**模式识别复习资料**

1. **选择题（6\*3）**
2. 下列属于非监督学习问题的是（  **A**  ）

A，根据1000位顾客对商场促销活动的购买行为对顾客进行分类；

B，根据AC米兰和国际米兰的既往战绩估计下一次“米兰德比”的结果；

C，根据往年水稻产量预测下一年度产量；

D，根据学生的期中考试成绩和平时作业成绩预测期末考试成绩；

**非监督学习 是指在没有类别信息情况下，通过对所研究对象的大量样本的数据分析实现对样本分类的一种数据处理方法**。

1. Logistic 回归的输出可以作为概率，假定训练好的Logistic 回归假设h（x）对新样本x的输出h（x）=0.2，意味着（ **A**  ）

A，p（y=1|x）=0.2 B，p（y=1|x）=0.8

C，p（y=0|x）=0.2 D，p（y=0|x）=0.1

1. 在训练Logistic 回归分类模型时，下列描述正确的是（ **C**  ）
2. 增加新的特征往往能减少过拟合；
3. 引入正则化约束往往能提升在训练集上的性能；
4. 增加新的特征往往能提升在训练集上的性能；
5. 引入正则化约束往往能降低在验证集上性能；

4，假设已经根据医学检验数据训练了一个线性Logistic回归分类模型来进行癌症筛查，该分类器的输出为，其决策函数为,如果增加threshold的值，关于该分类器的性能下列描述正确的是（  **C**  ）

A，召回率（Recall）增加 B，召回率（Recall）不变

* 1. 精度（Precision）增加 D，精度（Precision）不变

5，实际应用SVM进行图像中的手写字符识别时，下列描述正确的是（  **B** ）

1. 如果算法出现了过拟合，增加惩罚因子C的值可以提升性能；
2. 如果算法出现了欠拟合，收集并标注更多的图像可以提升性能；
3. 在训练完线性SVM分类器后，只能用0作为阈值来进行决策；
4. 训练样本数较少时采用非线性核SVM可以提升性能；

6，关于主元分析的应用，下列描述不正确的是（ **D** ）

A，数据压缩 B，数据降维 C，数据可视化 D，减少过拟合

1. **简答题（10\*5）**

**1.K均值算法的步骤和K怎么选取**

**k均值聚类算法（k-Means Clustering）步骤**：

(1) 随机选取k个对象作为初始聚类中心；

(2) 开始迭代，在迭代中对于每个样本点按照最小距离原则分配到最邻近聚类；

(3) 使用每个聚类的样本均值更新聚类中心；

(4) 重复(2)(3)迭代更新，直至误差小到某个值或者到达一定迭代次数，误差不变.

**K值的选取**：

**1**.与层次类聚算法结合，先通过层次类聚算法得出大致的类聚数目，并且获得一个初始聚类结果，然后再通过K-均值法改进聚类结果。  
 **2**.基于系统演化的方法，将数据集视为伪热力学系统，在分裂和合并的过程中，将系统演化到稳定平衡状态而确定K值。

**2.什么是模式识别，模式识别的意义**

**定义**：是通过计算机用数学技术方法来研究模式的自动处理和判读。是对表征事物或现象的各种形式的(数值的、文字的和逻辑关系的)信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。

**意义**：人们为了掌握客观事物，按事物相似的程度组成类别。模式识别的作用和目的就在于面对某一具体事物时将其正确地归入某一类别。

**3.模式识别预处理有哪些基本内容**

预处理是以去除噪声、加强有用的信息为目的, 并对输入测量仪器或其他因素所造成的退化现象进行复原的过程。

预处理一般有两种情况:

一是使数据的质量更好，比如用一些数字信号处理的方法去除信号中的噪声, 或者对一幅模糊的图像进行图像增强等, 确保有利于后期的模式识别工作；

另一种预处理是样本集的预处理, 比如样本集中异常值的剔除、类别的合并或分裂等。这一工作一般可以根据领域的专门知识进行, 也可以采用模式识别中的一些技术, 比如在进行后续工作之前先对样本集进行一次聚类分析。

