Part III-B: Deep Learning

Lecture by None

Note by THF

2025 年 3 月 14 日

目录

Learn 1

01.28

1 线性代数复习

Definition. 张量：多维数组或多维立方体（二维为矩阵，一维为向量，零维为常量），用字体不同的粗体表示：$\mathsf{\mathbf{A}}$

对比：矩阵 $\boldsymbol{A}$，张量 $\mathsf{\mathbf{A}}$

Definition. 向量：一般指列向量，转置后为横向量

Example. $\boldsymbol{x} = (1, 2, 3)^\top = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$，转置后：$\boldsymbol{x}^\top = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$

Definition. 矩阵乘法：左矩阵的行遍历右矩阵的列

Example. 两个向量的点乘（内积）：$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{y}$，如下图：

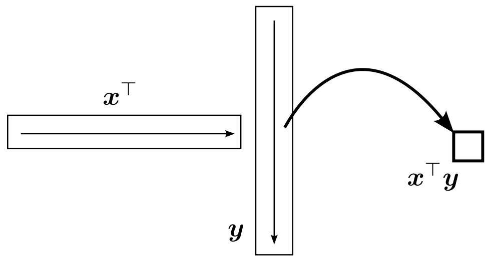

图 1: 点积

同理，两个向量的外积：$\boldsymbol{x} \times \boldsymbol{y} = \boldsymbol{x}\boldsymbol{y}^\top$ ，如下图：

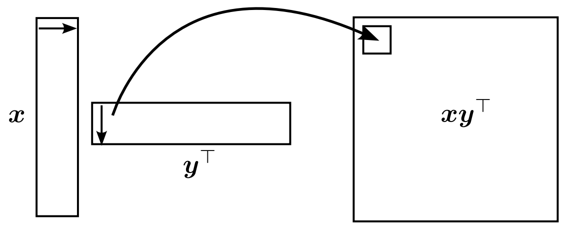

图 2: 外积

Definition. 广播：把向量转置后加到矩阵的每一行

Example.

$$\begin{array}{ccc} \boldsymbol{A} & +\boldsymbol{b} & =\boldsymbol{C} \\ \Rightarrow \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} & + \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}^\top & = \begin{bmatrix} a_{11}+b_1 & a_{12}+b_2 & a_{13}+b_3 \\ a_{21}+b_1 & a_{22}+b_2 & a_{23}+b_3 \end{bmatrix}. \end{array}$$

Definition. 范数：曼哈顿距离 $\Rightarrow$ 欧氏距离 $\Rightarrow \ldots$，L^p 范数形如：

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}.$$

Example. 曼哈顿距离：$L^1 : \|\boldsymbol{x}\|_1 = \sum_i |x_i|$ ，欧几里得距离：$L^2 : \|\boldsymbol{x}\|_2 = \sqrt{\sum_i x_i^2}$

Learn 2

01.30

Notation. L^2 范数可以写为 $\|\boldsymbol{x}\|_2$ ，也可以简写为 $\|\boldsymbol{x}\|$

Notation. L^1 范数常用于区分 0 值和距离 0 很近的值，使用 L^2 范数，距离太近几乎没有区别

Definition. 最大范数：$p \to \infty$ 时影响范数的最大因素为最大值：

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|.$$

Example. **Frobenius 范数：**使用范数衡量矩阵的大小，类似于 L^2 范数的计算：

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}.$$

Example. 使用范数表示向量的内积：

$$\boldsymbol{x}^\top \boldsymbol{y} = \|\boldsymbol{x}\| \cdot \|y\| \cos\theta.$$

即：

$$\boldsymbol{x} \cdot \boldsymbol{y} = xy \cos\theta.$$

1.1 特殊的矩阵和向量

Example. 对角矩阵，用 $\boldsymbol{D}$ 表示，除主对角线上的元素其他的元素均为零：

$$\boldsymbol{D} = \begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{bmatrix}.$$

Notation. 对角矩阵不一定是方阵，非方阵超出方阵的部分全部为 0

Definition. 对角转换 diag：$\text{diag}(\boldsymbol{v})$ 将 $\boldsymbol{v}$ 中的元素放入对角矩阵中

对角转换有以下性质：

- $\text{diag}(\boldsymbol{v})\boldsymbol{x} = \boldsymbol{v} \odot \boldsymbol{x}$
- $\text{diag}(\boldsymbol{v})^{-1} = \text{diag}(\begin{bmatrix} \frac{1}{v_1} & \dots & \frac{1}{v_n} \end{bmatrix}^\top)$

