speech and language processing 第五章报告

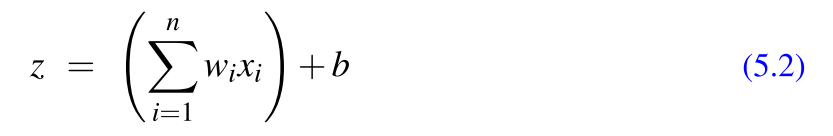
09017244 郑健雄

1. 本章概要：

Logistic Regression与朴素贝叶斯不同，是比较经典的判别模型。生成模型与判别模型是研究统计学习问题的两种思路，生成模型考虑的是一个类别是如何生成的，而判别模型则专注于可以将类别区分开来即可，比如生成模型分辨猫狗，知道什么是猫，什么是狗，而判别模型则不然，它只会看图像上的特征，比如如果狗都带项圈的话，那么它会将戴项圈的判别为狗。从数学上看，生成模型学习P(d|c)而判别模型则是学习P(c|d)。

典型的机器学习分类器一般由四个部件组成：

1. A feature representation of the input(输入的feature repression格式，比如向量)
2. A classification function that computes ˆy(一个分类函数，p(y|x)对应的公式，比如sigmoid和softmax)
3. An objective function for learning，比如交叉熵。
4. An algorithm for optimizing the objective function，比如随机梯度下降法。
5. 知识点总结：
   1. binary logistic regression
      1. 分类函数：sigmoid。logistic regression的思路主要是对于每个特征，都有一个权重值，代表其重要性以及对于正负方向的影响和一个偏差量。一般可以将预测结果标识如下：