1. **最大似然方法和贝叶斯估计方法的差异**

k均值聚类算法（k-means clustering algorithm）是一种[迭代](https://baike.baidu.com/item/%E8%BF%AD%E4%BB%A3/8415523?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95/_blank)求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95/_blank)。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，[误差](https://baike.baidu.com/item/%E8%AF%AF%E5%B7%AE/738024?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95/_blank)[平方和](https://baike.baidu.com/item/%E5%B9%B3%E6%96%B9%E5%92%8C/783894?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95/_blank)局部最小

**5.简述svm的基本思想**

支持向量机（SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器。思想是：假设样本空间上有两类点，我们希望找到一个划分超平面，将这两类样本分开，而划分超平面应该选择泛化能力最好的，也就是能使得两类样本中距离它最近的样本点距离最大。

**6.神经网络训练时是否可以将所有参数初始化0，为什么**

神经网络的训练时，不能将所有参数初始化为零，因为如果神经网络计算出来的输出值都一个样，那么反向传播算法计算出来的梯度值一样，并且参数更新值也一样，即如果权重初始化为同一个值，网络就不可能不对称（即是对称的）。

## 7.sigmoid用作激励函数时有什么缺点

（1）容易出现梯度消失的现象：当激活函数接近饱和区时，变化太缓慢，导数接近0，根据反向传递的数学依据是微积分求导的链式法则，当前导数需要之前各层导数的乘积，几个比较小的数相乘，导数结果很接近0，从而无法完成深层网络的训练。

（2）Sigmoid的输出不是0均值的：这会导致后层的神经元的输入是非0均值的信号，这会对梯度产生影响。以 f=sigmoid(wx+b)为例， 假设输入均为正数（或负数），那么对w的导数总是正数（或负数），这样在反向传播过程中要么都往正方向更新，要么都往负方向更新，导致有一种捆绑效果，使得收敛缓慢。

（3）幂运算相对耗时

## 8.k近邻的分类方法

**k-近邻：**一个样本x，要给它分类，即求出它的y，就从数据集中，在x的附近找离它最近的k个数据点，然后统计这k个数据点所属的分类与分类数量，若这k个数据点中，分类为c的数据点的数量最多，则将k分类为c。

从字义上看, 这个方法就是取未知样本 x 的 k 个近邻, 看这 k 个近邻中多数属于哪一类, 就把 x 归为哪一类。具体说就是在 N 个已知样本中, 找出 x 的 k 个 近邻。

## 9.2006年以后使得深度神经网络得到快速应用的技术突破在哪里，举四个例子

1. 图像字幕：可根据提供的图片，识别图片上的内容并匹配字幕，把它翻译为英文句子；
2. 风格迁移：利用深度神经网络改变图片的艺术风格，把梵高的星空风格移植到任意图片上；
3. 性取向判断：神经网络通过看照片中一个人的面部结构能够识该人的性取向，准确度达到87％；
4. 生成三维模型：利用神经网络搭建深度学习系统，只需要一幅二维草图，就能够自动生成三维模型；
5. 智慧医疗：利用ANN可以进行智能导诊、建立电子病历、医学影像分析、癌症早期筛查、假肢设计、移植时间优化等应用。

