Deep Learning 读书会第四次讨论记录（由@极视角小助手整理）

下面为 2016 年 12 月 4 日 Deep Learnin 读书会第四章 **Numerical Computation 数值优化**讨论交流笔记。由极视角小助手整理，**九问**和 **AG-GROUP 元芳**提出话题组织讨论。如有想加入读书会讨论的，请联系小助手（微信：Extreme-Vision）.

讨论话题

引言

第四章，正如章节的题目，数值计算，讲了一些计算机处理数据的技巧（计算机数学与理论数学的区别），讲解了数学中的数据敏感度等对于深度学习有很大影响的数学理论。同时提出了梯度下降法（BP 算法的核心）、牛顿法等优化算法，比较了两种算法的优劣，同时给出了最优化问题可以求取最优值的必要条件 KKT 条件。
（总结来自 AG-GROUP 元芳）

话题一. 为什么会出现 underflow 或 overflow? （AG-GROUP 元芳）

张小彬@Bruce

上溢是不是因为数值太大，比如 softmax 层要算自然指数就会很容易出现. 下溢就没遇到过了

九问

对，上溢是因为分母是接近 0？

$$\mathrm{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n}\exp(x_j)}.$$

枫

下溢分母接近 0 吧？

张小彬@Bruce

如果激活值 x 很大，分子分母都是 inf，那么结果就是 NaN

九问

当 exp(xj)中 xj 是负的非常大的时候，分母接近 0.所以容易溢出。书中提出了一个解决方案

can be resolved by instead evaluating $\mathrm{softmax}(\boldsymbol{z})$ where $\boldsymbol{z} = \boldsymbol{x} - \max_i x_i$.

用 Z 来代替原来的 x

枫

这个里面 x 减去的是什么？

$\max_i x_i$.

九问

减去最大的，那么肯定有一项为 1

减去的是最大的 xj 的值？
我理解的是这样，大家什么看法

张小彬@Bruce
恩，cs231n 也提到过，相当于分子分母同时约去一个公因子

九问
那么 underflow 怎么理解呢？

张小彬@Bruce
补充刚刚的 softmax 那个

$$\frac{e^{f_{y_i}}}{\sum_j e^{f_j}} = \frac{Ce^{f_{y_i}}}{C\sum_j e^{f_j}} = \frac{e^{f_{y_i}+\log C}}{\sum_j e^{f_j+\log C}}$$

九问
这个是哪里的呢？

张小彬@Bruce
cs231n

权
cs231n 吧

张小彬@Bruce

Practical issues: Numeric stability. When you're writing code for computing the Softmax function in practice, the intermediate terms $e^{f_{y_i}}$ and $\sum_j e^{f_j}$ may be very large due to the exponentials. Dividing large numbers can be numerically unstable, so it is important to use a normalization trick. Notice that if we multiply the top and bottom of the fraction by a constant C and push it into the sum, we get the following (mathematically equivalent) expression:

$$\frac{e^{f_{y_i}}}{\sum_j e^{f_j}} = \frac{Ce^{f_{y_i}}}{C\sum_j e^{f_j}} = \frac{e^{f_{y_i}+\log C}}{\sum_j e^{f_j+\log C}}$$

We are free to choose the value of C. This will not change any of the results, but we can use this value to improve the numerical stability of the computation. A common choice for C is to set $\log C = -\max_j f_j$. This simply states that we should shift the values inside the vector f so that the highest value is zero. In code:

```
f = np.array([123, 456, 789]) # example with 3 classes and each having large scores
p = np.exp(f) / np.sum(np.exp(f)) # Bad: Numeric problem, potential blowup

# instead: first shift the values of f so that the highest number is 0:
f -= np.max(f) # f becomes [-666, -333, 0]
p = np.exp(f) / np.sum(np.exp(f)) # safe to do, gives the correct answer
```