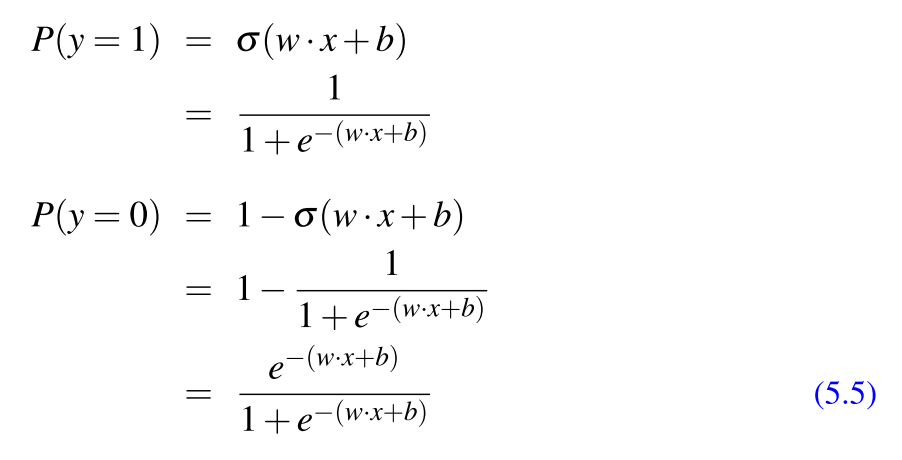




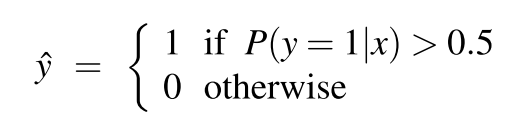
而为了将其转换为概率，使用了sigmoid：



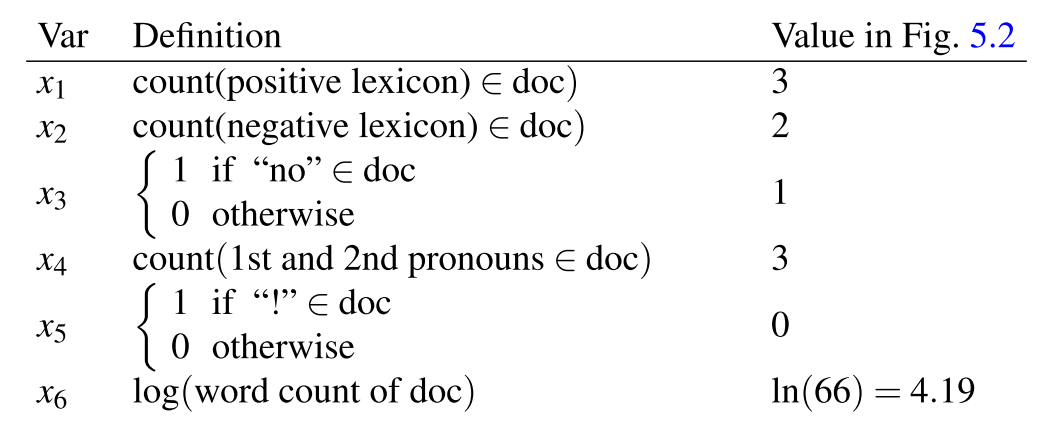
其进行二元分类的计算方法如下：



如果为1的概率大于0.5，则判断为正，反之思路一样：



* + 1. feature的设计：feature一般是可以通过人为设计，比如判断正负情绪可以设定如下的一些特征标准，但是这种设定一般较为主观，并且不一定全面。



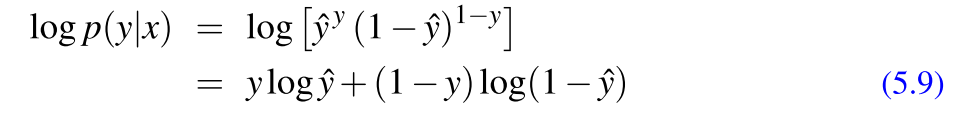
现在由于神经网络的普及，实际上特征可以通过神经网络输出向量来学习，比如word2vec这样的技术等，但是考虑到可解释性就比较麻烦。一般而言，logistic regression在大数据集上效果更好，而朴素贝叶斯则在小数据集上占优势。

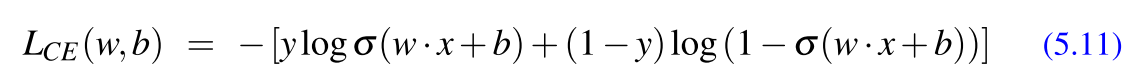
* + 1. 目标函数：

与分类函数不同，目标函数是用来进行优化的，机器学习问题一般可以理解为一个最优化问题，而在最优化问题中，我们要优化的目标是参数(w,b),比如进行最大化和最小化。一般而言，loss function和objective function可以相互转化，目标都是找出最优的参数，在logistic regression中，使用的是交叉熵损失函数作为目标来学习。



