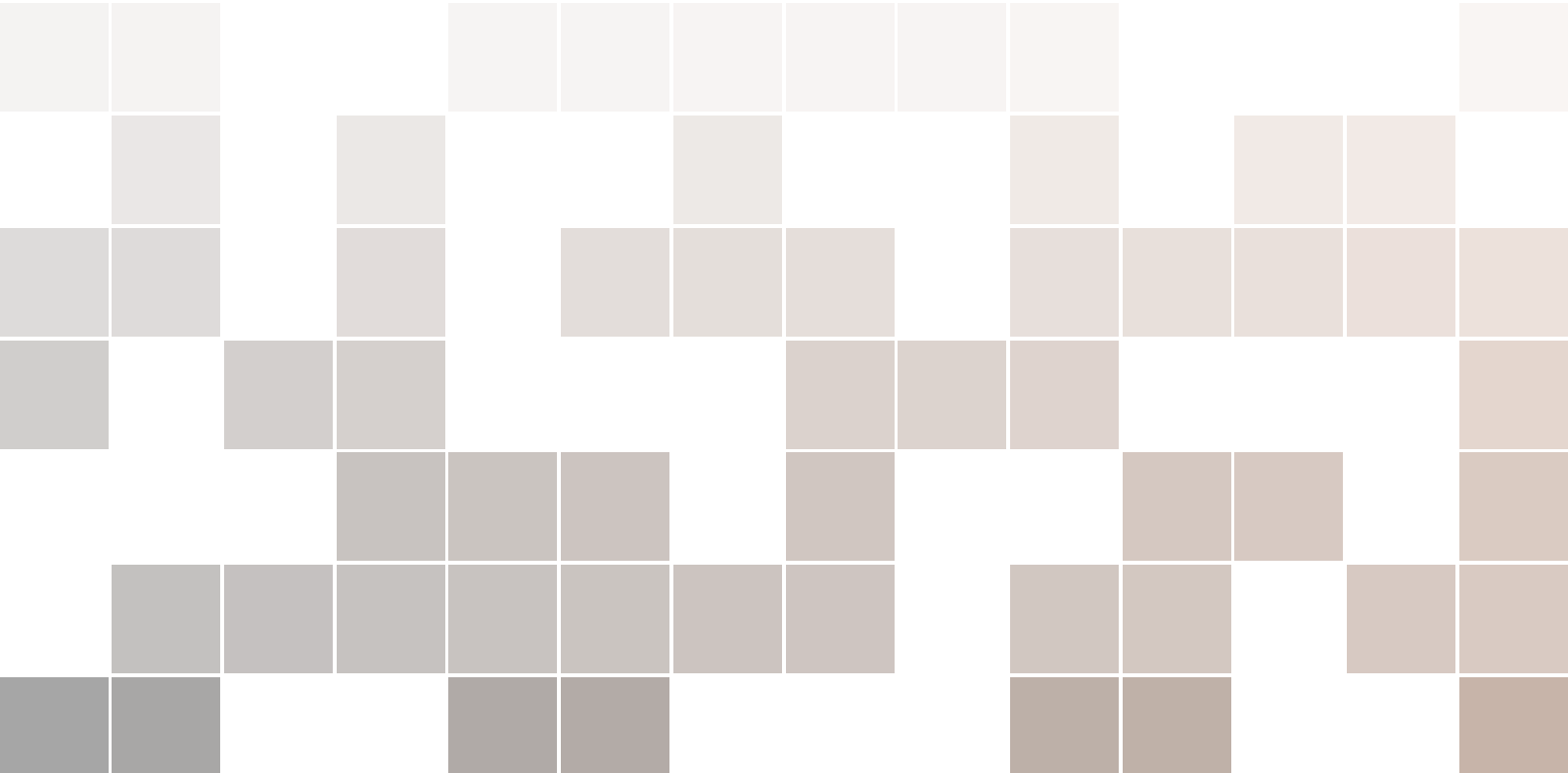




The Beauty of Embracing English

My English Note

Xia Wenxuan



Written by Xia Wenxuan, 2021

PUBLISHED BY MYSELF

<https://github.com/gegeji>

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, September 24, 2021

Contents

I	Basics of Linear Algebra	
1	对向量的介绍	7
1.1	Vector	7
1.2	Vector Space	8
1.3	向量运算	8
1.4	内积	9
1.4.1	常用的内积等式	10
1.5	Cauchy-Schwartz Inequality	10
1.6	浮点运算	10
2	Linear Function	11
2.1	Linear Function	11
2.2	泰勒展开	12
2.3	Regression Model	14
3	Norm and Distance	15
3.1	Norm	15
3.2	Root Mean Square Value (RMS)	17
3.3	Chebyshev's Inequality	17
3.4	Distance	17
3.4.1	Feature Distance and Nearest Neighbor	18

3.5	Standard Derivation	18
3.6	Angle	18
3.6.1	相关系数	19
4	以 k -Means 算法为例的优化问题	21
4.1	Convex Set	22
4.2	向量偏导	23
4.3	标量优化问题的例子	24
	Bibliography	25
	Articles	25
	Books	25

