智能之门

神经网络和深度学习入门

（基于Python的实现）

STEP 1 基本概念

第 3 章

损失函数

本部分主要介绍损失函数的概念和种类，着重说明了神经网络中目前最常用的均方差损失函数（用于回归任务）和交叉熵损失函数（用于分类任务）。

3.1 损失函数概论

- **“损失”，即所有样本的“误差”总和：**

$$J = \sum_{i=1}^{m} loss_i$$

在黑盒子的例子中，我们如果说“某个样本的损失”是不对的，只能说“某个样本的误差”，因为样本是一个一个计算的。如果我们把神经网络的参数调整到完全满足独立样本的输出误差为0，通常会令其它样本的误差变得更大，这样作为误差之和的损失函数值，就会变得更大。所以，我们通常会在根据某个样本的误差调整权重后，计算一下整体样本的损失函数值，来判定网络是不是已经训练到了可接受的状态。

- **损失函数的作用：计算神经网络每次迭代的前向计算结果与真实值的差距，从而指导下一步的训练向正确的方向进行。**

3.1 损失函数概论

➢ 损失函数使用步骤

- 用随机值初始化前向计算公式的参数。
- 代入样本，计算输出的预测值。
- 用损失函数计算预测值和标签值（真实值）的误差。
- 根据损失函数的导数，沿梯度最小方向将误差回传，修正前向计算公式中的各个权重值。
- 重复步骤2，直到损失函数值达到一个满意的值就停止迭代。