枫

这个好像分母还是 0

九问

能解释下嘛@张小彬@Bruce，加了个常数 logC

张小彬@Bruce

logC 就是你说的那个减去一个最大值，用 z 代替

分子分母都约去了 logC

九问

嗯，懂了

枫

这个 C 怎么选？

九问

这个很关键

张小彬@Bruce

代码的第三行， f -= np.max(f) 注意这里是向量的表达方式

九问
这个 C 就是 np.max(f)么

张小彬@Bruce
是 logC

人工智障 v1.04
logC 吧

九问
OK

人工智障 v1.04
它为什么要引入 log 来说明

九问
因为 softmax 分母就是指数形式呀

张小彬@Bruce
都可以吧，中间那个看公式
那 underflow 呢，我好像没遇到过，求解释？

九问
这个我也没遇到过。。有大神来解释下嘛

人工智障 v1.04
可能问题错了吧

张小彬@Bruce
就 softmax 而言，按理说结果都是正的才是啊，哦，是不是因为负无穷大，求指数以后就下溢了

九问
书中明确提出了 underflow 呀，只是没有举例子，很难理解
那个 underflow 是指，分子接近 0 吧

张小彬@Bruce

编程语言里 underflow 这个术语，指的是下溢到负无穷小，还是无穷接近零呢？还是说都可以？

清

我也觉得是这样

九问

接近 0

权

不一定是分子，对于一个变量，如果它很接近 0，就有可能被 rounded to zero，这种就是 underflow

九问

哦

清

We have an issue if our x’ s are very negative, because by exponentiating them we push them very close to zero, possibly resulting in underflow.

For instance, for a softmax:

$$softmax(x_i) = \frac{exp(x_i)}{\sum_{j=1}^{n} exp(x_j)}$$

We have an issue if our x’s are very negative, because by exponentiating them we push them very close to zero, possibly resulting in underflow. Similarly, with huge x, we might get overflow. We can do a bit of **normalization** to overcome this: by subtracting max(x), we don’t change the computation, but we ensure that in the numerator everything is <0, and in the denominator, at least one of the entries is exp(0)=1, which means we’re not going to get underflow!

人工智障 v1.04

可以让 x 加一个很小的值来解决吗

九问
类似那个 logC 就可以解决

枫
具体有什么解决办法？

九问
加上那个 logC 就可以，这个是一个常数，使得它不接近无穷小？

枫
underflow 是减一个最大的，overflow 加一个最小的？

人工智障 v1.04
用那个方法是两个问题都解决了吧

张小彬@Bruce
所以 softmax 下溢，就会输出 0/0，上溢就会输出 Inf/inf？

枫
关键是这个 C 从哪来？

九问
对，那个方法两个问题就解决了
logc=np.max(f)

枫
本来已经够小了 在减一个不是更小?

张小彬@Bruce
可以可以，很强大。
减去一个负数

话题二. condition number 的含义&condition 的作用（AG-GROUP 元芳&九问）

九问

条件数是衡量输入参数的微小变化对输出值的影响。我直接引用的 wiki 上的解释。举个例子吧，对于线性方程 Ax=b，如果条件数很大，那么 b 的小误差会导致 x 产生大误差，相反，如果条件数很小，那么 x 的误差相对于 b 的误差较小。这个条件数在衡量算法鲁棒性的时候非常有用。

清
解释的很棒

曲晓峰
pca 就是认为构造大条件数
人为构造

九问
比如线性回归中，我们需要需要学习一个鲁棒的算法使得学习参数，能够针对输入的数据有轻微的变化的时候，得到的输出近似一样。
因此条件数越小，说明学习到的参数泛化能力更强