Basics of Linear Algebra

1	对向量的介绍	7
1.1	Vector	
1.2	Vector Space	
1.3	向量运算	
1.4	内积	
1.5	Cauchy-Schwartz Inequality	
1.6	浮点运算	
2	Linear Function	11
2.1	Linear Function	
2.2	泰勒展开	
2.3	Regression Model	
3	Norm and Distance	15
3.1	Norm	
3.2	Root Mean Square Value (RMS)	
3.3	Chebyshev's Inequality	
3.4	Distance	
3.5	Standard Deviation	
3.6	Angle	
4	以 k-Means 算法为例的优化问题	21
4.1	Convex Set	
4.2	向量偏导	
4.3	标量优化问题的例子	
	Bibliography	25
	Articles	
	Books	

1. 对向量的介绍

1.1 Vector

Definition 1.1.1 — Vector. 一个有序的数字列表.

$$\begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{bmatrix} \text{ 或者 } \begin{pmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{pmatrix} \text{ 或者 } (-1.1, 0, 3.6, -7.2)$$

表中的数字是元素 (项、系数、分量)。元素的数量是向量的大小 (维数, 长度)。大小为 n 的向量称为 n 维向量。向量中的数字通常被称作标量。

用符号来表示向量, 比如 α , b , 一般小写字母表示. 其它表示形式 g, \vec{a}

Definition 1.1.2 — n 维向量 a 的第 i 元素. n 维向量 a 的第 i 元素表示为 a_i .

有时 i 指的是向量列表中的第 i 个向量.

Definition 1.1.3 — $a = b$. 对于所有 i , 如果有 $a_i = b_i$, 则称两个相同大小的向量 a 和 b 是相等的, 可写成 $a = b$

Definition 1.1.4 — stacked vector. 假设 b 、 c 、 d 是大小为 m 、 n 、 p 的向量

$$a = \begin{bmatrix} b \\ c \\ d \end{bmatrix}$$

$$a = (b_1, b_2, \dots, b_m, c_1, c_2, \dots, c_n, d_1, d_2, \dots, d_p)$$

Definition 1.1.5 — 零向量. 所有项为 0 的 n 维向量表示为 0_n 或者 0

Definition 1.1.6 — 全一向量. 所有项为 1 的 n 维向量表示为 $\mathbf{1}_n$ 或者 1

Definition 1.1.7 — 单位向量. 当第 i 项为 1, 其余项为 0 时表示为 e_i

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Definition 1.1.8 — 稀疏向量. 如果一个向量的许多项都是 0, 该向量为稀疏 (Sparse) 的。稀疏向量能在计算机上高效地存储和操作。

$\text{nnz}(x)$ 是指向量 x 中非零的项数 (number of non-zeros), 有时用 ℓ_0 表示。

向量 $x = (x_1, x_2)$ 可以在二维中表示一个位置或一个位移、图像、单词统计等。

1.2 Vector Space

Definition 1.2.1 — 向量空间 V . 设 V 是非空子集, P 是一数域, 向量空间 V 满足:

1. 向量加法: $V + V \rightarrow V$, 记作 $\forall x, y \in V$, 则 $x + y \in V$ (加法封闭)
2. 标量乘法: $F \times V \rightarrow V$, 记作 $\forall x \in V, \lambda \in P$, 则 $\lambda x \in V$ (乘法封闭)

上述两个运算满足下列八条规则 ($\forall x, y, z \in V, \lambda, \mu \in P$)

1. $x + y = y + x$ (交换律)
2. $x + (y + z) = (x + y) + z$ (结合律)
3. V 存在一个零元素, 记作 0 , $x + 0 = x$
4. 存在 x 的负元素, 记作 $-x$, 满足 $x + (-x) = 0$
5. $\forall x \in V$, 都有 $1x = x, 1 \in P$
6. $\lambda(\mu x) = (\lambda\mu)x$
7. $(\lambda + \mu)x = \lambda x + \mu x$
8. $\lambda(x + y) = \lambda x + \lambda y$

Corollary 1.2.1 向量空间也称为线性空间。

Corollary 1.2.2 如果 $x, y \in \mathbb{R}^2$, 则 $x + y \in \mathbb{R}^2, \lambda x \in \mathbb{R}^2 (\lambda \in \mathbb{R})$

1.3 向量运算

Definition 1.3.1 — 向量加法. n 维向量 a 和 b 可以相加, 求和形式表示为 $a + b$

设向量 a, b, c 是向量空间 V 的元素, 即 $a, b, c \in V$ 。

1. 交换律: $a + b = b + a$
2. 结合律: $(a + b) + c = a + (b + c)$ (因此可写成 $a + b + c$)
3. $a + 0 = 0 + a = a$
4. $a - a = 0$

Corollary 1.3.1 — 向量位移相加. 如果二维向量 a 和 b 都表示位移, 则它们的位移之和为 $a + b$

Definition 1.3.2 — 标量与向量的乘法.

$$\beta a = \begin{bmatrix} \beta a_1 \\ \vdots \\ \beta a_n \end{bmatrix}$$

标量 β, γ 与向量 a, b

1. 结合律: $(\beta\gamma)a = \beta(\gamma a)$
2. 左分配律: $(\beta + \gamma)a = \beta a + \gamma a$
3. 右分配律: $\beta(a + b) = \beta a + \beta b$

Definition 1.3.3 — 线性组合. 对于向量 a_1, \dots, a_m 和标量 β_1, \dots, β_m ,

$$\beta_1 a_1 + \dots + \beta_m a_m$$

是向量的线性组合。 β_1, \dots, β_m 是该向量的系数。

■ **Example 1.1** 对于任何向量 $b \in \mathbb{R}^n$, 有如下等式

$$b = b_1 e_1 + \dots + b_n e_n, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