其一般定义如上，其目标也很明确，让预测结果尽可能靠近真实结果。

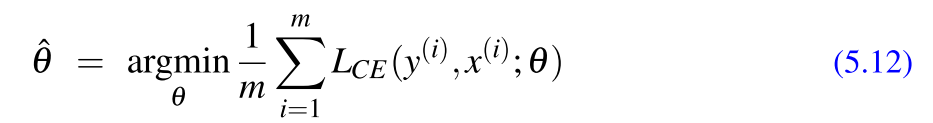




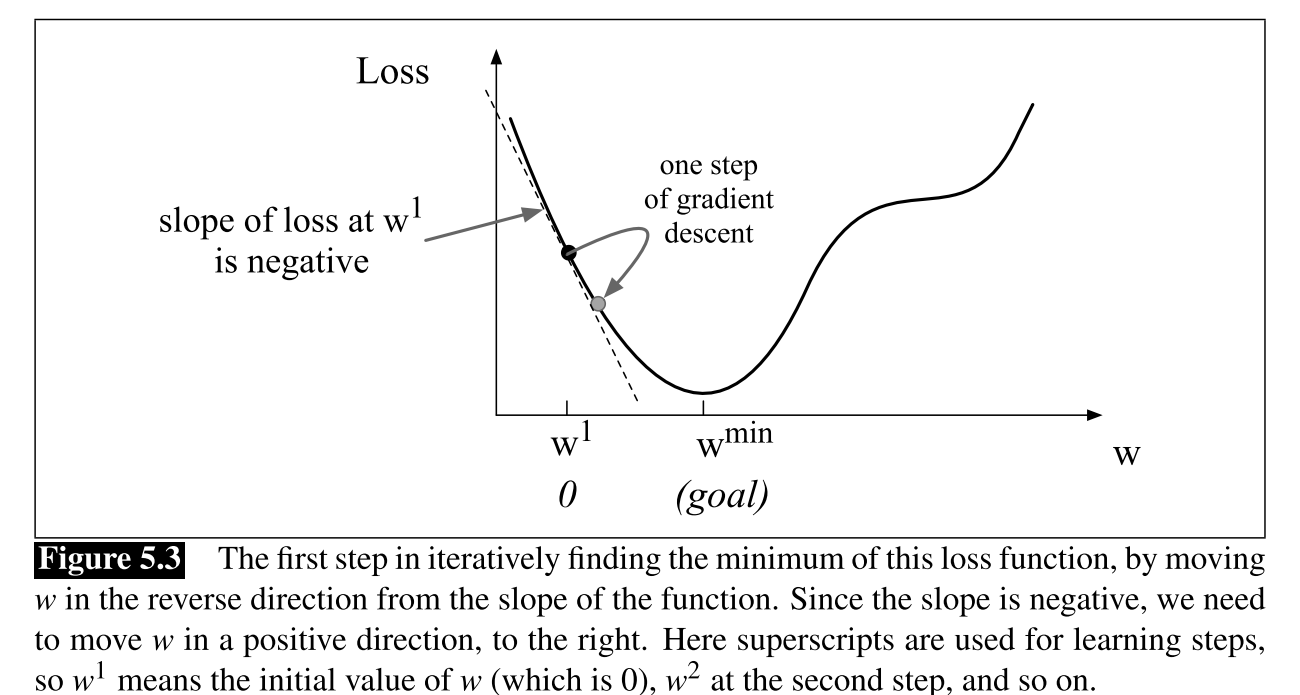
考虑到计算的方便性，一般使用对数形式来计算，而且可以看到最终使用的自变量是w和b，这是为了对其进行优化，因为y与x都是事实，是无法优化的。

* + 1. 梯度下降法：

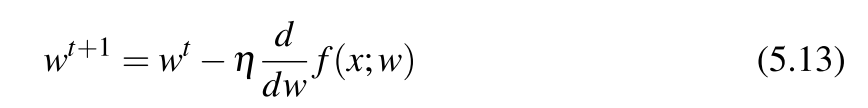
机器学习算法中，现在已经有了目标函数，但是如何让其最小化是个问题，这也是为什么提出优化算法的原因。一般我们的目标可以表示如下：