**10.简述隐马尔可夫模型的三个核心问题**

隐马尔可夫（HMM）理论有3个基本问题：

1、评估问题

给定一个HHM的模型和观测序列，如何高效的计算此模型产生的观测序列的概率。

2、解码问题

给定一个HHM的模型和观测序列，如何选择对应最佳的序列，使它在某状态下最优（出现的概率最大），比较好地解释观测值。

3、学习问题（训练问题）

给定观测序列，如何调整模型参数，以使得观察序列出现的概率最大化。

**三、计算题**

**1、假设我们有一个模型hθ（x）=g（θ0+θ1x1+θ2x2），其中函数y（z）= 为Logistic函数，并且参数向量θ=[-3,1,1]T。**

**(1)试给出对应的决策边界方程，并在图中画出对应的决策边界；**

**(2)试写出正则化Logistic回归的目标函数及求解最优参数的梯度下降算法。**

解：（1）

因为**hθ（x）=g（θ0+θ1x1+θ2x2），** **θ=[-3,1,1]T**

决策边界为：ATx -3+x1+x2=0

即：x1+x2=3

（2）

logistic regression 函数：和sigmoid函数一样，即

logistic regression的代价函数：

正则化目标函数：（加上蓝色的才正确）

利用梯度下降算法求参数：

Repeat{

}

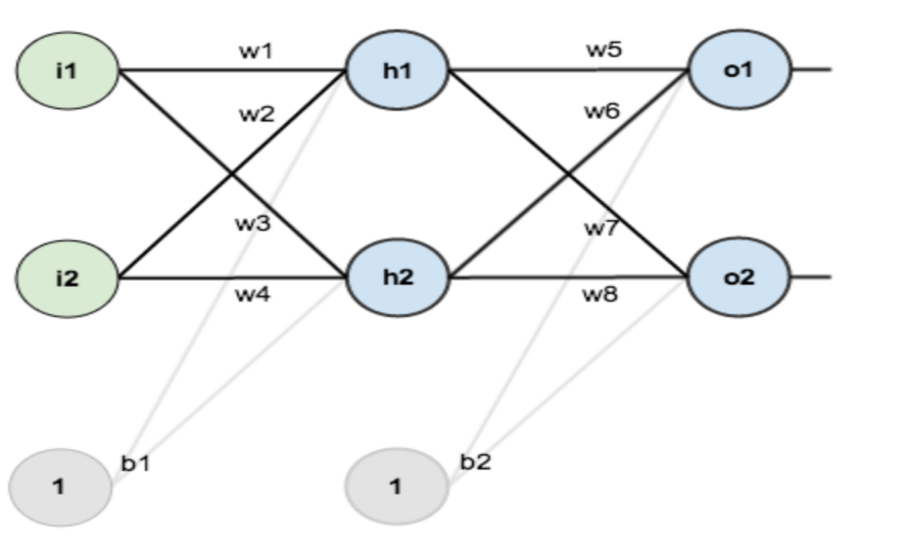
（加正则化）

决策边界：z = 0

(++=0)