张小彬@Bruce
哦，那个求解线性方程组的“病态矩阵问题” ill-condition matrix，是不是说的就是这个

九问
对，就是这个问题。也叫作 well-conditioned，反之如果是 ill-conditioned

张小彬@Bruce
condition number 是个标量数值是吧，是可以有公式计算的吗？

九问
有的

人工智障 v1.04
范数乘以逆的范数

张小彬@Bruce
矩阵范数？

九问

$$\max_{i,j}\left|\frac{\lambda_i}{\lambda_j}\right|.$$

指矩阵的特征值。λ_i

张小彬@Bruce

这里的 i, j 指的是什么？

权

不同的特征值的下标

九问

第几个特征值，一般矩阵有一个特征向量

曲晓峰

任意，然后找一个最大比率

夏天落大雨

遍历找最大的吗

张小彬@Bruce

是不是最大的特征值，除以最小的特征值

权

两两求最大

九问

对的

权
不一定是最大除以最小

张小彬@Bruce
加绝对值？

权
最后要取绝对值

张小彬@Bruce
绝对值最大的特征值，除以绝对值最小的特征值

枫
要是特征值只有一个不为 0 其他是 0 怎么办？

风牛也马
是的，condition number 一般在矩阵里被定义做最大 singular value 和最小 singular value 的比值

张小彬@Bruce
哦，奇异值，不是特征值。书里写的是特征值 eigenvalue。

话题三. 海森矩阵为什么可以帮助梯度下降法跨越鞍点（AG-GROUP 元芳）

九问
海森矩阵为什么可以帮助梯度下降法跨越鞍点，这个问题大家有谁能够解惑么

张小彬@Bruce
啥叫鞍点？

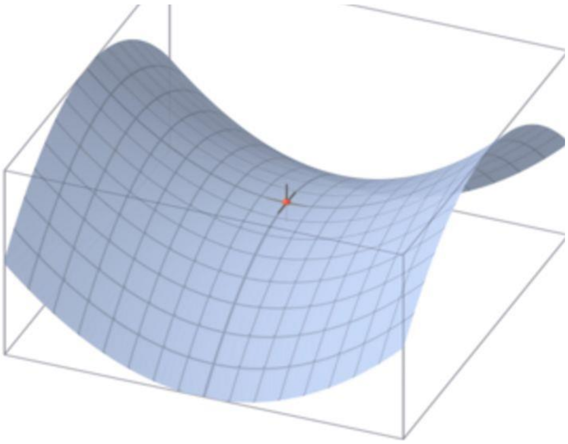
风牛也马

同问

WWN
一个方向稳定一个方向不稳定的奇点

九问
沿着某一方向是稳定的，另一条方向是不稳定的奇点，叫做鞍点，wiki 这么解释

清

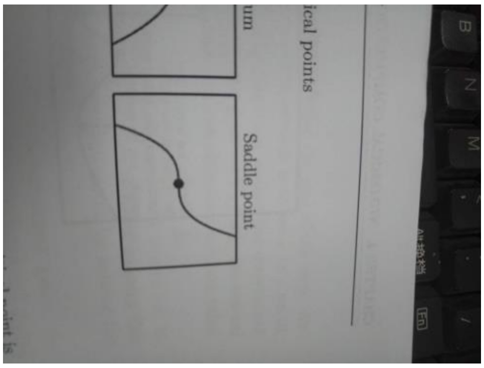


九问
有什么物理意义呢？

张小彬@Bruce
不是极大值，也不是极小值，但是导数为零？

清
对

WWN

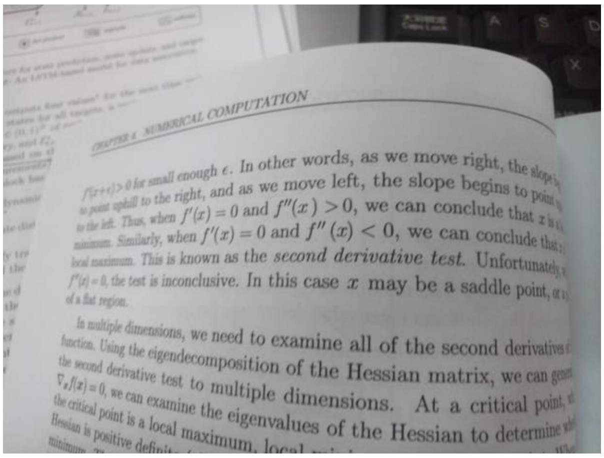


二微上就这个，好理解

张小彬@Bruce

那二阶导（海森矩阵）呢？有什么性质？

WWN



CHAPTER 4. NUMERICAL COMPUTATION

$f'(x+\epsilon) > 0$ for small enough ϵ. In other words, as we move right, the slope ... to point uphill to the right, and as we move left, the slope begins to point ... to the left. Thus, when $f'(x) = 0$ and $f''(x) > 0$, we can conclude that x is a ... minimum. Similarly, when $f'(x) = 0$ and $f''(x) < 0$, we can conclude that ... local maximum. This is known as the *second derivative test*. Unfortunately, ... $f''(x) = 0$, the test is inconclusive. In this case x may be a saddle point, or ... of a flat region.

