数学基础

贲晛烨 教授

山东大学信息科学与工程学院

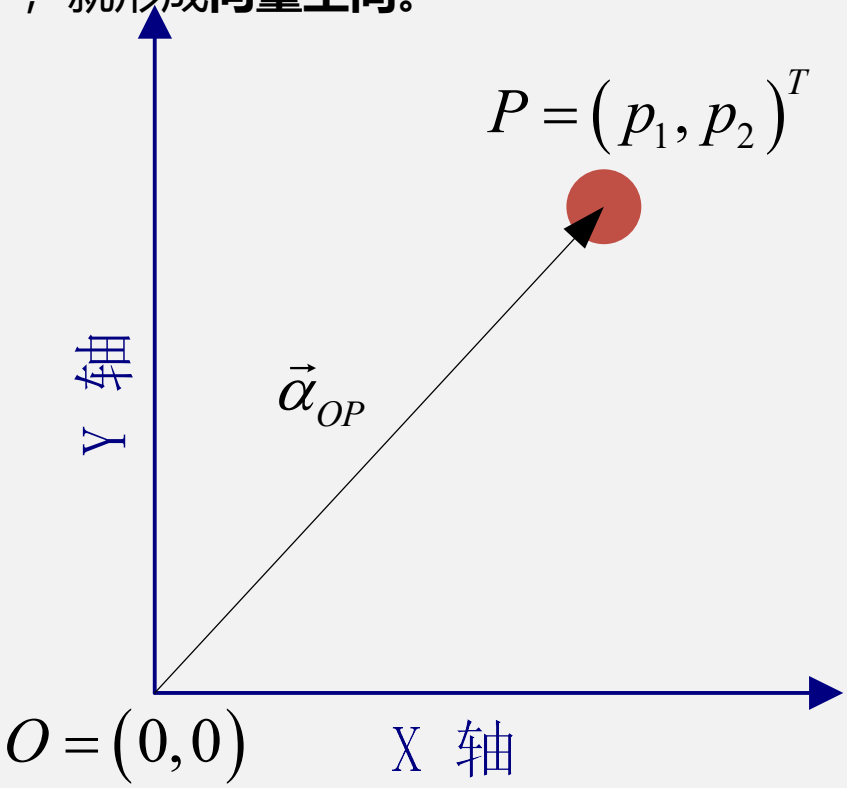
山东大学智慧法治大数据研究中心

2025.9.3

第二章　深度学习的数学基础

2.1.1向量空间

点空间中的每一个点与向量就建立了一一映射。因为向量与点之间的这种一一映射关系，可以把向量转化成几何空间中实在的点，利用点空间的方法来处理向量，这样处理就更加直观；或者把点空间的概念和方法推广到向量中，例如：借助几何中点空间的思路，我们把点空间的概念推广到向量中，就形成**向量空间**。



2.1.1向量空间

直观上，空间是一个几何的概念，但本质上，空间是由数据的运算规则确定的。数学上，空间不仅意味着定义了集合、集合成员、集合元素的运算及其运算规律；并且所有集合元素（即运算对象）按照这些运算规律运算后，运算结果仍然属于这个集合，即**运算具有封闭性**。空间就是由某些运算规则规定下形成的封闭集合，集合中的元素无论如何运算，结果仍然在该集合中。直观地看，就像密闭箱中的气体分子，无论如何运动都超不出箱体的范围。

2.1.1向量空间

给定一个非空集合是V和数域集合F，在V中定义了加法运算$+$，在V与F之间定义了数乘运算$\cdot$，$\alpha,\beta,\gamma \in V, k,l \in F$，如果该加法运算+和数乘运算·同时满足下面所有规则，则称V是F上的向量空间或线性空间。

（1）规则1：若$\alpha,\beta \in V$，则$\alpha+\beta \in V$

（2）规则2：若$\alpha,\beta \in V$，则$\alpha+\beta=\beta+\alpha$

（3）规则3：若$\alpha,\beta,\gamma \in V$，则$(\alpha+\beta)+\gamma=\alpha+(\beta+\gamma)$

（4）规则4：存在零元素$0 \in V$对 都有$0+\alpha=\alpha$

（5）规则5：对任意向量$\alpha \in V$都存在负元素$-\alpha \in V$使得$\alpha+(-\alpha)=0$

2.1.1向量空间

(6) 规则6：若$\alpha \in V, k \in F$，则$k \cdot \alpha \in V$,

(7) 规则7：若$\alpha, \beta \in V, k \in F$，则 $k \cdot (\alpha + \beta) = k \cdot \alpha + k \cdot \beta$