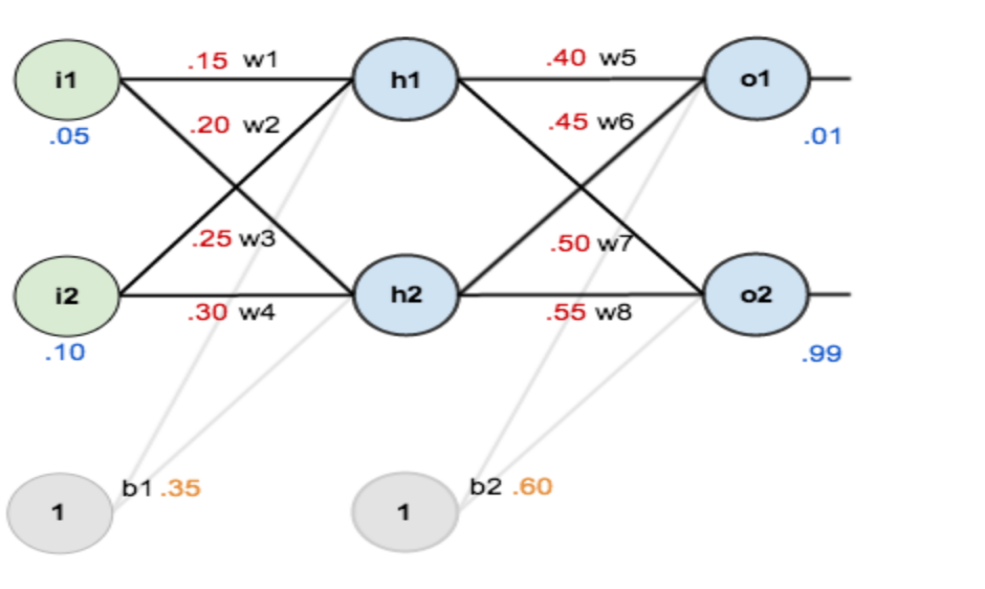
**2.熟悉神经网络的推导过程，三层弄懂就行**

假设，你有这样一个网络层：



　　第一层是输入层，包含两个神经元i1，i2，和截距项b1；第二层是隐含层，包含两个神经元h1,h2和截距项b2，第三层是输出o1,o2，每条线上标的wi是层与层之间连接的权重，激活函数我们默认为sigmoid函数。

　　现在对他们赋上初值，如下图：



　　其中，输入数据  i1=0.05，i2=0.10;

　　　　　输出数据 o1=0.01,o2=0.99;

　　　　　初始权重  w1=0.15,w2=0.20,w3=0.25,w4=0.30;

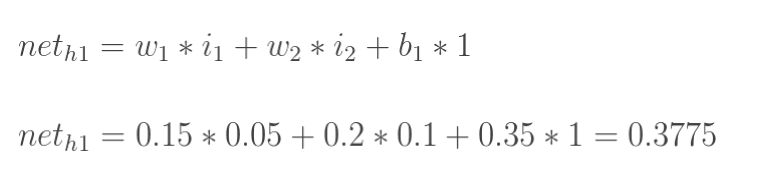
　　　　　　　　　  w5=0.40,w6=0.45,w7=0.50,w8=0.55

　　目标：给出输入数据i1,i2(0.05和0.10)，使输出尽可能与原始输出o1,o2(0.01和0.99)接近。

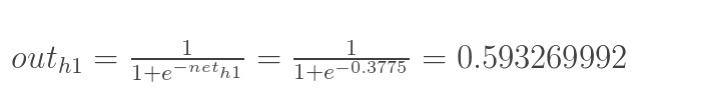
　　Step 1 前向传播

　　1.输入层---->隐含层：

　　计算神经元h1的输入加权和：



神经元h1的输出o1:(此处用到激活函数为sigmoid函数)：

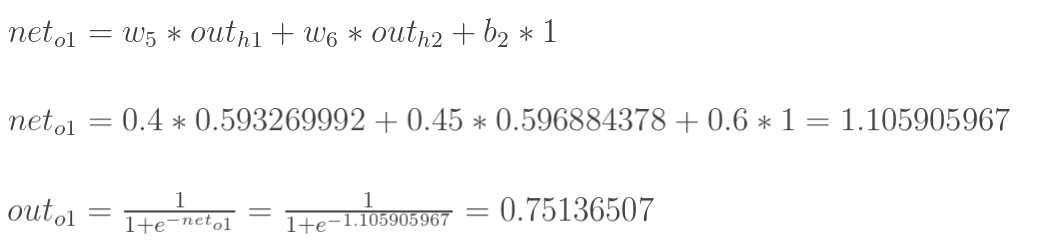


　　同理，可计算出神经元h2的输出o2：



　　2.隐含层---->输出层：

　　计算输出层神经元o1和o2的值：





这样前向传播的过程就结束了，我们得到输出值为[0.75136079 , 0.772928465]，与实际值[0.01 , 0.99]相差还很远，现在我们对误差进行反向传播，更新权值，重新计算输出。

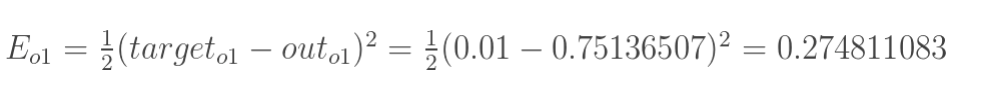
Step 2 反向传播

1.计算总误差

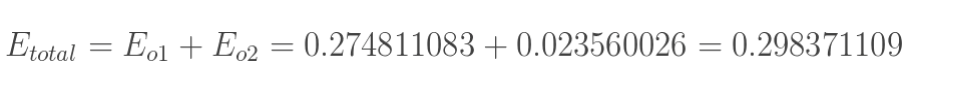
总误差：(square error)