In multiple dimensions, we need to examine all of the second derivatives ... function. Using the eigendecomposition of the Hessian matrix, we can gene... the second derivative test to multiple dimensions. At a critical point, ... $\nabla_x f(x) = 0$, we can examine the eigenvalues of the Hessian to determine ... the critical point is a local maximum, local ... Hessian is positive definit... minimum ...

清

鞍点附近 Hessian 矩阵有正的和负的特征值，是不定的

WWN

如果二阶导为零的话，就不一定是极大值或极小值喽

夏天落大雨

海森矩阵正定时局部最小值　负定时最大值

清

对 如果 Hessian 矩阵半正定 就可以

一定要正定吗

九问
海森矩阵正定代表什么意义呢

安兴乐
应该是一定要正定化的

夏天落大雨
半正定就行

清
一个$n \times n$的实对称矩阵M是**正定**的，当且仅当对于所有的非零实系数向量z，都有$z^T M z > 0$。其中z^T表示z的转置。

张小彬@Bruce
为什么要求要正定的？

夏天落大雨
正定就是特征值全部大于零

清
具体找一本最优化的书看一下

张小彬@Bruce
这个是定义

枫
函数的凹凸性 是不是就是可以通过 Hessian 矩阵判断？

九问
当海森矩阵正定时，代表着这个点是局部最小

张小彬@Bruce
海森矩阵正定，说明函数有全局最优解。否则可能存在极大值和极小值。

九问

特征值全是负的呢？

yc
我理解是这样的，如果是半正定，就说明可能有一维的特征值为 0，就是说这个维度上可能是平面或者 saddle point.所以要正定

张小彬@Bruce
“矩阵 A 是半正定矩阵” 等价于 “特征值全是非负实数”。正定的话，特征值全部大于零

清
鞍点附近 Hessian 矩阵不是负定，是不定的

yc
比如 y=x^2，沿着 z 轴拉开，在 H 在(0,0,0) 就是半正定的？（不知道说的对不对哦）

清
最好找个二维的吧。不然不好算 Hessian 矩阵

张小彬@Bruce
算二阶导也一样的 ,海森矩阵只是从函数 f(x) 到多元函数 f(x1,x2,...,xn) 的推广而已

话题四. 公式 4.21 怎么对 X 求导的，矩阵求导法则？（九问）

九问

Suppose we want to find the value of $\boldsymbol{x}$ that minimizes

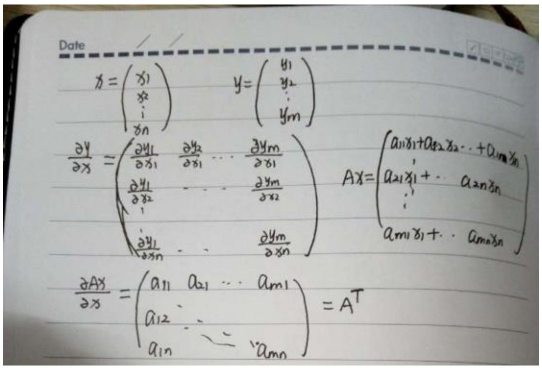
$$f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2^2. \tag{4.21}$$

There are specialized linear algebra algorithms that can solve this problem efficiently. However, we can also explore how to solve it using gradient-based optimization as a simple example of how these techniques work.

First, we need to obtain the gradient:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top\boldsymbol{b}. \tag{4.22}$$

九问
这个公式，我推了下，和书上结果不一样，我对矩阵求导不太熟，大家可以试试

夏天落大雨
Ax 对 x 求导是向量对向量求导 求导后就变成了 A 的转置


九问
上面是 L2 范式，不应该是 1/2(AX-b)(AX-b)T 对 x 求导么

夏天落大雨
1/2 消掉了

九问
我是通过这个公式求导，算出来结果不一样
@夏天落大雨 能否讲讲矩阵求导？
我把公式转成 1/2(AX-b)T(AX-b)这个，然后分别求导呀

夏天落大雨
1/2 和 2 范数求导后消掉剩下 Ax-b 再乘以复合函数 Ax 求导
不就是 A 的转置乘以 Ax-b

九问
他是个 L2 范数，能这么求么？

张小彬@Bruce

$$\frac{\partial(\mathbf{b}^T\mathbf{a})}{\partial\mathbf{a}} = \mathbf{b}$$

$$\frac{\partial(\mathbf{a}^T\mathbf{A}\mathbf{a})}{\partial\mathbf{a}} = (\mathbf{A}+\mathbf{A}^T)\mathbf{a}$$

