日乘子，求出拉格朗日函数的偏导数。



（3）解方程组：上述方程组给出了个方程（其中*n*）是变量的数量，解这个方程组可以得到和的值。

（4）检验解：确定每个解是最大值、最小值还是鞍点。

拉格朗日乘子法的核心思想是：在约束表面上，函数只有在其梯度与约束的梯度平行时才能达到极值。这是因为，如果的梯度不是约束的梯度的倍数，那么总有一种方法可以沿着约束表面稍微移动以改善的值。

### 1.1.9重识积分

积分的几何解释是：该函数曲线下的面积。

积分的物理解释是: 积分的物理意义随不同物理量而不同，比如对力在时间上积分就是某段时间内，力的冲量；如果是对力在空间上的积分，就是某段位移里，力做的功。

积分的代数解释是：更精细的乘法运算。

这里如何理解积分是更精细的乘法运算？我们还是放到路程速度时间这个物理系统中举个例子：假设现在速度*v* = 5m/s、时间*t* = 10s，行驶距离*s*怎么求？ 很简单正常的乘法即可处理（)。但是有个前提，即速度是恒定的。那如果是变速的呢？ 这就是积分的内容了。

假设现在的速度，求第10时刻行驶过的路程？积分的思想是将时间段尽可能的切分成小段，以每一小段起始时刻的顺时速度作为这一小段时间内的平均速度，最后把这些小段时间内各自形式的路程加起来就是10秒内行驶的总路程，具体解法如下：

1s时，小汽车的速度；2s时，小汽车的速度；3s时，小汽车的速度；依次算下去，10s时，小汽车的速度。再用乘法运算计算出每一小段时间（）的距离，即：



再将这些距离加起来，引入一个加和符号，即：



公式(1-30)中将10秒钟以一秒为间隔切分成10个小段时间，最后求得的路程是110m ，可视化图片如图1-13所示。可以想象，当时间间隔足够小时，这个速度函数图像近似三角形，此时的路程就是这个三角形的面积等于100m，刚刚近似求得的110m离这个正确答案已经非常接近了。

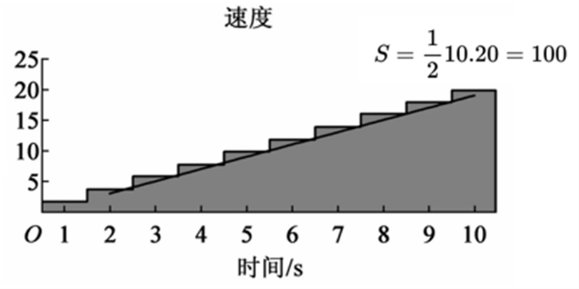


图1-13 积分近似求解路程

所以积分的本质就是更精细的乘法，最后求和就可以得到输出结果了。我们把整除的加减乘除归为初等数学，把微积分这种精细的乘除运算归为高等数学。

### 1.1.10不定积分和反导数

不定积分在机器学习中主要用于计算函数的原函数，尤其是在概率密度函数和累积分布函数之间的转换中。例如，在概率论和统计中，累积分布函数 (CDF) 是概率密度函数 (PDF) 的不定积分。对于某些模型，如概率图模型，我们可能需要在PDF和CDF之间进行转换。

不定积分和反导数是一个概念的两种不同说法，其本质就是求导过程的逆过程。“求导过程的逆过程”指的是，如果你先对函数进行微分得到其导数，然后再进行不定积分，你将得到原始函数（可能加上一个常数）。

不定积分表示的是一个函数的原函数集合，给定一个函数，其不定积分表示为：。结果是一个函数（或函数族），而不是一个具体的数值。常常在这个结果中加入一个常数项C，因为对结果进行微分后，任何常数都会消失。

反导数是描述导数的拟操作的另一种说法。如果函数是函数的反导数，那么。

其实可以从反函数的角度来理解反导数，例如指数函数和对数函数互为反函数，举个具体的例子：

已知一种病毒每分钟感染人数增加一倍，想知道六分钟能感染多少人，可以写成指数函数进行求解，即  。如果现在已知感染了64个人，想知道需要几分钟，那么则写成对数函数进行求解，即。从这个例子可以看出，所谓的 “反” 就是将已知条件和求解目标进行对换而已。

上面的类比其实并不严谨。反导数实际上是一个集合，它是一族函数。因为常数的求导结果为0，所以不定积分的求解结果上还要加一个常数项C，这个C可以等于任何常数，即：



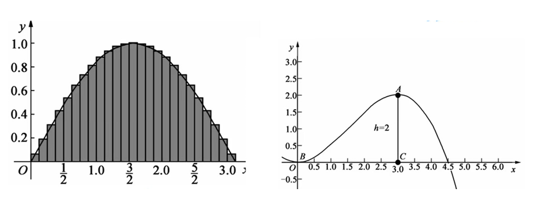
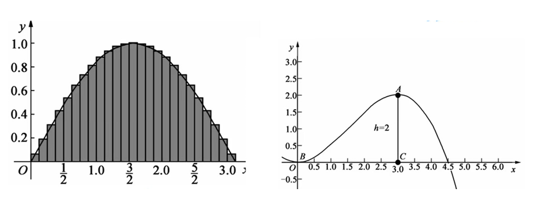
### 1.1.11定积分与牛顿-莱布尼茨公式