(8) 规则8：若$\alpha \in V, k, l \in F$，则$(k + l) \cdot \alpha = k \cdot \alpha + l \cdot \alpha$

(9) 规则9：若$\alpha \in V, k, l \in F$，则$k \cdot (l \cdot \alpha) = (kl) \cdot \alpha$

(10) 规则10：若$\alpha \in V$，则存在一个单位元素$1 \in F$使得$1 \cdot \alpha = \alpha$

2.1.1向量空间

通常，常见的线性空间如下所示。

（1）$R^{m\times n}$：所有$m\times n$的实矩阵在通常矩阵加法和数乘意义下对实数域R构成线性空间，通常记为$R^{m\times n}$。

（2）$F_n[x]$：次数小于等于n的全体实数多项式函数$F_n(x)=\sum_{i=0}^{n}a_ix^i, a_n\neq 0, a_i\in R, x\in R$集合(含0多项式)在通常的函数加法和函数数乘的意义下对实数域R构成线性空间，通常记为$F_n[x]$。

（3）Nul_A：线性方程$Ax=0$的解集合记作Nul_A，则在通常向量加法和数乘意义下Nul_A是实数域上的线性空间。

（4）Col_A：设$A=[a_1,\cdots,a_i,\cdots,a_n]\in R^{m\times n}$，则$A$的列的线性组合，即其生成空间，记为$Col_A=span(a_1,\cdots,a_i,\cdots,a_n)$，在通常向量加法和数乘意义下$Col_A$是实数域上的线性空间。

2.1.2矩阵分析

设F为数域，由F中任意数量/元素沿行列两个方向有序排列的m行n列的阵列/表格称为矩阵。
若第i行第j列的元素为a_{ij}，则矩阵可以记为$(a_{ij})_{m\times n}$，常记作

$$A=(a_{ij})_{m\times n}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

2.1.2矩阵分析

设$A=(a_{ij})_{m\times n}$和$B=(b_{ij})_{m\times n}$是两个$m\times n$矩阵，则有下面成立。

(1) 若 $A=B$，则有$a_{ij}=b_{ij},i=1,\cdots,m,j=1,\cdots,n$。

(2) 若$B=\alpha A$，则有$b_{ij}=\alpha a_{ij},i=1,\cdots,m,j=1,\cdots,n$。

(3) 若$C=(c_{ij})_{m\times n}=A+B$，则有$c_{ij}=a_{ij}+b_{ij},i=1,\cdots,m,j=1,\cdots,n$。

(3) 若$C=(c_{ij})_{m\times n}=AB$，则有$c_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj},i=1,\cdots,m,j=1,\cdots,n$。

(4) 若$B=A^T$，则有$b_{ij}=a_{ji},i=1,\cdots,m,j=1,\cdots,n$。

第二章　深度学习的数学基础

2.2.1概率与条件概率

概率公理化定义

设从事件/实验的样本空间Ω到闭区间[0, 1] 上的有界映射是 $P:\Omega \to [0,1]$，若事件/实验$A \subseteq \Omega$，并且满足以下三条件，则称$P(A) \in [0,1]$是事件/实验A的**概率**，

（1） $P(A) \in [0,1]$，即概率取值一定在闭区间[0, 1]中，称为**有界性公理**，本公理也说明了$P(A) \geq 0$，故有称为**非负性公理**

（2） $P(\Omega) = 1$，即必然事件概率为1，样本空间中总有某些样本是要发生的，样本空间中全部样本都不发生是不可能的，称为**规范性公理**

（3）设互不相容事件$A_k \subseteq \Omega$ （即若$i \neq j$则$A_i \cap A_j = \phi$）的和事件/实验的概率等各个事件/实验的概率和，即 $P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$，称为**可列可加性公理。**

2.2.1概率与条件概率

概率的最基本性质

（1）不可能事件的概率为0，即$P(\phi)=0$。

（2）有限可加性：n个（n是有限的）两两互不相容事件$A_k \subseteq \Omega$（即若$i \neq j$则$A_i \cap A_j = \phi$）的和事件（即$\cup_{k=1}^{\infty} A_k$）的概率 等于各个事件概率$P(A_k), k=1,2,\cdots,n$的和$\sum_{k=1}^{n} P(A_k)$，即：$P(\cup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k)$。

（3）单调性：若事件A是事件B的子集，则事件A发生的概率不大于事件B发生的概率。即，若$A \subseteq B \subseteq \Omega$，则有$P(A) \leqq P(B)$。

（4）互补性：若事件$\overline{A} \subseteq \Omega$是事件$A \subseteq \Omega$的对立事件，即$\overline{A} \cup A = \Omega$，则有$P(\overline{A}) + P(A) = 1$。