$$\frac{\partial}{\partial\mathbf{A}}\mathrm{tr}(\mathbf{B}\mathbf{A}) = \mathbf{B}^T$$

$$\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = \mathbf{A}^{-T} \triangleq (\mathbf{A}^{-1})^T$$

$$\mathrm{tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \mathrm{tr}(\mathbf{C}\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{C}\mathbf{A})$$

话题五．一阶优化算法与二阶优化算法的优劣对比（AG-GROUP 元芳）

九问

优化算法这块大家有什么想法呢？

张小彬@Bruce

一阶指的是梯度下降，二阶指的是牛顿法吗？

九问

对，我对牛顿法不太熟，只对 GD 比较熟

张小彬@Bruce

没用过牛顿法，感觉神经网络都是直接上 sgd 啊，有没有大神用过牛顿法

九问

是的，基本上都是 SGD

清

二阶优化算法得算海森矩阵，参数太多几乎不可能

九问

速度快，容易收敛

张小彬@Bruce
不是有拟牛顿法吗？

九问
拟牛顿法是怎么样的

清
自己搞一个近似的海森矩阵

张小彬@Bruce
什么 BFGS ，就是海森矩阵的逆矩阵太难求了，就用一个矩阵来迭代逼近它

清
其实除了计算复杂 存储也是个问题

九问
懂了

张小彬@Bruce

8.3.5 Quasi-Newton (variable metric) methods 拟牛顿法

所有二阶优化算法都是牛顿法，但是牛顿法的复杂度太高，因为计算海森矩阵的逆矩阵很麻烦，每次迭代都要计算一次。拟牛顿法的思想是用一个矩阵逐渐逼近海森矩阵和其逆矩阵，最常见的是 BFGS 算法，用一个矩阵 $\mathbf{B}_k \approx \mathbf{H}_k$ 来逼近，

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k} - \frac{(\mathbf{B}_k \mathbf{s}_k)(\mathbf{B}_k \mathbf{s}_k)^T}{\mathbf{s}_k^T \mathbf{B}_k \mathbf{s}_k}, \quad \text{其中，} \mathbf{s}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}, \mathbf{y}_k = \mathbf{g}_k - \mathbf{g}_{k-1}$$

矩阵 $\mathbf{B}$ 可以从 $\mathbf{B}_0 = \mathbf{I}$ 单位矩阵开始，BFGS 算法可以看做是向海森矩阵的对角低秩逼近（diagonal plus low-rank approximation）。

同样的，逆矩阵也可以通过类似的方式逼近 $\mathbf{C}_k \approx \mathbf{H}_k^{-1}$，

$$\mathbf{C}_{k+1} = \left(\mathbf{I} - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) \mathbf{C}_k \left(\mathbf{I} - \frac{\mathbf{y}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}\right) + \frac{\mathbf{s}_k \mathbf{s}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}$$

有时候数据很大，存储海森矩阵也很耗空间，需要 $O(D^2)$ 的空间复杂度，可以用**limited memory BFGS**, or **L-BFGS**，因为 $\mathbf{H}_k^{-1}\mathbf{g}_k$ 可以用 m 对最近的 $(\mathbf{s}_k, \mathbf{y}_k)$ 的内积，存储化简为 $O(mD)$ 的复杂度。