牛顿-莱布尼茨公式提供了一种计算定积分的方法，即通过求取两个不定积分的差值。在机器学习中，这常常用于计算概率或期望值。例如，在贝叶斯机器学习中，我们经常需要计算概率分布的期望值或方差。使用牛顿-莱布尼茨公式，我们可以通过求解不定积分来得到这些值。

牛顿-莱布尼茨公式揭示的是某函数在一个区间内的积分与该函数的反导数的关系，公式表示为：



举个例子，已知速度，求到的路程。求解方法有两种，一种是按照积分的定义，将这个时间段尽可能的细分，如图1-14(a)所示，将求解曲线下面积的问题转化成求解这些矩阵面积之和的问题。另一种方法就是利用牛顿莱布尼茨公式进行求解，如图1-14(b)所示。

[](http://124.220.164.99:8090/upload/2022/12/image.png)

(a)积分方法 (b) 牛顿莱布尼茨公式法

图1-14 路程求解

对速度求不定积分：



### 1.1.12微积分的基本定理

微积分不仅研究一个函数更深刻的性质（即更精细的乘除法），还研究不同函数之间的关系。举一个圆的例子，如果已知圆的周长，怎么求面积？如图1-15所示。



图1-15 积分近似求解圆面积

图1-15中，当知道周长求面积时就用到了积分的思想：即精细的乘法后求和。具体来说：将圆以同心圆的方式切分并展平，每一个切出来的环就类似于一个长方形。圆的面积就等于这些长方形面积的总和。当把这些长方形按照从低到高的顺序排列后，且切出来长方形的数量无限多时，这些长方形组成的形状类似于三角形，此时面积的求解就等于三角形的面积，即：圆的周长公式为，圆的面积公式为；化成三角形后边长分别为半径和周长，则面积为。

那么，当已知面积时，能不能求周长呢？当然可以，这其中就用到了求导的思想：即精细的除法，如图1-16所示。

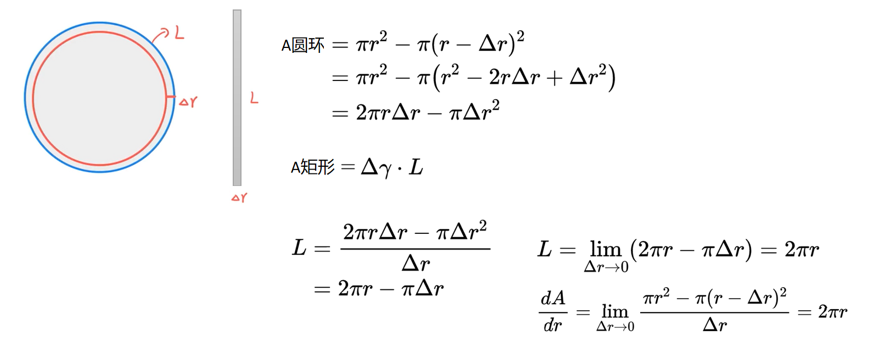
[](http://124.220.164.99:8090/upload/2022/11/image-1669322525501.png)

图1-16 微分近似求解圆的周长

圆环的面积为：



矩形的面积为：



圆环的面积约等于矩形的面积，即。约分后周长*L*等于。利用微分的思想，当无穷小时，周长 。

对圆环来说，当无穷小时，。

总结一下：

圆的面积公式: ；圆的周长公式: 。求微分等于；求积分等于。根据这个例子就可以引出微积分的基本定理了。

微积分的基本定理可以分为两部分。第一部分：如果是在闭区间上的一个原函数，那么，这意味着给定一个连续函数的导数，可以计算该函数在某个区间上的总变化。第二部分：如果在闭区间上是连续的，并且是在上的一个原函数，那么，这意味着给定一个函数，可以找到它的导数，并可以得到函数在每一点上的瞬时变化率。

考虑到圆的例子，这两部分都得到了很好的演示。当从面积函数到周长函数时，实际上是在应用微积分的基本定理的第二部分；当从周长函数到面积函数进行积分时，是在应用基本定理的第一部分。

在实际应用中，微积分的基本定理为我们提供了一种强大的工具，使我们能够在积分和微分之间进行转换。这在物理学、工程学、经济学等领域都是非常有价值的。例如：在物理学中，速度与位移之间的关系、电流与电荷之间的关系都可以用这个基本定理来描述；在生态学或生物学中，生物种群的增长率与其总体数量之间的关系也可以用这个定理来描述。

总的来说，微积分的基本定理为我们提供了一个统一的框架，使我们能够理解和应用微分和积分在各种不同背景下的关系。