Example. 对称矩阵：$\boldsymbol{A} = \boldsymbol{A}^\top$ ，常用于某些不依赖参数顺序的双参数函数，交换 i, j 后结果不变

Example. 单位向量：范数为 1 的向量，即：$\|\boldsymbol{x}\|_n = 1$

Example. 向量的正交：$\boldsymbol{x}^\top \boldsymbol{y} = 0$ ，代表着 $\|x\|_n$ 和 $\|y\|_n$ 都不为零时 $\theta = \frac{\pi}{2}$
标准正交：$\theta = \frac{\pi}{2}$ 且 $\|\boldsymbol{x}\|_n = \|\boldsymbol{y}\|_n = 1$

Example. 正交矩阵：行向量和列向量都分别标准正交，有以下性质：

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I} \Rightarrow \boldsymbol{A}^{-1} = \boldsymbol{A}^\top.$$

可得：正交矩阵求逆非常容易，可以用来求解线性方程组

1.2 特征分解

Definition. 特征向量和特征值：

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}.$$

则 λ 为 $\boldsymbol{A}$ 的特征值，$\boldsymbol{x}$ 为 $\boldsymbol{A}$ 的一个（右）特征向量

$$\boldsymbol{x}^\top \boldsymbol{A} = \lambda \boldsymbol{x}^\top.$$

称 $\boldsymbol{x}$ 为左特征向量

Definition. 特征分解：如果 $\boldsymbol{A}$ 有 n 个线性无关的特征向量 $\{\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(n)}\}$ ，对应 n 个特征值 $\{\lambda_1, \dots, \lambda_n\}$ ，将特征向量（列向量）排列为一个矩阵 $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}^{(1)} & \dots & \boldsymbol{v}^{(n)} \end{bmatrix}$，并将特征值排列为一个向量 $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \dots & \lambda_n \end{bmatrix}^\top$，则：

$$\boldsymbol{A} = \boldsymbol{V} \text{diag}(\boldsymbol{\lambda}) \boldsymbol{V}^{-1}.$$

特征向量、值、分解的性质：

- $\boldsymbol{A}$ 对 $\boldsymbol{x}$ 变换相当于对 $\boldsymbol{x}$ 缩放 λ 倍
- $\boldsymbol{v}$ 是 $\boldsymbol{A}$ 的特征向量，则 $c\boldsymbol{v}$ 也是，且特征值一样

- 实对称矩阵一定可以分解为实特征向量和实特征值，即 $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$

Notation. 实对称矩阵的分解：

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top.$$

其中 $\boldsymbol{Q}$ 为 $\boldsymbol{A}$ 的特征向量 $\boldsymbol{v}^{(i)}$ 组成的**正交矩阵**，$\boldsymbol{\Lambda}$ 为对角矩阵，且 $\Lambda_{i,i} = \lambda_i$ 所对应的特征向量是 $\boldsymbol{Q}$ 的第 i 个列向量 $\boldsymbol{Q}_{:,i}$

使用 $\boldsymbol{A}$ 进行矩阵乘法可以看作是：将空间各自沿 $\boldsymbol{v}^{(i)}$ 延展 λ_i 倍

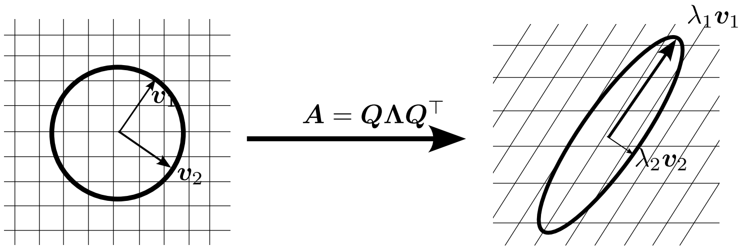

图 3: 实对称矩阵的乘法

Learn 3 01.31

Notation. $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ 表示二次型或者二次方程（$\boldsymbol{x}$ 为二维向量），如：

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \boldsymbol{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1 a + x_2 b & x_1 b + x_2 c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= a x_1^2 + 2b x_1 x_2 + c x_2^2. \end{aligned}$$

> **Definition.** 所有特征值都是非负数的矩阵为**半正定**矩阵，对于半正定矩阵有：
>
> $$\forall \boldsymbol{x}, \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} \geq 0.$$
>
> 对于正定矩阵（$A_{i,j} > 0$）有：
>
> $$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}.$$