3.1 损失函数概论

➢ 常用样本损失

- 符号规则：a 是预测值，y 是样本标签值，J 是损失函数值。

✓ 0-1误差（GOLD STANDARD LOSS）

$$loss = \begin{cases} 0, y = a \\ 1, y \neq a \end{cases}$$

✓ 绝对值损失（ABSOLUTE LOSS）

$$loss = |y - a|$$

✓ 铰链/折页损失或最大边界损失（HINGE LOSS）

$$loss = max(0,1 - y \cdot a), y = \pm 1$$

✓ 对数损失（LOG LOSS）

$$loss = -[y \ln a + (1 - y) \ln(1 - a)], y \in \{0,1\}$$

✓ 平方损失（SQUARED LOSS）

$$loss = (y - a)^2$$

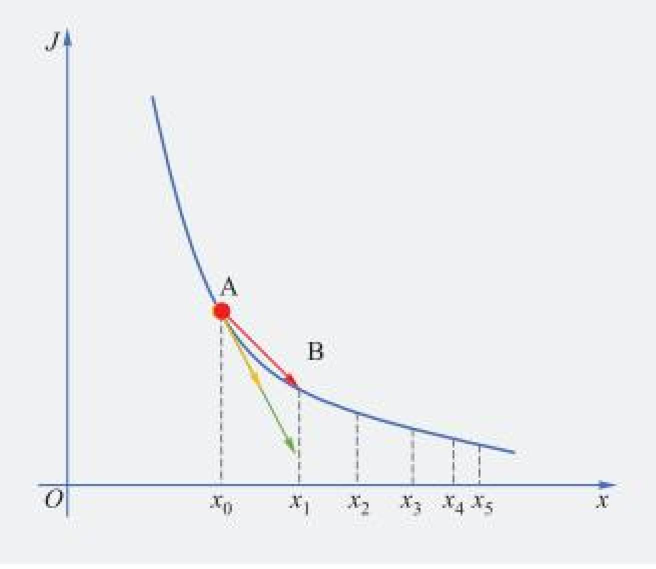
✓ 指数损失（EXPONENTIAL LOSS）

$$loss = e^{-y \cdot a}$$

3.1 损失函数概论

➢ 二维图像理解

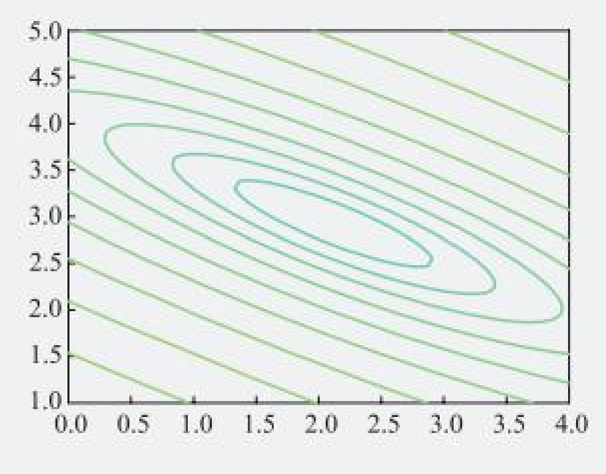
- 如右图，纵坐标是损失函数值，横坐标是变量。不断地改变变量的值，会造成损失函数值的上升或下降。而梯度下降算法会让计算沿着损失函数值下降的方向前进。
  - ✓ 假设我们的初始位置在A点，$x = x_0$，损失函数值（纵坐标）较大，回传给网络做训练；
  - ✓ 经过一次迭代后，我们移动到了B点，$x = x_1$，损失函数值也相应减小，再次回传重新训练；
  - ✓ 以此节奏不断向损失函数的最低点靠近，经历了x_2, x_3, x_4, x_5；
  - ✓ 直到损失值达到可接受的程度，比如x_5的位置，就停止训练。



3.1 损失函数概论

➢ 等高线图理解

- 如右图，横坐标是变量 w，纵坐标是变量 b。两个变量的组合形成的损失函数值，在图中对应处于等高线上的唯一的坐标点。 w, b 所有不同值的组合会形成一个损失函数值的矩阵，把矩阵中具有相同（相近）损失函数值的点连接起来，可以形成一个不规则椭圆，其圆心位置的损失值为 0，也是要逼近的目标位置。
  - ✓ 这个椭圆如同平面地图的等高线，来表示的一个洼地，中心位置比边缘位置要低，通过对损失函数值的计算，对损失函数的求导，会带领我们沿着等高线形成的梯子一步步下降，无限逼近中心点。



3.2 均方差损失函数

均方差函数是最直观的一个损失函数，计算预测值和真实值之间的欧氏距离，主要用于回归任务，公式如下：

$$loss = \frac{1}{2}(z-y)^2, \qquad J = \frac{1}{2m}\sum_{i=1}^{m}(z_i - y_i)^2$$

