softmax 回归的概率分类是如何实现计算损失的？

普通线性回归

函数 $\hat{y} = Xw + b$

损失 L_2 损失 $\frac{1}{2}(\hat{y} - y)^2$ 1/2是为了消掉2

更新参数 SGD

softmax

↓

用 one-hot 对类别编码.

用交叉熵来衡量损失

↓

衡量两个概率区别

交叉熵 以e为底

$$H(p, q) = \sum_i -p_i \log(q_i)$$

$$l(y, \hat{y}) = \sum_i -y_i \log(\hat{y}_i)$$

y 真实值 $\hat{y}$ 输出值

由于 one-hot 编码

y_i 只有一个元素为1

简化.

$$l(y, \hat{y}) = -\log \hat{y}_j$$ 真实类别的概率

综上

代码中要获取交叉熵结果仅知道真实类别的概率即可

为什么初始化参数要跟据输入输出来规定矩阵大小？

$y = xW + b$

W ↓ 权重

x 是一个批量的样本　输入层 $28 \times 28 = 768$ 个特征

输出层 10个类

batch_size 256

x 大小为 256×768　矩阵乘法　768×10 的 W

所以输入输出决定了初始化权重大小.

为什么 batch_size 中的梯度要 loss 求和计算?

1. 因为求梯度消耗资源.
2. loss 求和指求整个 batch_size 的总 loss　对不同参数求偏导.

↓

得到的也是对不同参数的总 batch_size 梯度

3. 在 SGD 中更新参数时除以 batch_size 以使其归一化.