2.2.1概率与条件概率

条件概率的定义
在样本空间Ω中，事件B发生的概率是$P(B)$，在事件B发生的条件下事件A也发生的概率称为条件概率，记作$P(A|B) = \frac{P(A \cap B)}{P(B)}$；同理，在事件A发生的条件下事件B发生的条件概率为$P(B|A) = \frac{P(A \cap B)}{P(A)}$。

2.2.2贝叶斯理论

全概率公式

贝叶斯理论在推断时的最大特点是该方法把推断目标的数据信息、主观经验、先验知识等各类事先已知信息抽象更新了先验概率，根据得到的后验概率对未知信息进行推断。

假设样本空间Ω的完备事件是$\theta_1, \theta_2, \cdots, \theta_n$，$X$是样本空间Ω内某任意事件，根据概率公理体系，易得：

$$P(X=x)=P\left(\sum_{i=1}^{n} x\theta_i\right)=\sum_{i=1}^{n} P(x\theta_i)=\sum_{i=1}^{n} P(x|\theta_i)P(\theta_i)$$

2.2.2贝叶斯理论

贝叶斯公式的基本形式

根据条件概率的定义，可以求解出任意一个的完备事件$\theta_1, \theta_2, \cdots, \theta_n$在事件$X = x$发生后的条件概率如下

$$P(\theta_i|X=x) = \frac{P(X=x,\theta_i)}{P(X=x)} = \frac{P(X=x|\theta_i)P(\theta_i)}{P(X=x)} = \frac{P(X=x|\theta_i)P(\theta_i)}{\sum_{i=1}^{n} P(X=x|\theta_i)P(\theta_i)} = \frac{P(x|\theta_i)P(\theta_i)}{\sum_{i=1}^{n} P(x|\theta_i)P(\theta_i)}$$

$P(\theta_i)$表示在不知道事件$X = x$发生的情况下事件θ_i的发生概率，代表着人们事先（此处主要是指在事件$X = x$发生之前）对θ_i的认识，故称为**先验概率**；但是，当人们获得了新信息后（此处主要是指已知了事件$X = x$的发生）会综合分析这些新信息（此处是指事件$X = x$的信息），从而会对事件θ_i的发生产生了新认识，即在事件$X = x$发生后的条件下事件θ_i发生的条件概率$P(\theta_i|X = x)$，故称为**后验概率**。

2.2.2贝叶斯理论

贝叶斯估计的基本形式

假设待估计的未知参数θ（因为估计对象往往是未知参数，故这里用未知参数θ代替估计对象）的先验分布是$\pi(\theta)$，在获得样本（因为结果事件往往是抽样样本的结果，故这里用样本/抽样来作为结果事件）x后，即在$X = x$的条件下的**条件分布记为$\pi(\theta|x)$**，则有

$$\pi\left(\theta|x\right)=\frac{h(x,\theta)}{m(x)}=\frac{f\left(x|\theta\right)\pi(\theta)}{\int_{\Theta} f\left(x|\theta\right)\pi(\theta)d\theta}$$

其中 π 指的是参数的概率分布，$\pi(\theta)$ 指的是先验概率，$\pi(\theta|x)$ 指的是后验概率，$f(x|\theta)$ 指的是我们观测到的样本的分布，也就是似然函数(likelihood)，记住 **竖线 | 左边的才是我们需要的。**其中积分求的区间 Θ 指的是参数 θ 所有可能取到的值的域，所以可以看出后验概率 $\pi(\theta|x)$ 是在知道 x 的前提下在 Θ 域内的一个关于 θ 的概率密度分布，每一个 θ 都有一个对应的可能性(也就是概率)。

2.2.2贝叶斯理论

贝叶斯估计的基本形式
贝叶斯理论认为，关于θ的一切统计和推断都是必须基于参数θ的后验分布$\pi(\theta|x)$，$\pi(\theta|x)$是贝叶斯推断的最主要依据和出发点。**用后验分布$\pi(\theta|x)$的均值作为未知参数θ的估计，称为后验期望估计**，其计算公式如下所示：

$$\hat{\theta} = E\left(\theta|x\right) = \int_{\Theta} \theta\pi\left(\theta|x\right)d\theta = \frac{\int_{\Theta} \theta h(x,\theta)d\theta}{m(x)} = \frac{\int_{\Theta} \theta f\left(x|\theta\right)\pi(\theta)d\theta}{\int_{\Theta} f\left(x|\theta\right)\pi(\theta)d\theta}$$

其中$h(x,\theta), m(x)$分别是联合分布和边缘分布，**而$\pi(\theta|x)$是用密度函数表示的贝叶斯公式**，或者叫作贝叶斯公式的密度函数形式。