但是有两个输出，所以分别计算o1和o2的误差，总误差为两者之和：

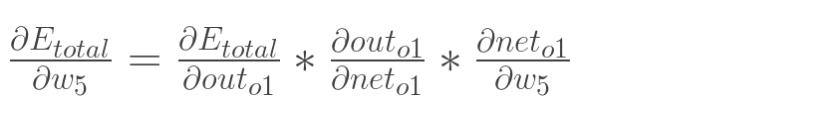




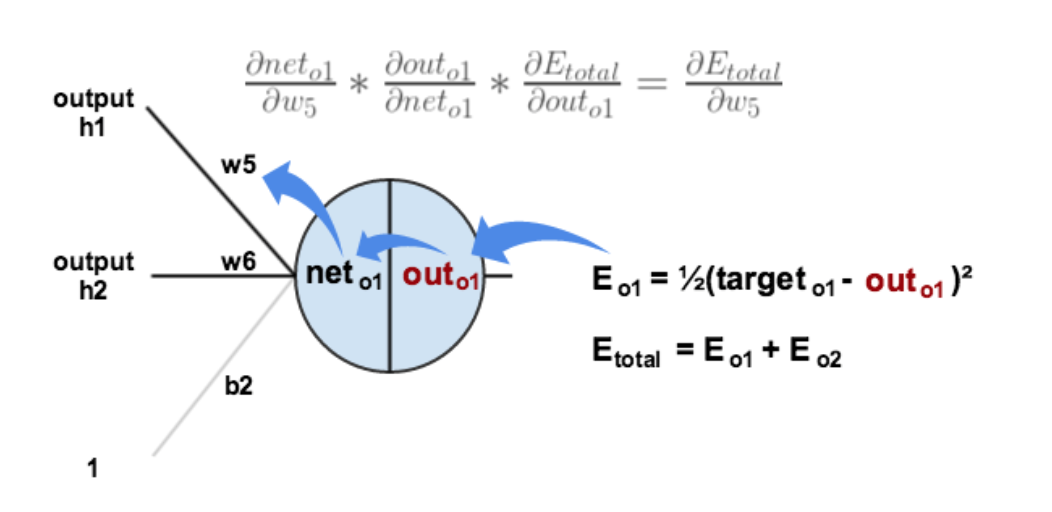


2.隐含层---->输出层的权值更新：

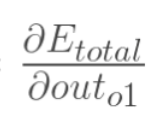
以权重参数w5为例，如果我们想知道w5对整体误差产生了多少影响，可以用整体误差对w5求偏导求出：（链式法则）