1.4 内积

Definition 1.4.1 — 内积. 在数域 \mathbb{R} 上的向量空间 V , 定义函数 $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$, 满足:

1. $\langle a, a \rangle \geq 0, \forall a \in V$, 当且仅当 $a = 0$ 时 $\langle a, a \rangle = 0$
2. $\langle \alpha a + \beta b, c \rangle = \alpha \langle a, c \rangle + \beta \langle b, c \rangle, \forall \alpha, \beta \in \mathbb{R}$, 且 $a, b, c \in V$
3. $\langle a, b \rangle = \langle b, a \rangle, \forall a, b \in V$

函数 $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ 成为内积。

■ **Example 1.2** 在向量空间 \mathbb{R}^n 上, 计算两个向量对应项相乘之后求和函数

$$\langle a, b \rangle = a_1 b_1 + a_2 b_2 + \dots + a_n b_n = a_b^T$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^n$$

Proof. $\langle a, a \rangle = a_1 a_1 + a_2 a_2 + \dots + a_n a_n = \sum_{i=1}^n a_i^2 \geq 0, \langle a, a \rangle = 0$, 则 $a = 0$

$$\langle \alpha a + \beta b, c \rangle = (\alpha a_1 + \beta b_1) c_1 + (\alpha a_2 + \beta b_2) c_2 + \dots + (\alpha a_n + \beta b_n) c_n$$

$$= \alpha \sum_{i=1}^n a_i c_i + \beta \sum_{i=1}^n b_i c_i$$

$$= \alpha \langle a, c \rangle + \beta \langle b, c \rangle$$

$$\langle a, b \rangle = a^T b = b^T a = \langle b, a \rangle$$

内积的性质: 交换律、结合律、分配律。

交换律: $a^T b = b^T a$

结合律: $(\gamma a)^T b = \gamma (a^T b)$

分配律: $(a + b)^T c = a^T c + b^T c$

1.4.1 常用的内积等式

Corollary 1.4.1 — 选出第 i 项.

$$e_i^T a = a_i$$

Corollary 1.4.2 — 向量每一项之和.

$$\mathbf{1}^T a = a_1 + \cdots + a_n$$

Corollary 1.4.3 — 向量每一项的平方和.

$$a^T a = a_1^2 + \cdots + a_n^2$$

1.5 Cauchy-Schwartz Inequality

Theorem 1.5.1 — **Cauchy-Schwartz Inequality**. 设 $\langle \cdot, \cdot \rangle$ 是向量空间 V 上的内积, $\forall x, y \in V$, 则有

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

Proof. 令 $\lambda \in \mathbb{R}$, 则有 $0 \leq \langle x + \lambda y, x + \lambda y \rangle = \langle x, x \rangle + \lambda \langle y, x \rangle + \lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle = \langle x, x \rangle + 2\lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle$

则有 $\lambda^2 \langle y, y \rangle + 2\lambda \langle y, x \rangle + \langle x, x \rangle \geq 0, \forall \lambda \in \mathbb{R}$.

$$\nabla = (2\langle y, x \rangle)^2 - 4\langle y, y \rangle \langle x, x \rangle \leq 0$$

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

当 $|\langle x, y \rangle|^2 = \langle x, x \rangle \langle y, y \rangle$ 时, 有 $\langle x, x \rangle^2 + 2\lambda \langle y, x \rangle + \lambda^2 \langle y, y \rangle = 0$

也即 $\langle x + \lambda y, x + \lambda y \rangle = 0$, 因此 $x + \lambda y = 0$, 即 $x = -\lambda y$. ■

1.6 浮点运算

计算机以浮点格式存储(实)数值。

基本的算术运算(加法, 乘法等)被称为浮点运算(flop)。

算法或操作的时间复杂度: 作为输入维数的函数所需要的浮点运算总数。

算法复杂度通常以非常粗略地近似估算。

(程序)执行时间的粗略估计: 计算机速度/flops

目前的计算机大约是 1Gflops/秒 (10^9 flops/秒)

Corollary 1.6.1 假设有 n 维向量 x 和 y :

- $x + y$ 需要 n 次加法, 所以时间复杂度为 (n) flops。
- $x^T y$ 需要 n 次乘法和 $n - 1$ 次加法, 所以时间复杂度为 $(2n - 1)$ flops。
- 对于 $x^T y$, 通常将其时间复杂度简化为 $2n$, 甚至为 n 。
- 当 x 或 y 是稀疏的时候, 算法的实际运算时间会比理论时间更少。

2. Linear Function

2.1 Linear Function

Definition 2.1.1 — Linear Function. f 是一个将 n 维向量映射成数的函数。

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

线性函数 f 满足以下两个性质 ($k \in \mathbb{R}, x, y \in \mathbb{R}^n$):

- 齐次性 (homogeneity): $f(kx) = kf(x)$
- 叠加性 (Additivity): $f(x + y) = f(x) + f(y)$