2.2.2信息论基础

信息必须满足以下四条公理：

（1）若信源符号a_i, a_j的概率是$p(a_i), p(a_j)$，且$p(a_i) > p(a_j)$，则$I(a_i) < I(a_j)$。

（2）若信源符号a_i的概率是$p(a_i)$，且$p(a_i) = 0$，即a_i是不可能事件，则$I(a_i) \to \infty$，不可能事件包含无穷大的信息量。

（3）若信源符号a_i的概率是$p(a_i)$，且$p(a_i) = 1$，即a_i是确定事件，则$I(a_i) = 0$，没有随机性的确定事实不含任何信息量。

（4）若信源符号a_i, a_j是统计独立的，例如，来自两个相互独立的信源，这两个消息总的信息量即联合信息量记为$I(a_i, a_j)$，则$I(a_i, a_j) = I(a_i) + I(a_j)$。

2.2.2信息论基础

信息熵：
若信源X可以随机地发出r个不同的符号，记为$a_i, i = 1,2,\cdots,r$，并且每一个符号a_i产生的概率是$p(a_i)$，显然每个符号a_i有自信息量$I(a_i) = -log_2(p(a_i))$。若在该信源 的概率空间$p(a_i), i = 1,2,\cdots,r$中统计所有符号$a_i, i = 1,2,\cdots,r$的平均信息量，并作为信源X的信息测度，称为信源X的信息熵，记作$H(X)$，即：

$$H(X) = -\sum_{i=1}^{r} p(a_i)\log_2(p(a_i))$$

2.2.2信息论基础

联合熵：

若符号a_i, b_j的联合概率是$p(a_i, b_j)$，则定义联合熵如下：

$$H(XY) = -\sum_{i=1}^{r}\sum_{j=1}^{s} p(a_i, b_j)\log p(a_i, b_j)$$

若 a_i, b_j分别是发送和接受的符号，则$H(XY)$表示发送了符号a_i并且一定能接受到符号b_j的后验平均不确定性，故也称为共熵。类似地，根据条件概率可以定义条件熵如下：

2.2.2信息论基础

根据条件概率可以定义条件熵如下：

$$H(X|Y)=\sum_{j=1}^{s}p(b_j)H(X|Y=b_j)$$

$$=-\sum_{j=1}^{s}p(b_j)\sum_{i=1}^{r}p(a_i|b_j)\log p(a_i|b_j)$$

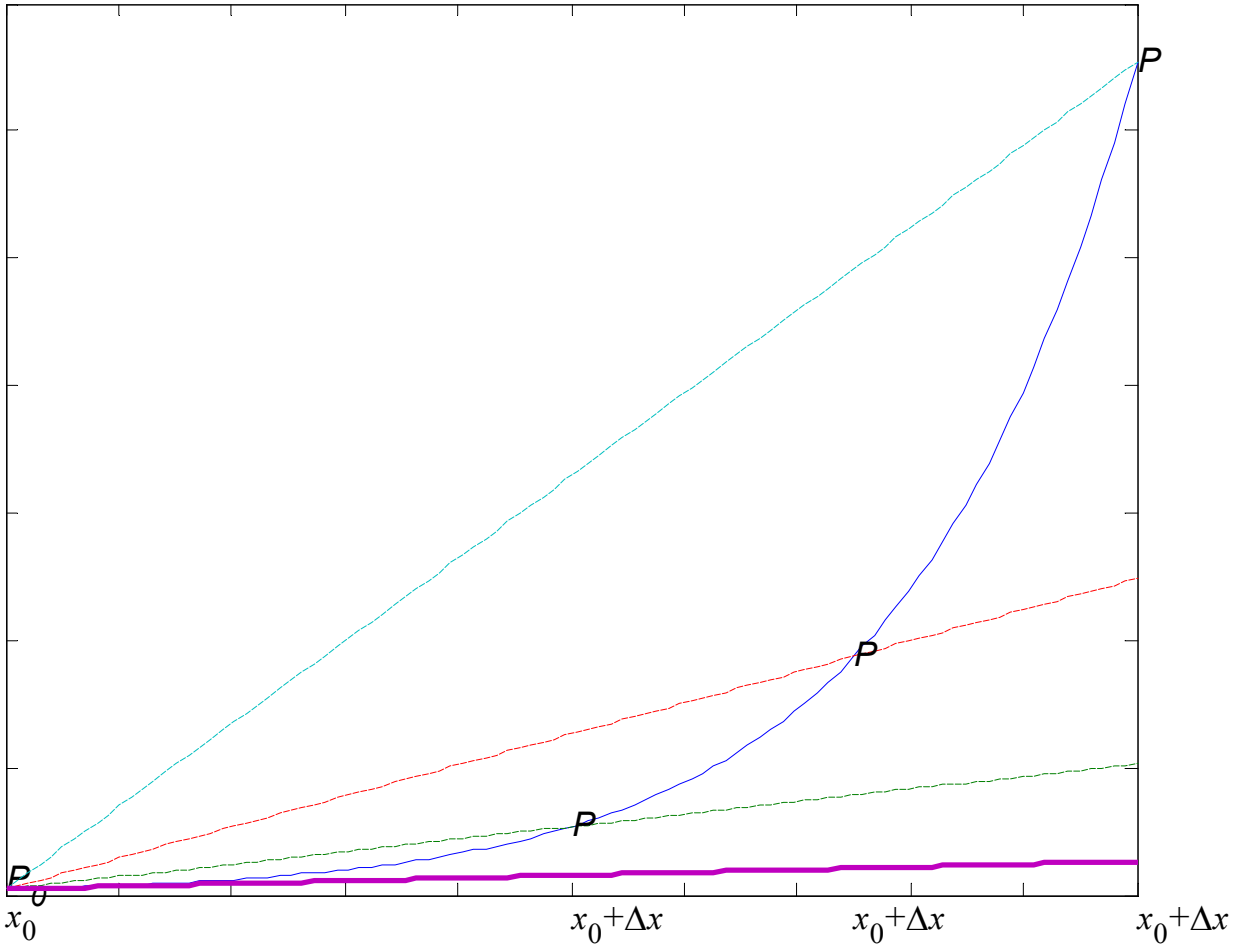
$$=-\sum_{j=1}^{s}\sum_{i=1}^{r}p(a_i|b_j)\log p(a_i|b_j)$$