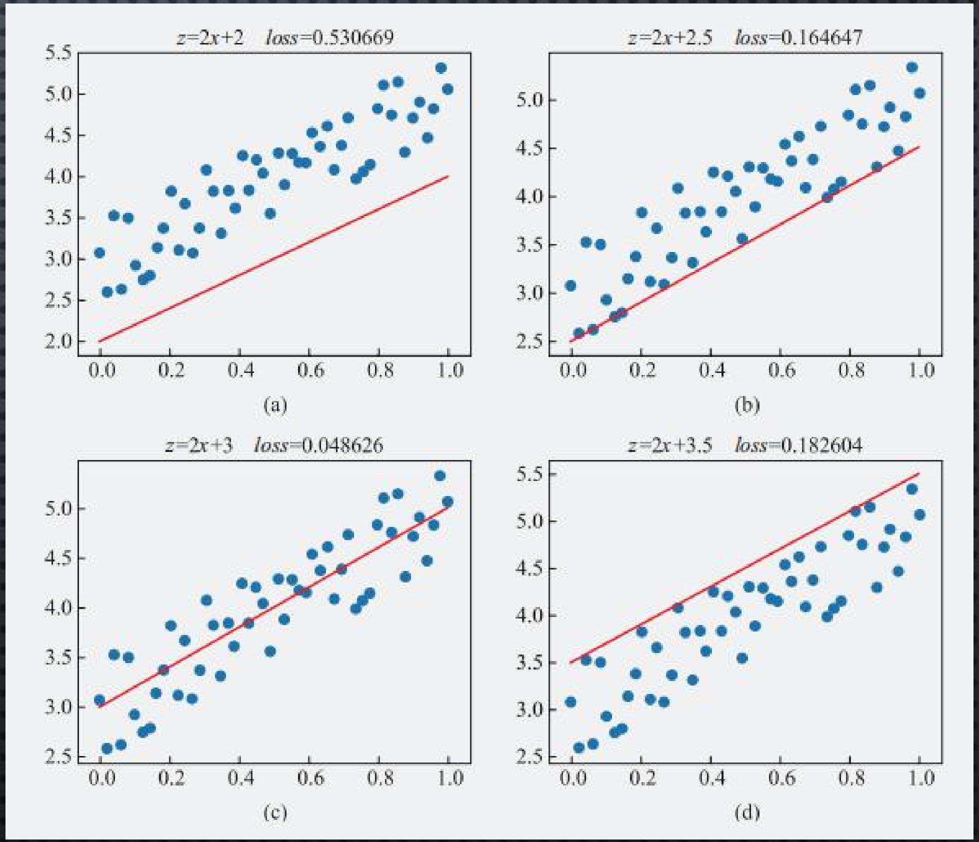
以下是绝对值损失函数与均方差损失函数的比较，可见MSE对某些偏离大的样本比较敏感，从而引起监督训练过程的足够重视，以便回传误差。

样本标签值	样本预测值	绝对值损失函数	均方差损失函数
[1,1,1]	[1,2,3]	|1-1|+|2-1|+|3-1|=3	$(1-1)^2+(2-1)^2+(3-1)^2=5$
[1,1,1]	[1,3,3]	|1-1|+|3-1|+|3-1|=4	$(1-1)^2+(3-1)^2+(3-1)^2=8$
		4/3=1.33	8/5=1.6

3.2 均方差损失函数

➢ 实际应用案例

- 假设有一组数据点，我们希望对其进行直线拟合。
  - ✓ (a) 初始情况下 $Loss = 0.53$。
  - ✓ (b) 直线略向上平移，$Loss = 0.16$，误差较(a)大幅减小。
  - ✓ (c) 直线继续向上平移，$Loss = 0.048$，此后还可以继续尝试平移（改变 b 值）或者变换角度（改变 w 值），得到更小的$Loss$。
  - ✓ (d) 直线偏离最佳位置，$Loss = 0.18$。