Notation. 当 $\boldsymbol{x}$ 是 $\boldsymbol{A}$ 的某个特征向量 $\boldsymbol{v}$ 时，由于：

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}.$$

且 $\boldsymbol{x}=\boldsymbol{v}=\begin{bmatrix}x_1 & x_2\end{bmatrix}^\top$ ，二次型可以写为：$f=\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v}$，带入得：

$$\begin{aligned} f &= \boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} \\ &= \boldsymbol{v}^\top\lambda\boldsymbol{v} \\ &= \lambda\boldsymbol{v}^\top\boldsymbol{v} \\ &= \lambda\left(x_1^2+x_2^2\right). \end{aligned}$$

如果限定 $\|\boldsymbol{x}\|_2=1$ ，即 $x_1^2+x_2^2=1$ ，则 $f=\lambda$ ，即 $\max(f)=\max(\lambda),\min(f)=\min(\lambda)$

1.3 奇异值分解

除特征分解外，还有一种分解叫做奇异值分解，将一个实对称矩阵分解为**奇异向量**和**奇异值**

Notation. 特征分解：$\boldsymbol{A}=\boldsymbol{V}\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-1}$

Definition. 奇异值分解（SVD）：

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \quad \text{or} \quad \boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

Notation. 每个实数矩阵都有一个奇异值分解

Learn 4

02.01

Notation. 奇异值分解出来的部分：

- $\boldsymbol{U}$ ： $m\times m$ 的方阵，为正交矩阵，其列向量为**左奇异向量**
- $\boldsymbol{D}$ ： $m\times n$ 的矩阵，为对角矩阵，其对角线值为**奇异值**
- $\boldsymbol{V}$ ： $n\times n$ 的方阵，为正交矩阵，其列向量为**右奇异向量**

Learn 5

02.18

奇异值分解的各部分来源

假设任意 $m\times n$ 矩阵 $\boldsymbol{A}$

1. 求出两个对称矩阵

由于：$\left(\boldsymbol{A}^\top\boldsymbol{A}\right)^\top=\boldsymbol{A}\boldsymbol{A}^\top$，即这两个矩阵（**方阵**）都是对称的，令 $\boldsymbol{S}_L: m\times m=\boldsymbol{A}\boldsymbol{A}^\top, \boldsymbol{S}_R: n\times n=\boldsymbol{A}^\top\boldsymbol{A}$ ，提取出两个对称矩阵的特征向量组：

$$\boldsymbol{S}_L\Rightarrow\begin{bmatrix}\boldsymbol{u_1} & \boldsymbol{u_2} & \ldots\boldsymbol{u_m}\end{bmatrix}=\boldsymbol{U} \quad \boldsymbol{S}_R\Rightarrow\begin{bmatrix}\boldsymbol{v_1} & \boldsymbol{v_2} & \ldots\boldsymbol{v_n}\end{bmatrix}=\boldsymbol{V}.$$

如果求出 $\boldsymbol{S}_L$ 和 $\boldsymbol{S}_R$ 的特征值，并顺序排列，若 $m>n$ ，则会有：$\boldsymbol{S}_L$ 的前 n 个特征值与 $\boldsymbol{S}_R$ 的 n 个特征值一一对应，$\boldsymbol{S}_L$ 的剩余特征值都为 0

Notation. $\boldsymbol{A}$ 的**非零奇异值的平方**是 $\boldsymbol{A}^\top\boldsymbol{A}$ 或 $\boldsymbol{A}\boldsymbol{A}^\top$ 的特征值

对非 0 的特征值开根号即为奇异值，即：$\lambda_i = \sigma_i^2$

Notation. $\boldsymbol{\Sigma}$ 中的奇异值为非负值，且按照降序排列，即：

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0.$$

其中：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}.$$

1.4 Moore-Penrose 伪逆

Learn 6

03.14

简单来说，伪逆是扩展到非方阵的矩阵的逆，正常的逆：

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \implies \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}.$$

Definition. 伪逆的定义：

$$\boldsymbol{A}^+ = \lim_{\alpha \searrow 0} \left(\boldsymbol{A}^\top\boldsymbol{A} + \alpha\boldsymbol{I}\right)^{-1} \boldsymbol{A}^\top.$$

实际的计算方法：

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^\top \quad \text{or} \quad \boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top.$$

其中的矩阵通过 SVD 而得：

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top.$$

当 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ 中 $m < n$ 时最常用该方法求解线性方程组；当 $m > n$ 时方程组可能没有解，但是伪逆依然可以解出一个 $\boldsymbol{x}_0$ ，且这个解最贴合原方程，即：

$$\|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{y}\|_2 = \min \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 .$$