第二章　深度学习的数学基础

2.3.1导数和偏导数

导数：
为了刻画所有物理量的瞬时变化率，数学中把它们作了归纳和抽象，并引入了导数的概念，用导数来定义一切物理量的变化率。设任意物理量用$y = f(x)$表示，其上任意点P_0记为$\left(x_0, f(x_0)\right)$，再在该点邻域附近取一点$P\left(x, f(x)\right)$做割线P_0P。若记$\Delta x = x - x_0, \Delta y = f(x + \Delta x) - f(x_0)$，显然，物理量$y = f(x)$在$[x_0, x]$之间的平均变化率为$\frac{\Delta y}{\Delta x} = \frac{f\left(x+\Delta x\right)-f(x_0)}{x-x_0}$，正是割线的斜率。当点$P$沿曲线移动，无限接近点$P_0$时，直线与曲线只有一个交点，割线变成了切线，相应地平均变化率也变成了瞬时变化率，其数值等于切线的斜率，这就是该点的导数。

2.3.1导数和偏导数



2.3.1导数和偏导数

导数：
设函数$f(x)$在点x_0的邻域内有定义，这样，当自变量从x_0变化到$x_0+\Delta x$时函数值的变化为$\Delta y=f(x+\Delta x)-f(x_0)$，若自变量的变化量$\Delta x$趋于无穷小时，比率$\frac{\Delta y}{\Delta x}$的极限存在，如下式所示，则该极限称为函数$f(x)$在点$x_0$的导数，通常记为$f'(x_0)$或$\frac{\mathrm{d}y}{\mathrm{d}x}|_{x=x_0}$，并称函数 在点 可导/可微；若极限不存在则称 在点 不可导。

$$\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=\lim_{\Delta x\to 0}\frac{\Delta y=f\left(x_0+\Delta x\right)-f\left(x_0\right)}{\Delta x}$$

2.3.1导数和偏导数

常见函数的导数

序号	$f(x)$	$f'(x)$	序号	$f(x)$	$f'(x)$
1	C	0	5	$\log_a x$	$\frac{1}{x \ln a}$
2	x^α	$\alpha x^{\alpha-1}$	6	$\ln x$	$\frac{1}{x}$
3	a^x	$a^x lnx$	7	$\sin x$	$\cos x$
4	e^x	e^x	8	$\cos x$	$-\sin x$