■ **Example 2.1** 求平均值: $f(x) = \frac{1}{n} \sum_{i=1}^n x_i$ 为线性函数。 ■

■ **Example 2.2** 求最大值: $f(x) = \max \{x_1, x_2, \dots, x_n\}$ 并不是线性函数。 ■

Proof. 令 $x = (1, -1), y = (-1, 1), \alpha = 0.5, \beta = 0.5$, 有 $f(\alpha x + \beta y) = 0 \neq \alpha f(x) + \beta f(y) = 1$

$$\begin{aligned} f(x + y) &= \max \{x_1 + y_1, x_2 + y_2, \dots, x_n + y_n\} \\ &\leq \max \{x_1, x_2, \dots, x_n\} + \max \{y_1, y_2, \dots, y_n\} \\ &\leq f(x) + f(y) \end{aligned}$$

■

Theorem 2.1.1 设 $\alpha_1, \dots, \alpha_m \in \mathbb{R}, u_1, \dots, u_m \in \mathbb{R}^n$, 则线性函数 f 满足

$$\begin{aligned} f(\alpha_1 u_1 + \alpha_2 u_2 + \dots + \alpha_m u_m) &= f(\alpha_1 u_1) + f(\alpha_2 u_2 + \dots + \alpha_m u_m) \\ &= \alpha_1 f(u_1) + f(\alpha_2 u_2 + \dots + \alpha_m u_m) \\ &= \alpha_1 f(u_1) + \alpha_2 f(u_2) + \dots + \alpha_m f(u_m) \end{aligned}$$

Definition 2.1.2 — 内积函数 (inner product function). 对于 n 维向量 a , 满足以下形式的函数被称为内积函数

$$f(x) = a^T x = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$$

上述 $f(x)$ 可以看作是每项 x_i 的加权之和。

Corollary 2.1.2 内积函数都是线性的。

Proof.

$$\begin{aligned} f(\alpha x + \beta y) &= a^T (\alpha x + \beta y) \\ &= a^T (\alpha x) + a^T (\beta y) \\ &= \alpha (a^T x) + \beta (a^T y) \\ &= \alpha f(x) + \beta f(y) \end{aligned}$$

Definition 2.1.3 — 仿射函数 (affine function). 其一般形式为 $f(x) = a^T x + b$, 其中 $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ 为标量。

Theorem 2.1.3 函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 为仿射函数需要满足

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y), \alpha + \beta = 1, \alpha, \beta \in \mathbb{R}, x, y \in \mathbb{R}^n$$

2.2 泰勒展开

Definition 2.2.1 — 函数 f 第 i 个分量的一阶偏导数。

$$\begin{aligned} \frac{\partial f}{\partial z_i}(z) &= \lim_{t \rightarrow 0} \frac{f(z_1, \dots, z_{i-1}, z_i + t, z_{i+1}, \dots, z_n) - f(z)}{t} \\ &= \lim_{t \rightarrow 0} \frac{f(z + te_i) - f(z)}{t} \end{aligned}$$

Definition 2.2.2 — f 在点 z 的梯度。

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f}{\partial z_1}(z) \\ \vdots \\ \frac{\partial f}{\partial z_n}(z) \end{bmatrix}$$

Definition 2.2.3 — Taylor's Approximation.

$$\begin{aligned} f(x) &= f(z) + \frac{\partial f}{\partial x_1}(z)(x_1 - z_1) + \frac{\partial f}{\partial x_2}(z)(x_2 - z_2) + \dots + \frac{\partial f}{\partial x_n}(z)(x_n - z_n) \\ &\quad + \frac{1}{2!} \sum_{j=1}^n \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(z)(x_i - z_i)(x_j - z_j) + \dots \end{aligned}$$

■ **Example 2.3** 泰勒公式利用多项式在一点附近逼近函数

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots + (-1)^{k-1} \frac{x^{2k-1}}{(2k-1)!} + \frac{\sin \left[\xi + (2k+1) \frac{\pi}{2} \right]}{(2k+1)!} x^{2k+1}$$

一次逼近: $\sin x \approx x$

三次逼近: $\sin x \approx x - \frac{x^3}{3!}$ ■

Proof.

$$f(x) = P_n(x) + R_n(x)$$

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n$$

$$R_n(x) = o((x - x_0)^n)$$

$$f(x) \approx P_n(x)$$

$$\therefore P_n(x_0) = f(x_0), P'_n(x_0) = f'(x_0), P''_n(x_0) = f''(x_0), \dots, P_n^{(n)}(x_0) = f^{(n)}(x_0)$$

$$\text{要求 } P_n(x_0) = f(x_0) \Rightarrow a_0 = f(x_0)$$

$$P'_n(x) = a_1 + 2a_2(x - x_0) + \cdots + na_n(x - x_0)^{n-1} \Rightarrow a_1 = f'(x_0)$$

$$\text{依此类推. } a_n = \frac{f^{(n)}(x_0)}{n!}$$
 ■

Corollary 2.2.1 — n 阶泰勒多项式.