当 $m = n$ 时求解与使用正常逆求解一致。

1.5 迹

Notation. 迹即对角线求和：

$$\mathrm{Tr}\boldsymbol{A} = \sum_i A_{i,i}.$$

迹运算可以用来表示 Frobenius 范数。回顾 Frobenius 范数：

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}.$$

设 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，由于：

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^\top &= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & \ddots & & \\ \vdots & & \ddots & \\ A_{m1} & & & A_{mn} \end{bmatrix} \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & \ddots & & \\ \vdots & & \ddots & \\ A_{1n} & & & A_{mn} \end{bmatrix} \\
&= \begin{bmatrix} A_{11}^2 + A_{12}^2 + \ldots + A_{1n}^2 & & & \\ & A_{21}^2 + A_{22}^2 + \ldots + A_{2n}^2 & & \\ & & \ddots & \\ & & & A_{m1}^2 + A_{m2}^2 + \ldots + A_{mn}^2 \end{bmatrix} \\
&\implies \mathrm{Tr}\boldsymbol{A} = \sum_i \sum_j A_{ij}^2.
\end{aligned}$$

Notation. 迹运算的性质：**轮换对称**，即：

$$\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}\right) = \mathrm{Tr}\left(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}\right) = \mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}\right).$$

并且如果矩阵的形状改变，如 $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{B} \in \mathbb{R}^{n\times m}$，相乘后 $\boldsymbol{A}\boldsymbol{B} \in \mathbb{R}^{m\times m}$ 而 $\boldsymbol{B}\boldsymbol{A} \in \mathbb{R}^{n\times n}$，但是 $\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right) = \mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{A}\right)$ 仍然成立。

拓展到多个矩阵相乘：

$$\mathrm{Tr}\left(\prod_{i=1}^{n} \boldsymbol{F}^{(i)}\right) = \mathrm{Tr}\left(\boldsymbol{F}^{(n)} \prod_{i=1}^{n-1} \boldsymbol{F}^{(i)}\right).$$

Notation.

$$a = \mathrm{Tr}\left(a\right).$$

1.6 行列式

代表一个矩阵内列向量围成的空间大小，如：

$$\boldsymbol{A} \in \mathbb{R}^{2\times 2} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

代表二维空间中由 $\begin{bmatrix} A_{11} & A_{21} \end{bmatrix}^\top$ 和 $\begin{bmatrix} A_{12} & A_{22} \end{bmatrix}^\top$ 所围成的空间大小（所围成的平行四边形大小）

Notation. 如果行列式为 0, 说明该线性空间至少朝一个维度坍缩了（3 维 → 2 维 ~ 0 维）

1.7 主成分分析

假设有一组二维数据，这些数据有一个比较明显的走向，因此我们希望将这个明显的走向体现出来，而忽略其他的部分来实现数据的压缩。通过找到数据中心并旋转坐标系来拟合一个损失最小的坍缩方向。

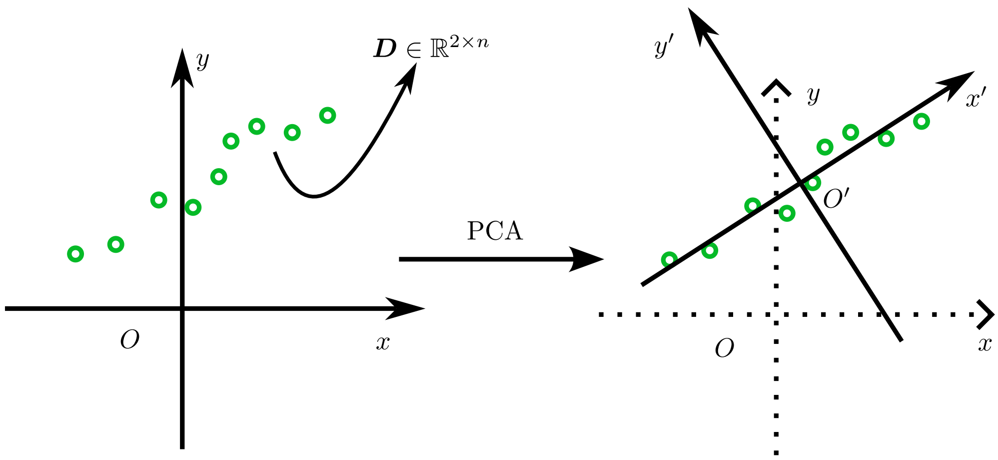

图 4: 主成分分析图示

其中，x' 为数据的**主成分**