AG-GROUP 元芳

海森矩阵太占用资源了
而且要注意一点，牛顿法是一步到位的，可是书上说了，海森矩阵指示的是最陡峭的方向，并不一定是最优解方向

张小彬@Bruce
优化非凸的目标函数就会这样

话题六. KKT 条件什么时候是最优化问题可以求取最优值的充要条件？（元芳）

AG-GROUP 元芳
这个问题其实答案是凸优化时候是充要，可是为啥我就不知道了

清
这个我记得是最优化里面给几个限制条件 然后用拉格朗日乘子法的时候用的
深度学习在什么情况下会用到啊

张小彬@Bruce
KKT 是解决带约束的最优化问题，不知道什么时候会带约束？SVM 里好像会有用到

枫
我怎么觉得像是把不等式约束问题转化成等式约束问题求解的条件，也就是可以用 lagrangian 求解的条件？

九问
只在 SVM 里面见到过。

AG-GROUP 元芳
其实书上就是说如何判断最优化问题是否有最优解
判断 deep learning 是否能 work？

夏天落大雨
约束条件是等式时直接用拉格朗日乘子，好像如果还有不等式时就要用到 kkt 了

张小彬@Bruce
约束条件可以是等式约束，也可以是不等式约束
可以参考李航《统计学习方法》的附录 C

枫
好像只有不等式约束才出现 KKT 条件

曲晓峰
最优化里面，梯度的阶越高收敛速度越快。二阶要明显好于一阶。只不过求导会很难，甚至是不存在。

张小彬@Bruce
KKT 条件的结果就是目标函数对变量 x 的偏导为零，对等式约束的拉格朗日因子偏导为零，对不等式约束的拉格朗日因子偏导为零

话题七. 书中 Page91 中提到 H 的 condtion number 是啥意义，还有 Jacobi 的 condition number 呢？（yc）

yc
Fig 4.6

张小彬@Bruce
4.6 不是海森矩阵吗

yc
对，问的就是 Hessian 的 condition number 具体是啥意义

清
其实海森矩阵的条件数越大就表示我们算出来的值越不安全，模型的参数稍微改变一下 值就会有很大的变化

yc
那我能不能这样说，Jacobi 也是类似？

张小彬@Bruce
能不能这样子理解？condition number 衡量的是对矩阵求逆结果到底靠谱不

谱，鲁棒不鲁棒，海森矩阵和雅各比矩阵都有求逆的过程

yc
H 的 con num 是曲率的 恶劣程度
J 的 con num 是梯度的 恶劣程度

清
我如果没记错的话雅可比是一阶的

AG-GROUP 元芳
所以不还是蝴蝶效应的大小么
变化一点点对结果的影响大小

yc
@张，我也不清楚啊，con num 和 inv 有关吗

清
我们的结果一般是一个数值，不是一个向量。怎么求雅可比矩阵？

张小彬@Bruce
Ax = b 的结果就是 x = A^-1 b，那么 ill-condition 正好对应矩阵的逆；就是说 A 的一点点变动，都会导致 A^-1 的很大变动。这是我的理解

yc
好的

清
新的思路哎

安兴乐
雅可比在高数下有个详细求法

张小彬@Bruce
雅各比矩阵是 f: x-> y，但是 x 和 y 都是多元的，所以一阶导数是个矩阵

清

之前没见过，学到了

张小彬@Bruce
如果是 f : x-> y，而 x 是多元的，y 是个数值的话，那么一阶导数梯度，是个向量；二阶偏导是海森矩阵，是矩阵。

清
对

张小彬@Bruce
那个雅各比矩阵好像在 RNN 的一些论文里看到过 ,因为 RNN 处理的就是 序列到序列的映射

清
最主要深度学习中损失函数的输出是一个数值不会使用雅可比矩阵的

张小彬@Bruce
不过具体我也不知道干嘛用的。

For feedforward neural networks, the *Jacobian* J is the matrix of partial derivatives of the network output vector y with respect to the input vector x:

$$J_{ki} = \frac{\partial y_k}{\partial x_i} \tag{3.36}$$

These derivatives measure the relative sensitivity of the outputs to small changes in the inputs, and can therefore be used, for example, to detect irrelevant inputs.

yc
BP 会用到吗？

Yisen
不会吧

安兴乐
说实话，写简单 BP 的时候我没有用到

End

写在最后

非常感谢此次进行讨论交流的朋友们以及群内支持的朋友们 ,希望我们读书会能让大家学到更多，并且讨论后可以对原书有更独到的理解。

#广告时间#
免费线上技术分享，视觉前沿资讯关注请关注极市平台公众号。