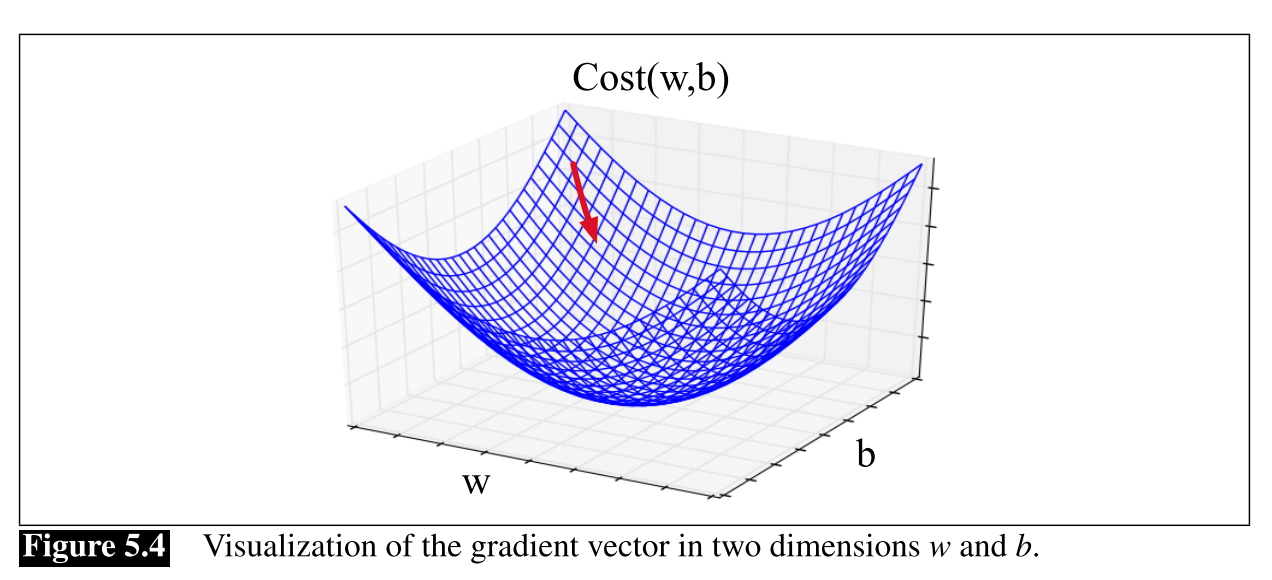


考虑到目标函数的性质是凸函数(只有一个全局最优解，不会出现局部最优，神经网络则容易出现局部最优)，那么一个比较好的办法其实是利用导数的性质来更新，这个过程就像下山一样。