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

$$\text{where } a_n = \frac{f^{(n)}(x_0)}{n!}$$

Corollary 2.2.2 — 对于高阶余项的公式. 带拉格朗日余项的泰勒公式

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1} \quad (\xi \text{ 在 } x_0 \text{ 与 } x \text{ 之间})$$

Corollary 2.2.3 — 麦克劳林 (Maclaurin) 公式. 在零点展开麦克劳林 (Maclaurin) 公式

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1} \quad (0 < \theta < 1)$$

Definition 2.2.4 — 一阶泰勒公式. 假设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$, 函数 f 在 z 点可导

$$\hat{f}(x) = f(z) + \frac{\partial f}{\partial x_1}(z)(x_1 - z_1) + \cdots + \frac{\partial f}{\partial x_n}(z)(x_n - z_n)$$

当 x 非常接近 z 时, $\hat{f}(x)$ 也非常接近 $f(z)$ 。 $\hat{f}(x)$ 是关于 x 的一个仿射函数。

Corollary 2.2.4 — 一阶泰勒公式的内积形式.

$$\hat{f}(x) = f(z) + \nabla f(z)^T (x - z) \quad \nabla f(z) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(z) \\ \vdots \\ \frac{\partial f}{\partial x_n}(z) \end{bmatrix}$$

一维时, $\hat{f}(x) = f(z) + f'(z)(x - z)$

■ **Example 2.4**

$$f(x) = x_1 - 3x_2 + e^{2x_1+x_2-1}$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \end{bmatrix} = \begin{bmatrix} 1 + 2e^{2x_1+x_2-1} \\ -3 + e^{2x_1+x_2-1} \end{bmatrix}$$

函数 f 在 0 点的一阶泰勒公式为:

$$\hat{f}(x) = f(0) + \nabla f(0)^T (x - 0) = e^{-1} + (1 + 2e^{-1})x_1 + (-3 + e^{-1})x_2$$

■

2.3 Regression Model

Definition 2.3.1 — **Regression Model.** 回归模型 (regression model) 为关于 x 的仿射函数

$$\hat{y} = x^T \beta + v$$

x 是特征向量 (*feature vector*), 它的元素 x_i 称为回归元 (*regressors*)。n 维向量 β 是权重向量 (*weight vector*)。标量 v 是偏移量 (*offset*)。标量 \hat{y} 是预测值 (*prediction*)。表示某个实际结果或因变量, 用 y 表示。

3. Norm and Distance

3.1 Norm

Definition 3.1.1 — Vector Norm. 在向量空间中存在一个函数 $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$, 且满足以下条件

- 齐次性: $\|\alpha x\| = |\alpha| \|x\|$, $\alpha \in \mathbb{R}$ 且 $x \in \mathbb{R}^n$;
- 三角不等式: $\|x + y\| \leq \|x\| + \|y\|$, $x, y \in \mathbb{R}^n$;
- 非负性: $\|x\| \geq 0$, $x \in \mathbb{R}^n$ 且 $\|x\| = 0 \Leftrightarrow x = 0$;

则称 $\|\cdot\|$ 为向量范数。

■ **Example 3.1 — ℓ_1 -范数 (曼哈顿范数, Manhattan norm) .**

$$\|x\|_1 = |x_1| + |x_2| + \dots + |x_n| \quad x, y \in \mathbb{R}^n, \alpha \in \mathbb{R}$$

Proof.

$$\|\alpha x\|_1 = |\alpha x_1| + |\alpha x_2| + \dots + |\alpha x_n| = |\alpha| \|x\|_1 \geq 0$$

$$\|x + y\|_1 = |x_1 + y_1| + \dots + |x_n + y_n| \leq |x_1| + |y_1| + \dots + |x_n| + |y_n| = \|x\|_1 + \|y\|_1$$

■ **Example 3.2 — ℓ_2 -范数 (欧几里得范数, Euclidean norm) .**

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{x^T x} = (\langle x, x \rangle)^{\frac{1}{2}}$$

Proof.

$$\|\alpha x\|_2 = (\langle \alpha x, \alpha x \rangle)^{\frac{1}{2}} = |\alpha| (\langle x, x \rangle)^{\frac{1}{2}} = |\alpha| \|x\|_2$$

$$\begin{aligned}
\|x + y\|_2^2 &= \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\
&= \|x\|_2^2 + 2\langle x, y \rangle + \|y\|_2^2 \leq \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 \\
&= (\|x\|_2 + \|y\|_2)^2
\end{aligned}$$

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$$

■

Corollary 3.1.1 — 柯西—施瓦茨不等式.

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle = \|x\|_2^2 \|y\|_2^2$$

Definition 3.1.2 — ℓ_∞ -范数.

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|, x \in \mathbb{R}^n$$

Proof.

$$\begin{aligned}
\max_{1 \leq i \leq n} |x_i| &\leq (|x_1|^p + \cdots + |x_i|^p + \cdots + |x_n|^p)^{1/p} \\
&\leq \left(n \max_{1 \leq i \leq n} |x_i|^p \right)^{1/p} \\
&= n^{1/p} \max_{1 \leq i \leq n} |x_i| \\
&\rightarrow \max_{1 \leq i \leq n} |x_i| \quad (p \rightarrow \infty)
\end{aligned}$$

■

Definition 3.1.3 — ℓ_p -范数.

$$\|x\|_p = \left(x_1^p + x_2^p + \cdots + x_n^p \right)^{\frac{1}{p}}, \quad x \in \mathbb{R}^n, p \geq 1$$

ℓ_1 范数 $\|x\|_1$, ℓ_2 -范数 $\|x\|_2$, ℓ_∞ -范数是 ℓ_p -范数的特例。

证明可以使用以下两条不等式

Theorem 3.1.2 — Minkowski Inequality.

