无约束的矩阵低秩近似
伪逆与SVD奇异值分解
LoRA微调和MLA的数学基础

东川路第一可爱猫猫虫

主要内容

- 矩阵的范数
- Frobenius范数
- 伪逆（广义逆）
- 最优化求伪逆
- SVD分解求低秩近似
- Eckart-Young-Mirsky 定理

矩阵的范数

- 范数是向量或矩阵的一种度量

  范数是将矩阵映射为非负实数的函数

  若$N(A) = ||A||$满足正定性，齐次性，三角不等式，则称$N(A) = ||A||$是矩阵A的**范数**

- Frobenius范数

$$||M||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} M_{i,j}^2}$$

  把矩阵的所有元素的平方加起来，开根号

  简称F范数

正交变换不改变 F 范数

- $\left||A|\right|_F^2 = tr(A^TA)$

  proof：A^TA展开，其主对角线上元素为第i行元素平方和，再对主对角线上元素求和即得到$\left||A|\right|_F^2$

- 对任意矩阵 $(A \in R^{m\times n})$，左乘或右乘正交阵均称为正交变换
- 以左乘正交阵为例来证明

  $$\left||QA|\right|_F^2 = tr[(QA)^T(QA)] = tr[A^TQ^TQA] = tr[A^TA] = \left||A|\right|_F^2$$

  右乘正交阵的情况同理可证

广义逆（伪逆）

- 逆矩阵

  若矩阵AB=M，且A为可逆的方阵，则有B=$A^{-1}M$称为A的逆

  那如果A不可逆呢？甚至A不是方阵呢？

- 广义逆

  1920年Moore提出广义逆

  1954年Penrose独立地给出了更完备的定义

$$AXA = A,$$
$$XAX = X,$$
$$(AX)^* = AX,$$
$$(XA)^* = XA,$$

这四个等式存在唯一解，这个解X就是矩阵A的伪逆

《A GENERALIZED INVERSE FOR MATRICES》

另一种方式求伪逆

- 在已知矩阵A和M的情况下，求出一个B使得AB与M间误差最小

  A的伪逆=$argmin_B \left|\left|AB-M\right|\right|_F^2$

  其中A∈Rn×r,B∈Rr×m,M∈Rn×m

  上式的运算结果是一个使得$\mathrm{L}=\left|\left|AB-M\right|\right|_F^2$最小的矩阵B

  为使$\mathrm{L}=\left|\left|AB-M\right|\right|_F^2$最小，求L对矩阵B的导数

- 标量函数f的矩阵X的**导数**

  $$\frac{\partial f}{\partial X}=\left[\frac{\partial f}{\partial x_{i,j}}\right]_{m\times n}$$

求$\mathrm{L} = \left|\left|AB - M\right|\right|_F^2$对矩阵B的导数

- 链式法则

令E=AB-M

$$\frac{\partial L}{\partial B_{i,j}} = \sum_{k,l} \frac{\partial L}{E_{k,l}} \frac{\partial E_{k,l}}{\partial B_{i,j}}$$

这里$E_{k,l}$代表矩阵E在(k,l)位置的元素，$B_{i,j}$代表矩阵B在(ki,j)位置的元素

接下来分别求$\frac{\partial L}{E_{k,l}}$和$\frac{\partial E_{k,l}}{\partial B_{i,j}}$

求第一部分 $\frac{\partial L}{E_{k,l}}$

- 由$\mathrm{L} = \left|\left|AB - M\right|\right|_F^2$以及F范数的定义

  $\mathrm{L} = \left|\left|\mathrm{E}\right|\right|_F^2 = \sum_{i,j} E_{i,j}^2$

  多元函数对一个元求偏导，只有(I,j)=(k,l)时，$\frac{\partial L}{E_{k,l}}$才不为0

  也就是说，L关于$E_{k,l}$的导数就是$E_{k,l}^2$关于$E_{k,l}$的导数

  $\frac{\partial L}{\partial E_{k,l}} = 2E_{k,l}$ ，$\frac{\partial L}{\partial E} = 2E$

求第二部分 $\frac{\partial E_{k,l}}{\partial B_{i,j}}$

- E=AB-M 而M是常数矩阵，与B无关 因此 $\frac{\partial E}{\partial B}=\frac{\partial AB}{\partial B}$
- $A_{n\times r}B_{r\times m}$在(k,l)位置上的元素为？

  这个元素实际上是A的第k行向量与B的第l列向量做内积

  $\left(A_{k,1},\dots A_{k,r}\right)\left(B_{1,l},\dots,B_{r,l}\right)=\sum_m A_{k,m}\,B_{m,l}$
- 由此E在(k,l)位置的元素为

  $\sum_m A_{k,m}\,B_{m,l}-M_{k,l}$

  若$\mathrm{l}\neq \mathrm{j}$，则$\frac{\partial E_{k,l}}{\partial B_{i,j}}=0$；若$\mathrm{l}=\mathrm{j}$，则$\frac{\partial E_{k,l}}{\partial B_{i,j}}=A_{k,i}$

  $\frac{\partial E_{k,l}}{\partial B_{i,j}}=A_{k,\mathrm{i}}\delta_{l,j}$　其中$\delta_{l,j}$用于判定l与j是否相等