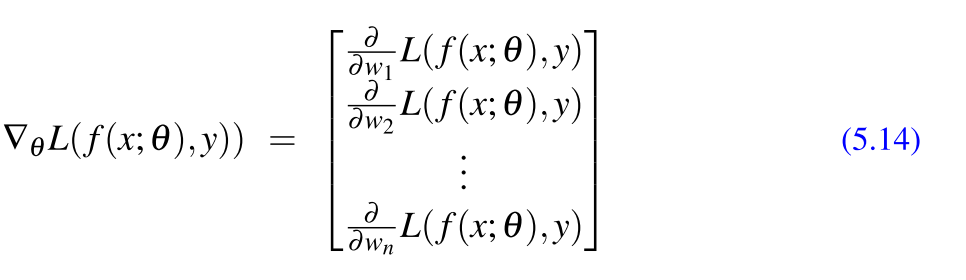


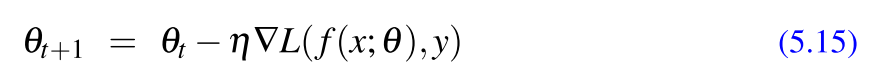
通过计算梯度，找到更新的方向，然后借助学习率，对权重进行更新。





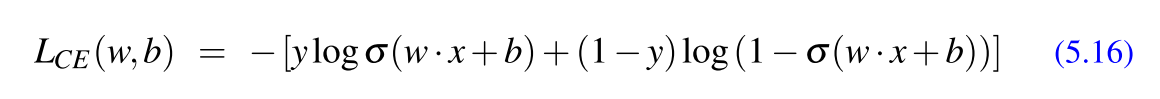
直观上看，梯度下降可以理解为上面的形式。

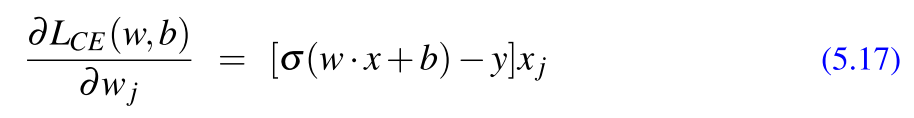




按照惯例，可以将权值向量计算与更新理解为上图，而f(x;theta)代表的是在给定参数theta(代表所有等待学习的参数)下的计算结果。

搞清楚梯度下降，在这个算法中，其可表示为：

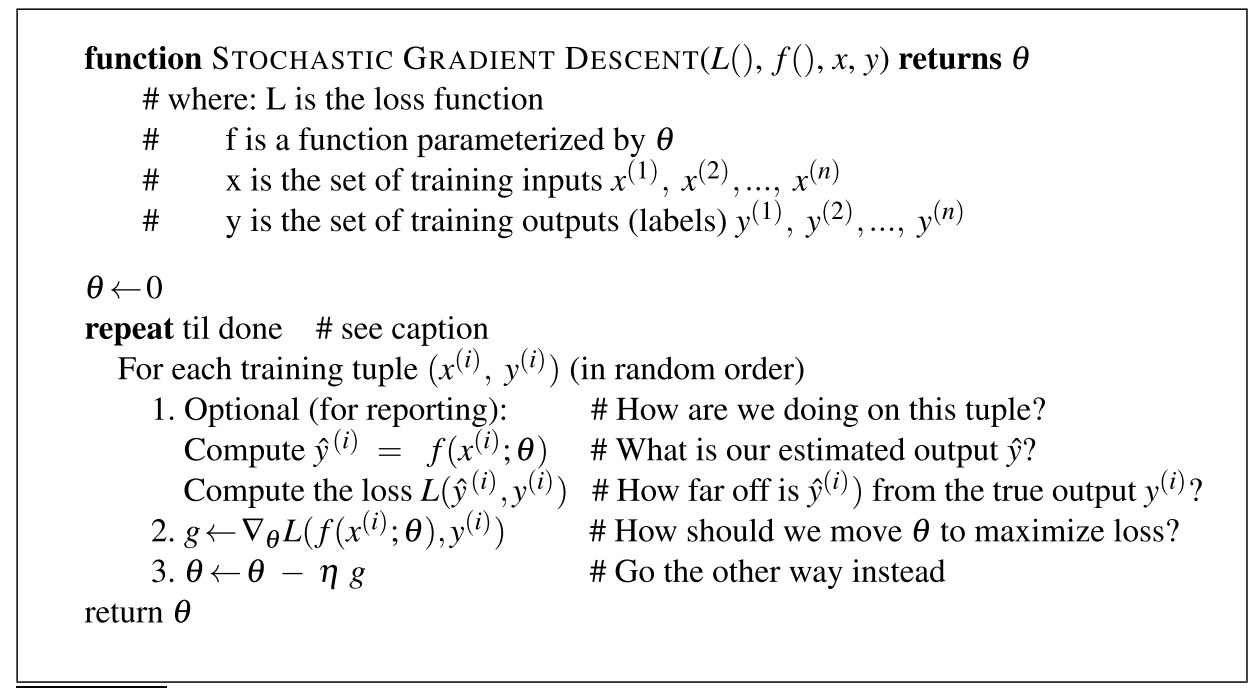




梯度的计算用到了链式法则以及lnx的性质，此处不再展开，直观上记忆可以理解为预测与真实情况的差值乘以权值对应feature的值。

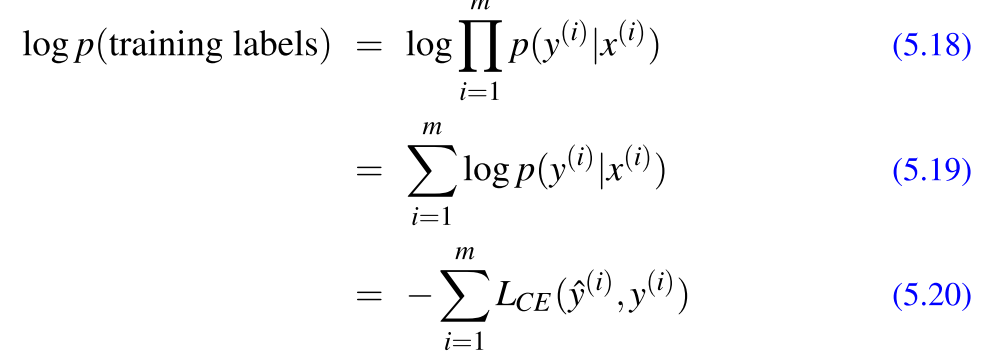
* + 1. The Stochastic Gradient Descent Algorithm：

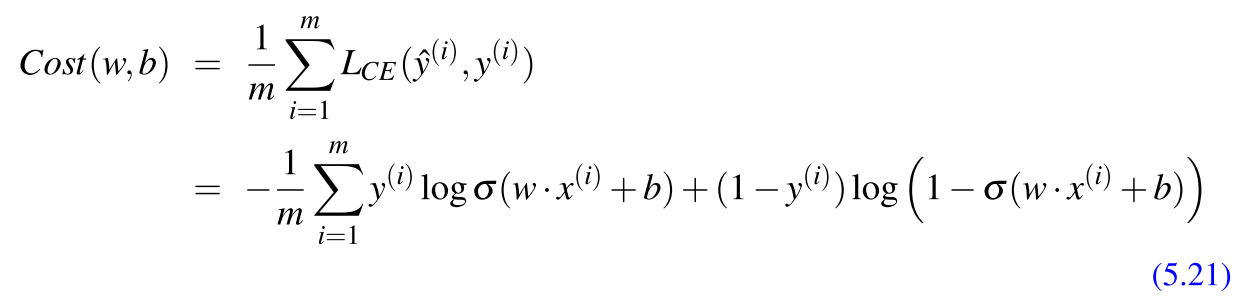
一般情况下，如果使用上面的模式进行更新，需要计算所有的样本，得到一个完美的方向，但是这非常费时间，随机梯度下降的好处很明显，其每次都根据一个样本来进行一些局部的更新，虽然不是很稳定。