$$\left(\sum_{i=1}^n |x_i + y_i|^p \right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |y_i|^p \right)^{\frac{1}{p}}, p \geq 1, x, y \in \mathbb{R}^n$$

Theorem 3.1.3 — Hölder Inequality.

$$\sum_{i=1}^n |x_i y_i| \leq \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} \left(\sum_{i=1}^n |y_i|^q \right)^{1/q}, \frac{1}{p} + \frac{1}{q} = 1, 1 < p, q < \infty$$

3.2 Root Mean Square Value (RMS)

Definition 3.2.1 — 向量 x 的均方值 (mean-square value). 向量 $x \in \mathbb{R}^n$ 的均方值 (mean-square value)

$$\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} = \frac{\|x\|_2^2}{n}$$

Definition 3.2.2 — n 维向量 x 的均方根 (root-mean-square value, RMS).

$$\text{rms}(x) = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \frac{\|x\|_2}{\sqrt{n}}$$

$\text{rms}(x)$ 给出了 $|x_i|$ 的“典型” (typical) 值。例如, $\text{rms}(\mathbf{1}) = 1$ (与 n 无关)。均方根 (RMS) 值对于比较不同长度的向量大小是比较有用的。

3.3 Chebyshev's Inequality

Theorem 3.3.1 — Chebyshev's Inequality.

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

$$P(|X - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}$$

Theorem 3.3.2 — Chebyshev's Inequality. 假设 k 为向量 x 分量满足条件 $|x_i| \geq a$ 的个数, 即 $x_i^2 \geq a^2$ 的个数。

因此: $\|x\|_2^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \geq ka^2$

将 a^2 移项, 可得到 $k \leq \frac{\|x\|_2^2}{a^2}$

满足 $|x_i| \geq a$ 的 x_i 数量不会超过 $\frac{\|x\|_2^2}{a^2}$

Corollary 3.3.3 — Chebyshev's Inequality Using RMS.

$$\text{rms}(x) = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \frac{\|x\|_2}{\sqrt{n}}$$

$|x_i| \geq a$ 的项数占整体的比例不会超过 $\left(\frac{\text{rms}(x)}{a}\right)^2$, 即 $\frac{k}{n} \leq \left(\frac{\text{rms}(x)}{a}\right)^2$

3.4 Distance

Definition 3.4.1 — Euclidean distance. n 维向量 a 和 b 之间的欧氏距离

$$\text{dist}(a, b) = \|a - b\|_2$$

Definition 3.4.2 — RMS deviation. $\text{rms}(a - b)$ 是 a 和 b 之间的均方根偏差。

Theorem 3.4.1 — Triangular Inequality.

$$\|a - c\|_2 = \|(a - b) + (b - c)\|_2 \leq \|a - b\|_2 + \|b - c\|_2$$

3.4.1 Feature Distance and Nearest Neighbor

Definition 3.4.3 — Feature Distance. 如果 x 和 y 分别为两个实体的特征向量, 那么它们的特征距离 (feature distance) 为 $\|x - y\|_2$

Definition 3.4.4 给定向量 x , 一个组向量 Z_1, \dots, Z_m , 当 \hat{q}_j 满足:

$$\|x - z_j\|_2 \leq \|x - z_i\|_2, \quad i = 1, \dots, m$$

则称 z_j 是 x 的最近邻 (nearest neighbor)

3.5 Standard Derivation

Definition 3.5.1 — 算术平均值. 对于 n 维向量 x

$$\text{avg}(x) = \frac{\mathbf{1}^T x}{n}$$

Definition 3.5.2 — De-meaned Vector.

$$\tilde{x} = x - \text{avg}(x)\mathbf{1}$$

因此 $\text{avg}(\tilde{x}) = 0$

Definition 3.5.3 — x 的标准差.

$$\text{std}(x) = \text{rms}(\tilde{x}) = \frac{\|x - (\mathbf{1}^T x / n) \mathbf{1}\|_2}{\sqrt{n}}$$

$\text{std}(x)$ 表示数据元素的变化程度。对于常数 α , 当且仅当 $x = \alpha\mathbf{1}$ 时, $\text{std}(x) = 0$.

Theorem 3.5.1

$$\text{rms}(x)^2 = \text{avg}(x)^2 + \text{std}(x)^2$$

3.6 Angle

Definition 3.6.1 — 两个非零向量 a 和 b 之间的角 (angle).

$$\angle(a, b) = \arccos\left(\frac{a^T b}{\|a\|_2 \|b\|_2}\right)$$

$\angle(a, b)$ 的取值范围为 $[0, \pi]$, 且满足

$$a^T b = \|a\|_2 \|b\|_2 \cos(\angle(a, b))$$

在二维和三维向量之中, 这里的角与普通角度 (ordinary angle) 是一致的。

- $\theta = \frac{\pi}{2} = 90^\circ$: a 和 b 为正交, 写作 $a \perp b$ ($a^T b = 0$)。
- $\theta = 0$: a 和 b 为同向的 ($a^T b = \|a\|_2 \|b\|_2$)。
- $\theta = \pi = 180^\circ$: a 和 b 为反向的 ($a^T b = -\|a\|_2 \|b\|_2$)。

- $\theta < \frac{\pi}{2} = 90^\circ$: a 和 b 成锐角 ($a^T b > 0$)。
- $\theta > \frac{\pi}{2} = 90^\circ$: a 和 b 成钝角 ($a^T b < 0$)。

Definition 3.6.2 — 球面的距离.

$$R\angle(a, b)$$

3.6.1 相关系数

给定向量 a 和 b ，其去均值向量为：

$$\tilde{a} = a - \text{avg}(a)1, \tilde{b} = b - \text{avg}(b)1$$

Definition 3.6.3 — a 和 b 的相关系数.

$$\rho = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\|_2 \|\tilde{b}\|_2} = \cos \angle(\tilde{a}, \tilde{b})$$

where $\tilde{a} \neq 0, \tilde{b} \neq 0$.