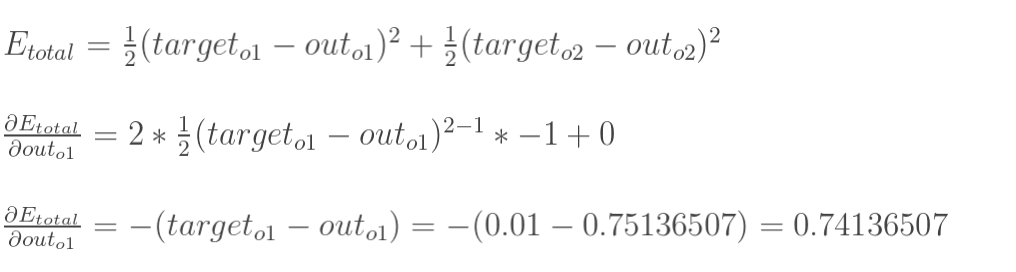


下面的图可以更直观的看清楚误差是怎样反向传播的：

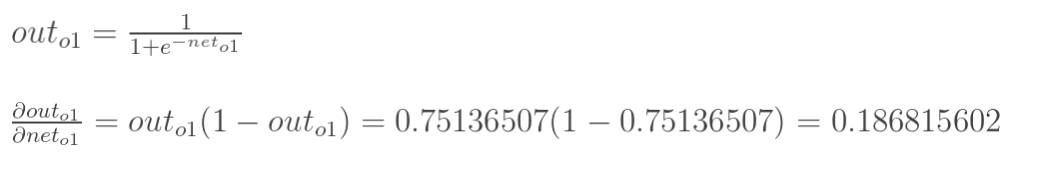


现在我们来分别计算每个式子的值：

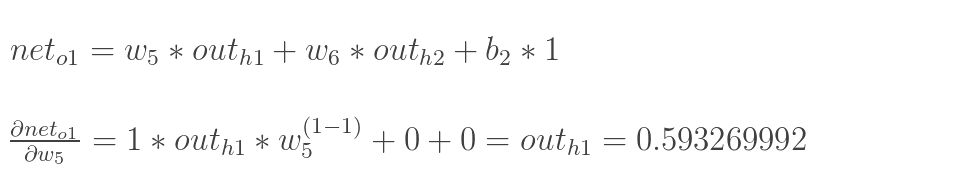
计算：



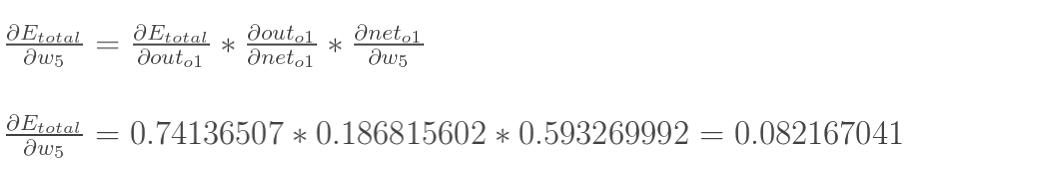
计算：



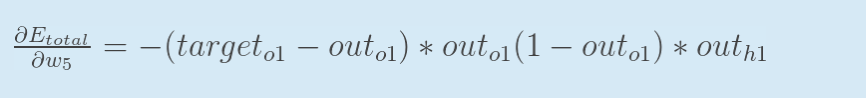
计算：



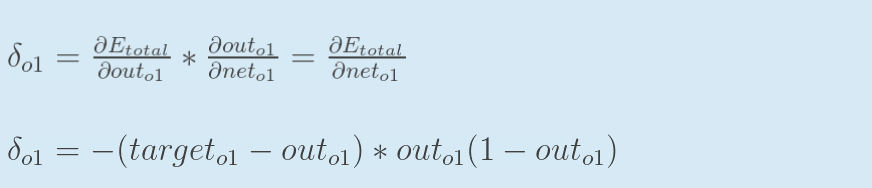
最后三者相乘：



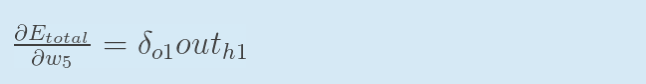
这样我们就计算出整体误差E(total)对w5的偏导值。