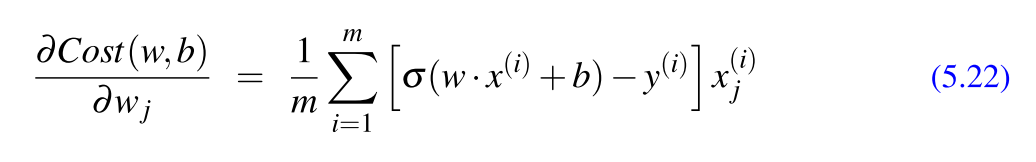


* + 1. Mini-batch：

Mini batch介于batch和随机梯度下降之间，其一次更新会使用一组样本，比如512，1024，这样速度快而且效率高。



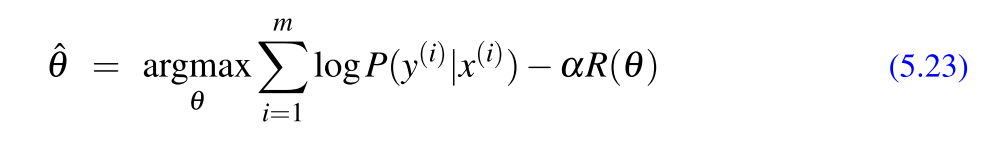




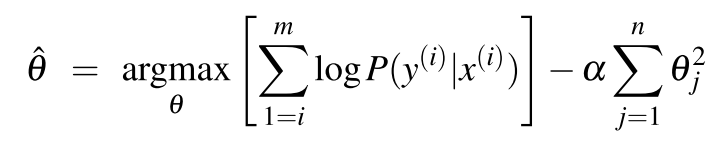
可以看到其计算的代价函数是一个平局值，而更新的参数也是如此，不是单独一个样本，而是多个样本效果的叠加。

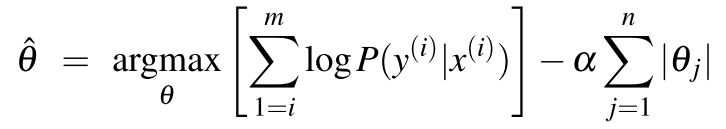
* + 1. 正则化：

一般机器学习都会面临过拟合的问题，也就是过于符合训练集。其结果就是对于真实的情况不能很好地进行预测，为了避免过拟合，引入了正则化。



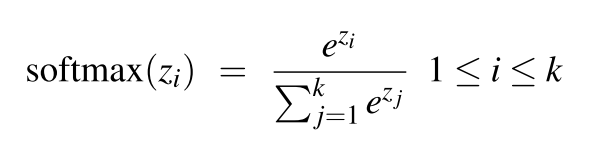
也就是在代价中引入一个惩罚项，一般而言，如果权值中的项比较小，其过拟合就不是很严重，而如果过大，那么损失函数也会变大，产生惩罚。一般来说，正则化函数可以用L2 regularization和L1 regularization.