■ **Example 3.3** 高度相关的向量：

- 邻近地区的降雨时间序列。
- 类型密切相关文档的单词计数向量。
- 同行业中类似公司的日收益。

比较不相关的向量：

- 无关的向量。
- 音频信号 (比如，在多轨录音中的不同轨)。

负相关的向量：

- 深圳与墨尔本的每天气温变化

■

4. 以 k -Means 算法为例的优化问题

Problem 4.1 假设 N 个样本向量 $x_1, \dots, x_N \in \mathbb{R}^n$, 需要找到中心向量 z 满足

$$\min_{z \in \mathbb{R}^n} \sum_{i=1}^N \|x_i - z\|_2^2$$

Definition 4.0.1 — 高阶无穷小记号 o . 设 x, y 是同一变化过程中的无穷小, 即 $x \rightarrow 0, y \rightarrow 0$, 如果它们极限

$$\lim \frac{y}{x} = 0$$

则称 y 是 x 的高阶无穷小, 记作 $y = o(x)$.

Corollary 4.0.1

$$\lim \frac{y}{Cx} = \frac{1}{C} \lim \frac{y}{x} = 0$$

也即则称 y 是 Cx 的高阶无穷小, 记作 $y = o(Cx)$ 。

Proposition 4.0.2 — 优化求解的必要条件. 假设函数 f 在 \hat{x} 可微, 则有

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} f(x) \Rightarrow \nabla f(\hat{x}) = 0$$

Proof. 假设函数 f 在 \hat{x} 一阶泰勒展开, 有

$$f(x) = f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + o(\|x - \hat{x}\|_2)$$

假设 $\delta f(\hat{x}) \neq 0$, 则令 $\tilde{x} = \hat{x} - t \nabla f(\hat{x}), t > 0$, 可得

$$f(\tilde{x}) = f(\hat{x}) - t \|\nabla f(\hat{x})\|_2^2 + o(t \|\nabla f(\hat{x})\|_2)$$

当 $t \rightarrow 0$ 则 $t \|\nabla f(\hat{x})\|_2 \rightarrow 0$, 高阶无穷小 $o'(t \|\nabla f(\hat{x})\|_2) \rightarrow 0$

当 t 足够小时, 存在 $t \|\nabla f(\hat{x})\|_2 \geq o(t \|\nabla f(\hat{x})\|_2)$, 即

$$-t\|\nabla f(\hat{x})\|_2^2 + o(t\|\nabla f(\hat{x})\|_2) \leq 0$$

$$f(\tilde{x}) = f(\hat{x}) - t\|\nabla f(\hat{x})\|_2^2 + o(t\|\nabla f(\hat{x})\|_2) \leq f(\hat{x})$$

与 $\hat{x} = \arg \min_{\mathbf{R}^n} f(x)$ 矛盾。

$\nabla f(\hat{x}) = 0$, 是最优问题解的必要条件。通常 $\nabla f(\hat{x}) = 0 \not\Rightarrow \hat{x} = \arg \min_{\mathbf{R}^n} f(x)$ 。 ■

■ **Example 4.1**

$$f(x) = -x^2, \quad x \in \mathbf{R}, \hat{x} = \arg \min_{\mathbf{R}} f(x)$$

$\nabla f(\hat{x}) = 0$, 则有 $-2\hat{x} = 0$, 即 $\hat{x} = 0$

$$f(\hat{x}) = 0 \geq f(x), \quad x \in \mathbf{R}$$

(最大值!) ■

4.1 Convex Set

Definition 4.1.1 — 凸集. $\forall x, y \in \Omega, \alpha \in \mathbf{R}, 0 \leq \alpha \leq 1$ 有

$$\alpha x + (1 - \alpha)y \in \Omega$$

则定义域 $\Omega \in \mathbf{R}^n$ 称为凸的 (Convex) 集合
(域内两点连线之间都属于这个域)

Definition 4.1.2 — 凸函数. 设函数 $f(x)$ 定义于称为凸的定义域 $\Omega \in \mathbf{R}^n$ 满足

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y), \forall x, y \in \Omega, \alpha \in \mathbf{R}, 0 \leq \alpha \leq 1$$

称其为凸函数。

■ **Example 4.2**

$$f(x) = x^2, x \in \mathbf{R}$$

$$\begin{aligned} f(\alpha x + (1 - \alpha)y) &= (\alpha x + (1 - \alpha)y)^2 \\ &= \alpha^2 x^2 + 2\alpha(1 - \alpha)xy + (1 - \alpha)^2 y^2 \\ &= \alpha x^2 + (1 - \alpha)y^2 + (\alpha^2 - \alpha)x^2 + (\alpha^2 - \alpha)y^2 + 2\alpha(1 - \alpha)xy \\ &= \alpha x^2 + (1 - \alpha)y^2 - \alpha(1 - \alpha)(x - y)^2 \\ &\leq \alpha x^2 + (1 - \alpha)y^2 = \alpha f(x) + (1 - \alpha)f(y) \end{aligned}$$