求$\frac{\partial L}{\partial B_{i,j}} = \sum_{k,l} \frac{\partial L}{E_{k,l}} \frac{\partial E_{k,l}}{\partial B_{i,j}}$

- $\frac{\partial L}{\partial B_{i,j}} = 2\sum_{k,l} E_{k,l} A_{k,i} \delta_{l,j} = 2\sum_{k} E_{k,j} A_{k,i}$
- 注意到$\sum_{k} E_{k,j} A_{k,i}$是$A^T E$在(i,j)的元素

  因此有$\frac{\partial L}{\partial B_{i,j}} = 2(A^T E)_{i,j}$

  $\frac{\partial L}{\partial B} = 2(A^T E) = 2A^T(AB - M)$

  同理，$\frac{\partial L}{\partial \mathrm{A}} = 2(AB - M)\mathrm{B}^T$

要求min，令偏导=0

- $\frac{\partial L}{\partial B} = 2A^T(AB - M) = 0$ ， $\frac{\partial L}{\partial \mathrm{A}} = 2(AB - M)\mathrm{B}^T$

  有$A^T AB = A^T M$ ， $ABB^T = MB^T$

  只要$A^T A$可逆，我们就求出了B=$(A^T A)^{-1} A^T M$

  这就是我们求得的$argmin_B \left|\left|AB - M\right|\right|_F^2$，A的伪逆

  以上就是给定矩阵M和A时，优化$\left|\left|AB - M\right|\right|_F^2$的最优解

A和B均未知的话，该怎么办

- 刚才我们在已知矩阵A和M的情况下，求出了$argmin_B \left|\left|AB - M\right|\right|_F^2$
- 现在考虑A、B未知时候的最优解

  即求$argmin_{A,B} \left|\left|AB - M\right|\right|_F^2$

  其中，A∈Rn×r,B∈Rr×m,M∈Rn×m,r<min(n,m)
- SVD奇异值分解

  奇异值分解在实矩阵域的定理：

  对于任意矩阵M∈ $R^{n\times m}$，都能找到M=$U\Sigma V^T$

  其中U、V为$n \times n$、$m \times m$正交阵，Σ是非负对角阵

  对角线元素称为奇异值，默认从大到小排序

用SVD求解$argmin_{\mathrm{A},B}\left||AB-M|\right|_F^2$

- 将M分解为$U\Sigma V^T$

$$\text{则}\left||AB-M|\right|_F^2=\left||UU^TABVV^T-U\Sigma V^T|\right|_F^2$$
$$=\left||U(U^TABV-\Sigma)V^T|\right|_F^2$$
$$=\left||U^TABV-\Sigma|\right|_F^2$$

问题被化简了

- 低秩近似

  低秩近似的目标是求“M的最优r秩近似”

  即求秩不超过r的矩阵X，使得$\left|\left|X-M\right|\right|_F^2$最小

  矩阵的秩分解性质：秩为r的矩阵可分解为 “$m \times r$列满秩矩阵” 与 “$r \times n$行满秩矩阵” 的乘积

  因此问题可以转化为$argmin_{\mathrm{A},B}\left|\left|AB-M\right|\right|_F^2$

  由刚才的SVD分解，我们进一步把问题转化为了：

$$argmin_{\mathrm{A},B}\left|\left|U^TABV-\Sigma\right|\right|_F^2$$

问题的进一步化简

- 令$Y = U^T ABV$
- 我们注意到

  Y的秩与AB相同（正交变换不改变矩阵的秩）⩽r

  任意秩不超过r的矩阵Y，都能找到A和B，这是因为U和V都是可逆的，AB=UYV^T，而UYV^T秩⩽r，又可以分解为A和B
- 由此，问题被简化成了**非负对角阵Σ的最优r秩近似**

非负对角阵Σ的最优r秩近似

- 我们要求出一个秩≤r的矩阵Σ_r，以最小化$\left|\left|\Sigma_r - \Sigma\right|\right|_F^2$

  其中Σ为已知的非负对角阵
- 当然我们考虑的是低秩近似，如果Σ秩≤r，那就取$\Sigma_r = \Sigma$，这样$\left|\left|\Sigma_r - \Sigma\right|\right|_F^2$直接为0了，结论很显然，没有讨论价值
- 所以我们考虑Σ秩 > r的情况

  为达到目标，Σ_r的非对角线元素均需为0，消除非对角线误差

  对角线上，保留前 r 个最大的奇异值，其余为0
- 结论

  非负对角阵的最优r秩近似就是只保留对角线最大的r个元素的矩阵

基于 SVD 的Eckart-Young-Mirsky 定理

- 如果M∈ $R^{m\times n}$的SVD奇异值分解为$U\Sigma V^T$，那么M的最优r秩近似为$U_{[:n,:r]}\Sigma_{[:r,:r]}V^T_{[:m,:r]}$

  $U_{[:n,:r]}$ 取正交矩阵U的前 r 列（保留前 r 个 “左特征方向”）

  $\Sigma_{[:r,:r]}$取对角阵Σ的前 r 阶对角子阵,保留前 r 个最大的奇异值

  $V^T_{[:m,:r]}$取正交矩阵V^T的前 r 列（保留前 r 个 “右特征方向”）

- 普遍意义在于:

  任何矩阵（无论是否对角、是否方阵）的最优 r 秩近似，都可以通过其 SVD 分解，“截取前 r 个最大奇异值及对应方向” 得到