2.3.1导数和偏导数

偏导数：
设函数$z = f(x, y)$在点(x_0, y_0)的邻域内有定义，这样，当自变量x从x_0变化到$x_0 + \Delta x$时而y固定在y_0时,函数值的变化为$\Delta y = f(x + \Delta x, y_0) - f(x_0, y_0)$，若自变量的变化量$\Delta x$趋于无穷小时，比率$\frac{\Delta_x z}{\Delta x}$的极限存在，如下式所示，则该极限称为函数$f(x, y)$在点$(x_0, y_0)$的导数，通常记为$f'(x_0, y_0)$或$\frac{\partial y}{\partial x}|_{(x_0, y_0)}$，并称函数 $z = f(x, y)$在点(x_0, y_0)可导/可微；若极限不存在则称$z = f(x, y)$在点(x_0, y_0)不可导。

$$\left.\frac{\partial z}{\partial x}\right|_{(x_0, y_0)} = \lim_{\Delta x \to 0} \frac{\Delta_x z}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}$$

2.3.2梯度和海森矩阵

梯度：
根据上面偏导数的定义，当对某个变量求偏导数时函数中所有其它变量要被当作常数。即把函数当作只含该变量的一元函数，然后根据一元函数的求导法则进行求导即可。这样二元函数的偏导数定义可以推广到三元函数和多元函数，它们的偏导数求解都是类似的。
记n元实函数$f: R^n \to R$为$f(X)$，其中$X = (x_1, x_2, \cdots, x_n)'$是$n$维自变量。如果$f(X)$在每一个分量$x_i, i = 1,2, \cdots, n$一阶可导，即偏导数$\frac{\partial y}{\partial x_i}, i = 1,2, \cdots, n$都存在，则称$f(X)$在点$X$处一阶可导，并且把偏导数组成的向量$\nabla f(X) = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_n},\right)'$称为$f(X)$在点$X$处的一阶导数，也即梯度，常记为$\nabla f(X)$。

2.3.2梯度和海森矩阵

海森矩阵：
记n元实函数$f: R^n \to R$为$f(X)$，其中$X = (x_1, x_2, \cdots, x_n)'$是$n$维自变量。如果$f(X)$在每一个分量$x_i, i = 1,2, \cdots, n$二阶可导，即二阶偏导数$\frac{\partial^2 y}{\partial^2 x_i x_j}, i = 1,2, \cdots, n, j = 1,2, \cdots, n$都存在，则称$f(X)$在点$X$处二阶可导，并且把偏导数组成的矩阵称为 $f(X)$点X处的二阶导数，也即海森(Hesse)矩阵，常记为$\nabla^2 f(X)$。

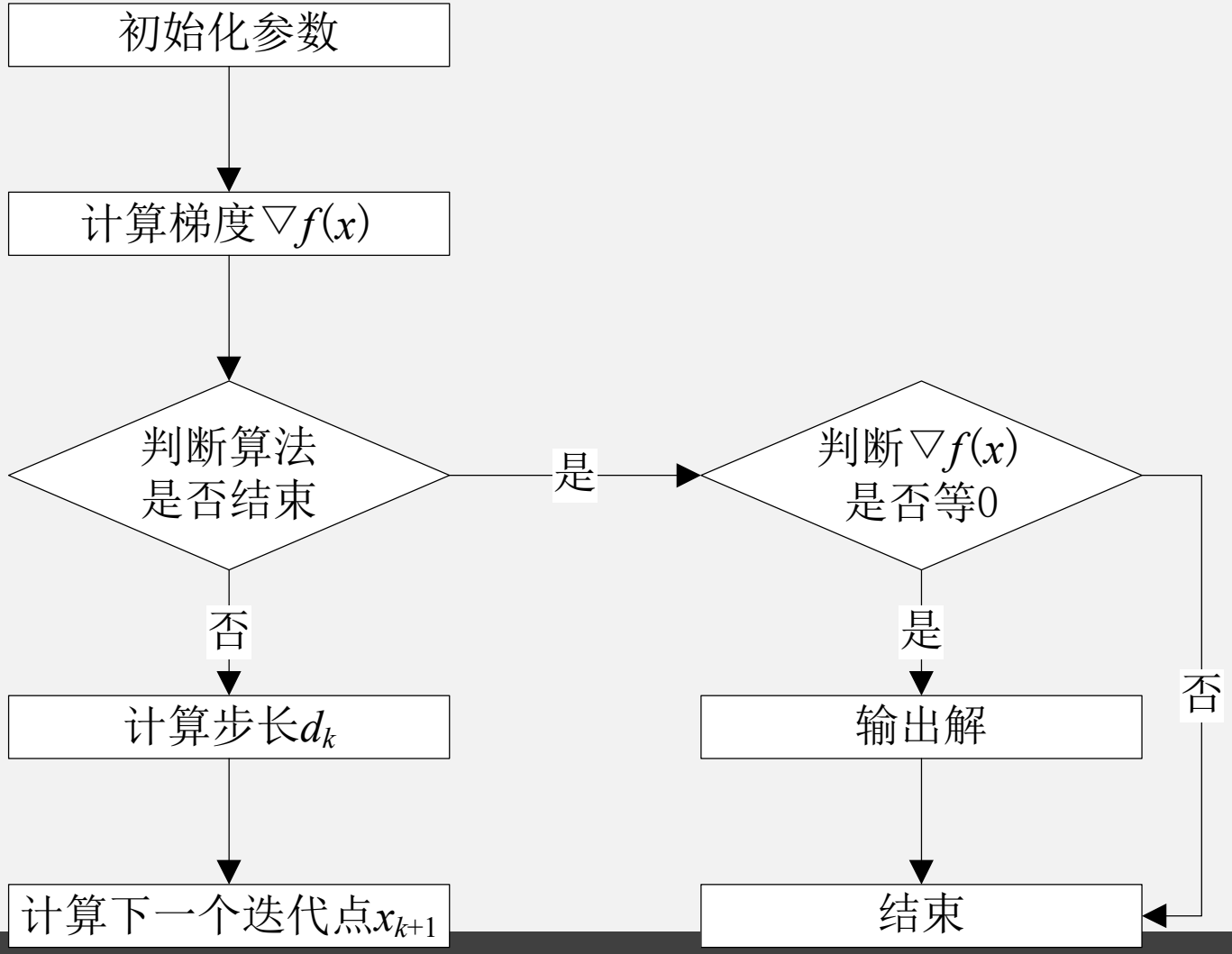
$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