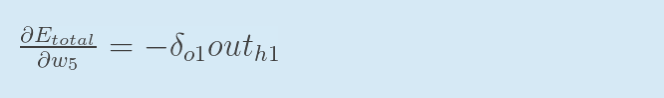
为了表达方便，用来表示输出层的误差：



因此，整体误差E(total)对w5的偏导公式可以写成：



如果输出层误差计为负的话，也可以写成：



最后我们来更新w5的值：



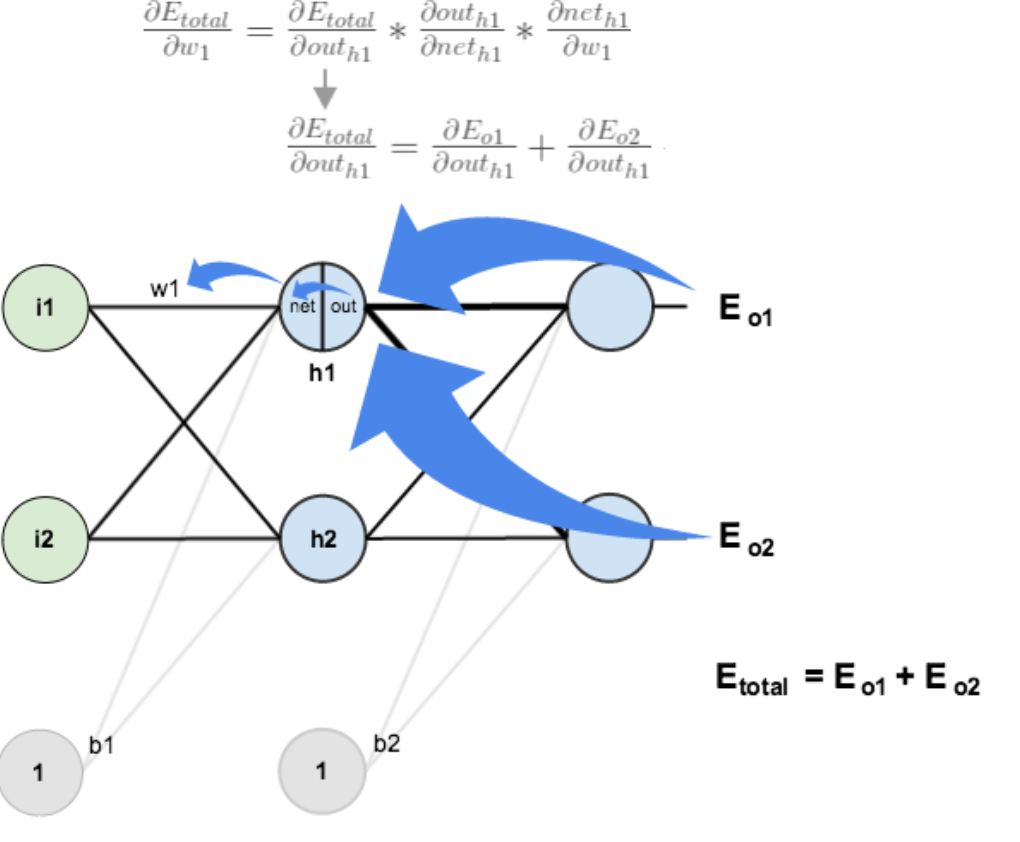
（其中，是学习速率，这里我们取0.5）

同理，可更新w6,w7,w8:



3.隐含层---->隐含层的权值更新：

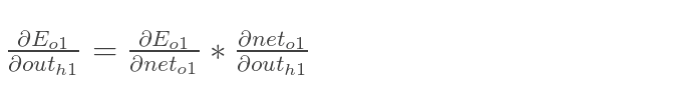
　方法其实与上面说的差不多，但是有个地方需要变一下，在上文计算总误差对w5的偏导时，是从out(o1)---->net(o1)---->w5,但是在隐含层之间的权值更新时，是out(h1)---->net(h1)---->w1,而out(h1)会接受E(o1)和E(o2)两个地方传来的误差，所以这个地方两个都要计算。