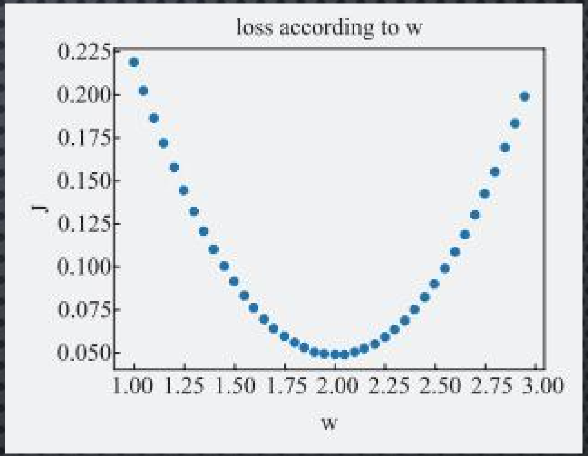


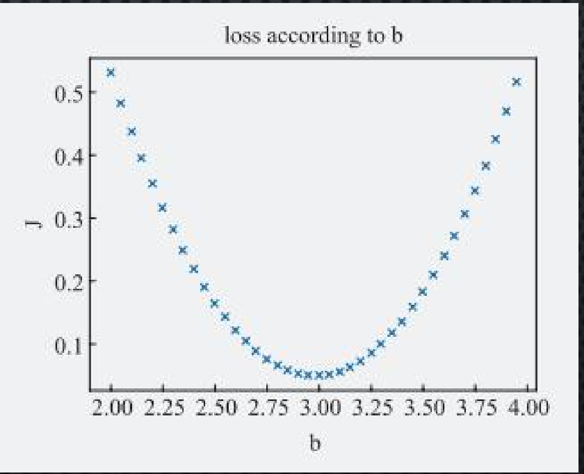
3.2 均方差损失函数

➢ **导数公式**

$$\frac{\partial J}{\partial a_i} = a_i - y_i$$

- 导数值可取全体实数值，被反向传播到前面的计算过程中以后，就会引导训练过程朝正确的方向尝试。
- 右面两图为本例中，损失函数值 J 分别随参数 w 和 b 的变化情况。

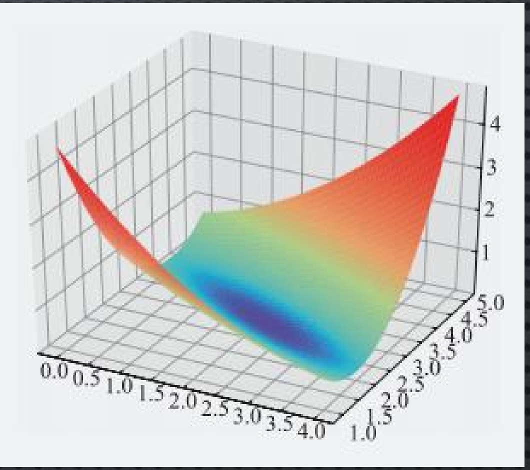


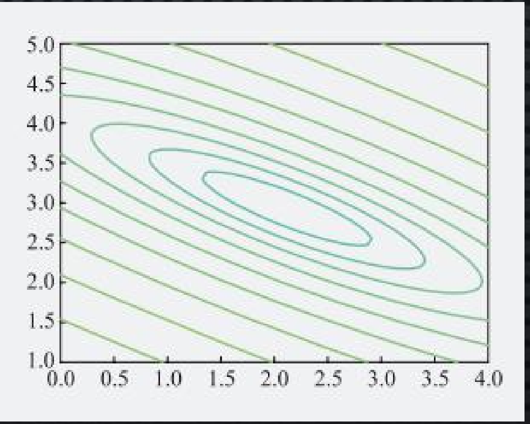


3.2 均方差损失函数

➢ 可视化

- 三维可视化
  - ✓ 横坐标为w，纵坐标为b，针对每一个w,b的组合计算出一个损失函数值，用三维图的高度来表示这个损失函数值。右图中的底部并非一个平面，而是一个有些下凹的曲面，只不过曲率较小。
- 二维可视化
  - ✓ 在平面地图中，经常会看到用等高线的方式来表示海拔高度值，右下图即为三维可视化图在平面上的投影。