■ **Example 4.3** $f(x) = \|x\|$, 其中 $\|\cdot\|$ 表示 \mathbf{R}^n 上的向量范数, $x \in \mathbf{R}^n$. ■

Proof.

$$\|\alpha x + (1 - \alpha)y\| \leq \|\alpha x\| + \|(1 - \alpha)y\| = |\alpha|\|x\| + |1 - \alpha|\|y\|$$

■ **Example 4.4**

$$f(x) = \|x\|_2^2, x \in \mathbf{R}^n$$

Theorem 4.1.1 — 可微函数 f 是凸函数的充要条件.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y$$

Proof. 首先, 证明一维情况 $f: \mathbb{R} \rightarrow \mathbb{R}, \alpha \in [0, 1]$.

\Rightarrow 充分条件: $f(\alpha x + (1 - \alpha)y) = f(x + (1 - \alpha)(y - x)) \leq \alpha f(x) + (1 - \alpha)f(y)$, 有

$$f(y) \geq f(x) + \frac{f(x + (1 - \alpha)(y - x)) - f(x)}{(1 - \alpha)(y - x)}(y - x)$$

令 $\alpha \rightarrow 1^-$, 则有 $f(y) \geq f(x) + f'(x)(y - x)$.

\Leftarrow 必要条件: 令 $y \neq x, z = \alpha x + (1 - \alpha)y$ 则有

$$f(x) \geq f(z) + f'(z)(x - z), f(y) \geq f(z) + f'(z)(y - z)$$

可得

$$\begin{aligned} \alpha f(x) + (1 - \alpha)f(y) &\geq f(z) + \alpha f'(z)(x - z) + (1 - \alpha)f'(z)(y - z) \\ &= f(z) + f'(z)(\alpha x + (1 - \alpha)y - z) \\ &= f(z) \end{aligned}$$

证明 n 维情况 $f: \mathbb{R}^n \rightarrow \mathbb{R}$.

\Rightarrow 充分条件: 令 $g(t) = f(tx + (1 - t)y), t \in \mathbb{R}$, 则 $g'(t) = \langle \nabla f(tx + (1 - t)y), x - y \rangle$ 由于 f 是凸函数, 证明 $g(t)$ 也是凸函数; 并可得 $g(0) \geq g(1) + g'(1)(-1)$, 得证.

\Leftarrow 必要条件: 与一维类似。 ■

Theorem 4.1.2 如果可微函数 f 是凸函数, 则有

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} f(x) \Leftrightarrow \nabla f(\hat{x}) = 0$$

Proof. 已证 $\hat{x} = \arg \min_{x \in \mathbb{R}^n} f(x) \Rightarrow$ 可得 $\nabla f(\hat{x}) = 0$

只需证 $\nabla f(\hat{x}) = 0 \Rightarrow \hat{x} = \arg \min_{x \in \mathbb{R}^n} f(x)$.

由于函数 f 是可微凸的, 则有 $\forall x \in \mathbb{R}^n$,

$$\begin{aligned} f(x) &\geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle \\ &\geq f(\hat{x}) + \langle 0, x - \hat{x} \rangle \geq f(\hat{x}) \end{aligned}$$

可得 $f(x) \geq f(\hat{x}), \hat{x} = \arg \min_{x \in \mathbb{R}^n} f(x)$. ■

4.2 向量偏导

Definition 4.2.1 — 向量对向量的导数.

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f(z)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z)}{\partial z_n} \end{bmatrix}$$

■ **Example 4.5**

$$f(z) = x^T z + z^T z = \sum_{i=1}^n \{x_i z_i + z_i^2\}$$

$$\nabla f(z) = \begin{bmatrix} \frac{\partial f(z)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z)}{\partial z_n} \end{bmatrix} = \begin{bmatrix} x_1 + 2z_1 \\ \vdots \\ x_n + 2z_n \end{bmatrix} = x + 2z$$

■

问题4.1中已知目标函数是凸函数。(见4.2, 4.3, 4.1)
则可以求解

$$f(z) = \sum_{i=1}^N \|x_i - z\|_2^2 = \sum_{i=1}^N \langle x_i - z, x_i - z \rangle = \sum_{i=1}^N \{x_i^T x_i - 2x_i^T z + z_i^T z\}$$

利用等价条件4.1.1

$$\nabla f(z) = \sum_{i=1}^N \{-2x_i + 2z\} = 0$$

(求导 4.2.1)

$$z = \frac{1}{N} \sum_{i=1}^N x_i$$

4.3 标量优化问题的例子

Problem 4.2 假设 $a, b \in \mathbb{R}^n, a \neq 0, t \in \mathbb{R}$, 当 t 多大时, ta 到 b 之间的距离最小

$$\min_t \|ta - b\|_2^2$$

$$f(t) = \|ta - b\|_2^2 = \langle ta - b, ta - b \rangle = t^2 a^T a - 2ta^T b + b^T b$$

$$\nabla f(t) = 2ta^T a - 2a^T b = 0$$

$$t = \frac{a^T b}{a^T a} = \frac{a^T b}{\|a\|_2^2}$$



Bibliography

Articles

Books