2.3.3最速下降法

假设最速下降法求解函数$f(X)$时第k次迭代到了点x_k，则选择第k次迭代的搜索方向d_k为最速下降方向，**即搜索方向是负梯度方向**$-\nabla f(X)$，也就是令$d_k = -\nabla f(X)$。然后从x_k出发沿方向d_k搜索函数 的最小值，也就是在射线 $x_k + \lambda d_k$（其中$\lambda > 0$是射线的参数变量，表示与点x_k的距离）上找一点使得该点的函数值$f(X)$最小。假设该最小点距当前点x_k的距离是λ_k，则可以表示为：$f(x_k + \lambda_k d_k) = \min_{\lambda \geq 0} f(x_k + \lambda d_k)$，其中$\lambda_k$也称为搜索的步长。求步长$\lambda_k$通常被当作一个线搜索问题，也就是一元函数的优化问题。一旦求出了步长，就确定了沿方向d_k能找到的最小函数值$x_k + \lambda_k d_k$，下一次迭代就以该点为起点，即$x_{k+1} = x_k + \lambda_k d_k$。

2.3.3最速下降法

最速下降算法的流程图



2.3.3最速下降法

最速下降算法的步骤

Step 1. 初始化算法的参数。

Step 2. 计算目标函数$f(X)$在x_k的梯度$\nabla f(X)$。通常，不同函数的梯度是不一样的，有时也可以用梯度的近似值代替梯度的精确值。

Step 3. 判断算法迭代是否满足终止条件，若满足则转步6，否则转步4。迭代终止条件在不同条件下也有所不同。

Step 4. 令搜索方向$d_k = -\nabla f(X)$，选择某种一维线搜索方法计算算法步长λ_k。

Step 5. 令$x_{k+1} = x_k + \lambda_k d_k$，转步2重复上述计算过程。

Step 6. 判断梯度在x_k是否为0。实际问题中只需要 近似接近0即可。若满足，则可以认为当前点 是目标函数的最小值。这样就可以结束迭代，停止算法。