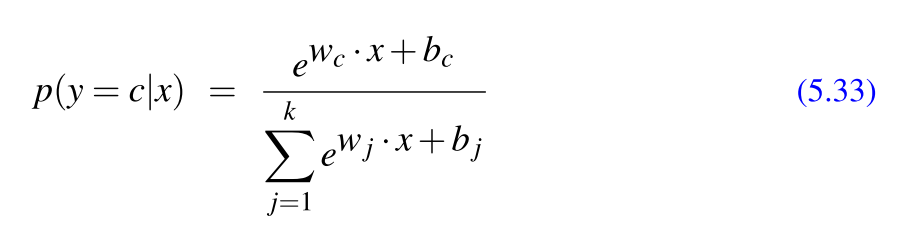


* 1. Multinomial logistic regression：

一般还有多元分类问题，其与二元分类类似。其预测类别使用了softmax函数。

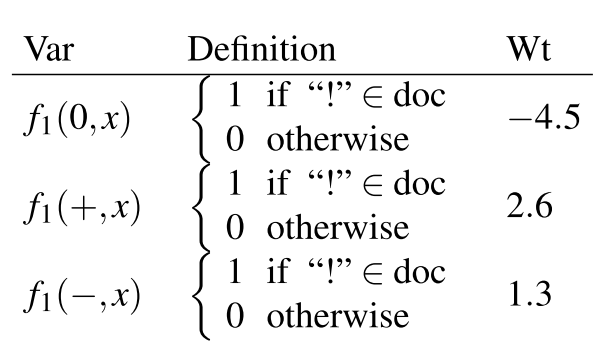


而其区别在于对于每个类别c，都有一个独特的w向量和偏差b。

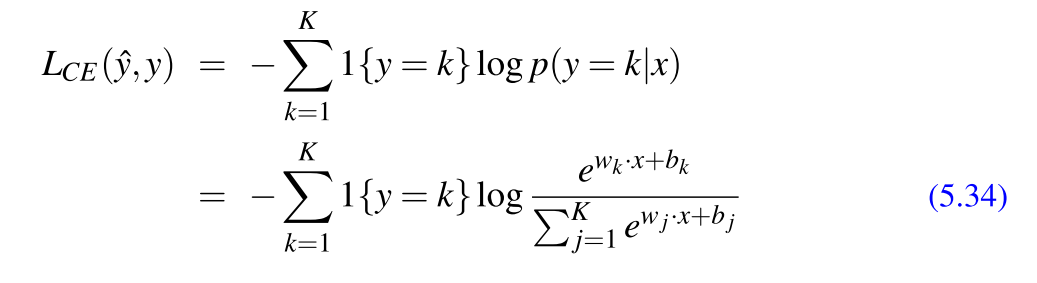


这个很好理解，毕竟不同类别的话，看重的特征也不同，使用sigmoid反而效果不好。

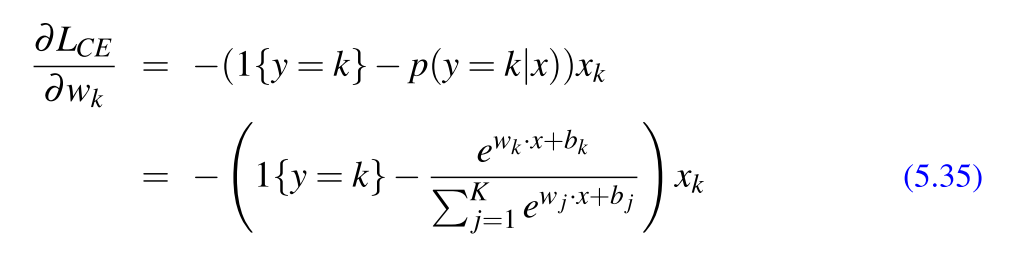
when we’re discussing features we will use the notation fi(c,x), meaning feature i for a particular class c for a given observation x.



学习过程也类似：



计算的梯度和二元的情况也十分类似：



logistic regression模型的价值在于其可以衡量特征的重要程度，比如说其可以看出是否出现否定词对于一个句子情感正负的表达的影响大不大，有没有意义。而这种可解释性也是机器学习过程中十分重要的要素。

1. 本章摘要：
   1. 二元logistic regression是一种监督学习算法，是判别模型，其组成成分也比较简单。sigmoid作为分类函数，feature的抽取可以人为，也可以通过机器学习得出，损失函数是交叉熵而学习算法是随机梯度下降或者mini-batch。
   2. 多元logistic regression和二元基本一样，其只是将sigmoid换成了softmax函数，并且没有使用交叉熵来进行判别，而是借助了softmax的性质。
   3. logistic regression可以用在很多应用中，也可以说明某些特征的价值，从而提供很好的解释性。