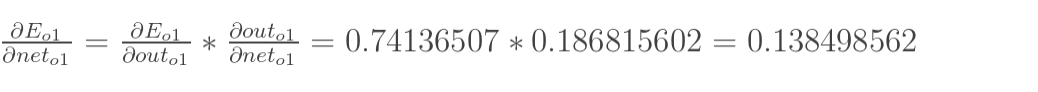


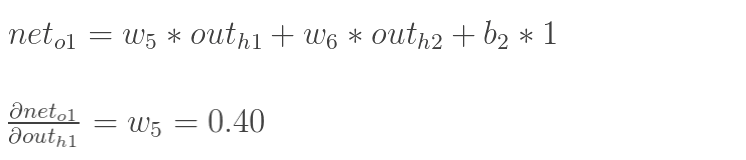
计算：



先计算：





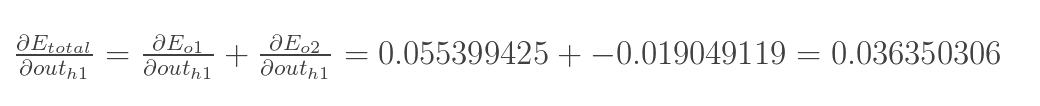




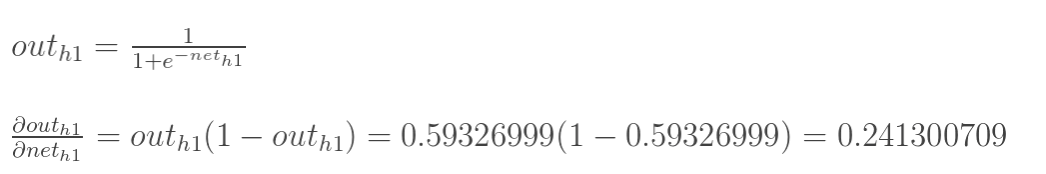
同理，计算出：



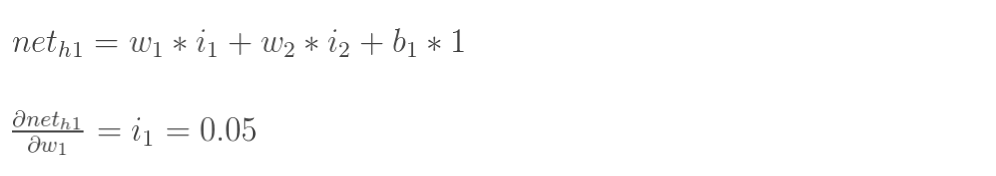
两者相加得到总值：



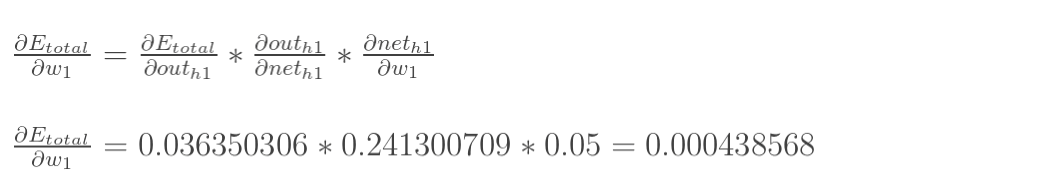
再计算：



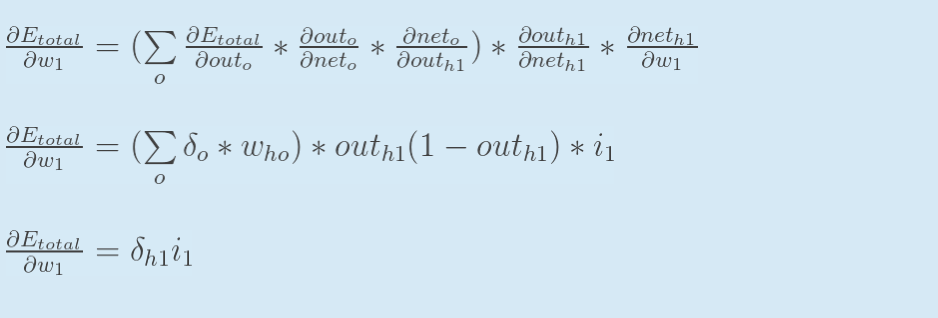
再计算：



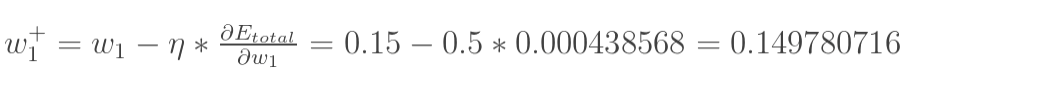
最后，三者相乘：



 为了简化公式，用sigma(h1)表示隐含层单元h1的误差：



最后，更新w1的权值：



同理，额可更新w2,w3,w4的权值：



**3.**