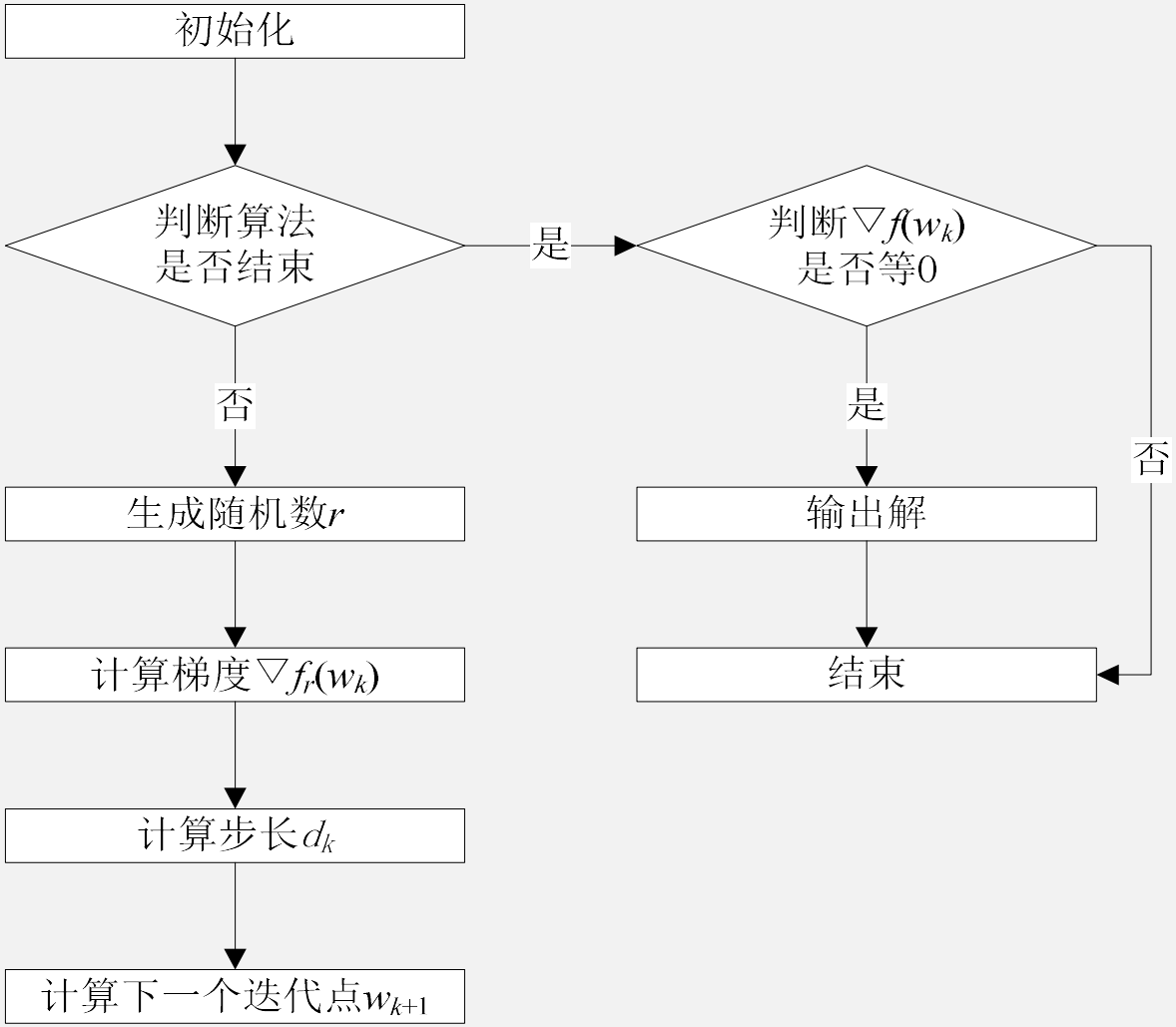
2.3.4随机梯度下降算法

随机梯度下降算法

如前所述，经典梯度下降算法虽然具有广泛地适用性，但是求解机器学习领域中的训练问题效率非常低，有必要进一步改进。根据概率统计学中的大数定理，当样本量很大或趋于无穷时大量样本的均值与任意一个样本母体近似相等。注意到我们需要求解的正好是梯度关于N个样本的均值，这样如果把每一个样本当作随机的，则在大样本条件下，任意一个样本的梯度与N个样本梯度的均值近似相等。这样，用一个随机样本的梯度来代替N个样本梯度的均值不仅是可行的，而且减少了计算量提高了计算效率。因为样本是已知的自然也是确定的，为了让已知样本具有随机性，通常采用无放回抽样策略，即从样本集中随机选择一个样本用它的梯度来代替所有样本梯度的均值。这样就增加了随机性，确定的梯度下降法就变成了随机梯度下降算法。

2.3.4随机梯度下降算法

随机梯度下降算法的的流程图



2.3.4随机梯度下降算法

随机梯度下降算法的的步骤
Step 1. 初始化算法的参数。
Step 2. 判断迭代是否满足结束条件，若满足则转步7；若不满足则转步3。
Step 3. 产生随机数$r \in [1,2,\cdots,N]$，即选择来了函数$f_r(w)$。
Step 4. 计算函数$f_r(w)$在x_k的梯度$\nabla f_r(x_k)$。
Step 5. 令搜索方向$d_k = \nabla f_r(x_k)$，选择某种一维线搜索方法计算迭代步长λ_k。
Step 6. 令$x_{k+1} = x_k + \lambda_k d_k$，步2重复上述计算过程。
Step 7. 判断梯度在x_k是否为0。实际问题中只需要 近似接近0即可，即 是否满足。若满足，则可以认为当前点 是目标函数的极小值。这样就可以结束迭代，停止算法。

感谢聆听