3.3 交叉熵损失函数

信息量

- 一个事件 x 发生的概率 $p(x)$ 越大，那么它一旦发生时的信息量 $I(x)$ 就越大。

$$I(x) = -\ln p(x)$$

熵

- 事件发生信息量的期望。

$$H(x) = -\sum_{j=1}^{n} p(x_j) \ln p(x_j)$$

相对熵/KL散度

- 衡量两个概率分布的差异，相当于信息论范畴的均方差。

$$D_{KL}(p\|q) = \sum_{j=1}^{n} p(x_j) \ln \frac{p(x_j)}{q(x_j)}$$

交叉熵

- KL散度与熵值之和，即为香农信息论中的重要概念——交叉熵，以度量两个概率分布间的差异性信息。

$$H(p,q) = -\sum_{j=1}^{n} p(x_j) \ln q(x_j)$$

3.3 交叉熵损失函数

在机器学习中，需要评估标签值和预测值之间的差距，由于数据总体分布的熵值确定，因而可直接用交叉熵代替KL散度作为分类任务的损失函数。

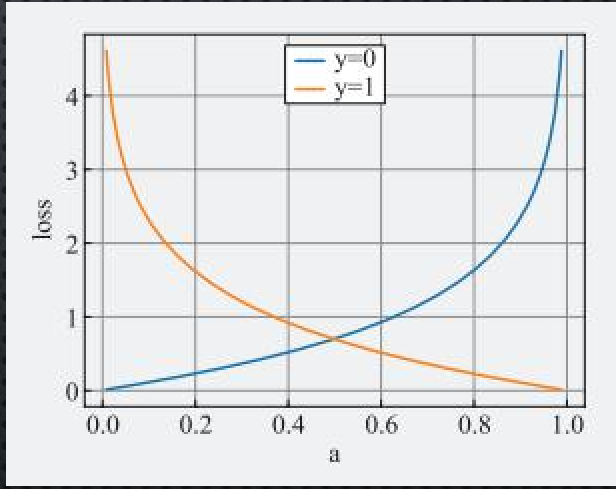
- 交叉熵损失函数（单样本、多样本）分别表示如下，其中m代表样本个数，n代表分类个数：

$$loss = -\sum_{j=1}^{n} y_j \ln a_j, \qquad J = -\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} \ln a_{ij}$$

- 特别地，二分类任务的损失函数可表示如下，易见预测输出越接近实际输出，损失函数值越小，训练结果越准确，如右图：

$$loss = -[y \ln a + (1-y)\ln(1-a)]$$

$$J = -\sum_{i=1}^{m}[y_i \ln a_i + (1-y_i)\ln(1-a_i)]$$



3.3 交叉熵损失函数

➢ **二分类任务**

- 假设学会了某门课程的标签值为1，没有学会的标签值为0。建立一个预测器，对一个特定的学员，根据出勤率、课堂表现、作业情况、学习能力等来预测其学会该课程的概率。
  - ✓ 对于学员甲，预测其学会的概率为0.6，而实际上该学员通过了考试，所以，学员甲的交叉熵损失函数值是：

$$loss_1 = -[1 \times \ln 0.6 + 0 \times \ln 0.4] = 0.51$$

  - ✓ 对于学员乙，预测其学会的概率为0.7，而实际上该学员也通过了考试。所以，学员乙的交叉熵损失函数值是：

$$loss_2 = -[1 \times \ln 0.7 + 0 \times \ln 0.3] = 0.36$$

◆ 预测值越接近真实标签值，交叉熵损失函数值越小，反向传播的力度越小。

3.3 交叉熵损失函数

- **多分类任务**
  - **假设期末考试有三种情况：**
    - ✓ **优秀，标签值 OneHot 编码为 [1,0,0]。**
    - ✓ **及格，标签值 OneHot 编码为 [0,1,0]。**
    - ✓ **不及格，标签值 OneHot 编码为 [0,0,1]。**
    - 假设预测学员丙的成绩为优秀、及格、不及格的概率为 [0.2,0.5,0.3]，而真实情况是该学员不及格，则得到的交叉熵是：

$$loss_3 = -[0 \times \ln 0.2 + 0 \times \ln 0.5 + 1 \times \ln 0.3] = 1.2$$

    - 假设我们预测学员丁的成绩为优秀、及格、不及格的概率为：[0.2,0.2,0.6]，而真实情况是该学员不及格，则得到的交叉熵是：

$$loss_4 = -[0 \times \ln 0.2 + 0 \times \ln 0.2 + 1 \times \ln 0.6] = 0.51$$

- 预测值越接近真实标签值，交叉熵损失函数值越小，反向传播的力度越小。

3.3 交叉熵损失函数

- **QUESTION**

  - 为什么不能使用均方差损失函数作为分类问题的损失函数？
    - ✓ 凸性与最优解
    - ✓ 求导运算的复杂性和运算量

THE END

谢谢！