第三周学习汇报

分类问题

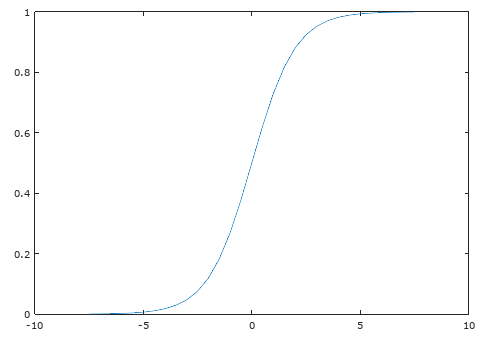
分类问题相比于线性问题，分类问题的结果是有限个，是离散的，线性问题的结果是无限的，是连续的。但是在我看来，分类问题是线性问题的一种特殊情况，所以可以用线性问题的预测函数来预测分类问题，但是并不是太准确。所以分类问题有了自己的解决方案：逻辑回归方法。

逻辑函数（S型函数）

线性回归预测函数有时会出现的情况，需要一个能够满足，分类问题预测函数如下：



图像显示如下：



从图像上可得，满足的要求。由所得的结果是一种概率，是指分类结果为1的概率，如=0.7，则说明分类结果为1的概率为70%，那么相反结果为0的概率则为30%。

决策边界

当0.5时，认为y=0; 当0.5时，认为y=1。

由图像可知，所以讨论的值。而当时，此时就是决策边界的函数。。

代价函数

使用其概率的-log函数作为代价函数，即使用NLL作为其代价函数。



通过代价函数可得到以下结论：（也可由图像可知）

1. 当时，=0
2. 当
3. 

简化代价函数：



向量化代价函数：



梯度下降法

使用梯度下降法用以找到参数，定义如下：（与线性回归类似）



代入得：



向量化得：



参数的其他高级计算方法介绍：

其他高级计算方法有共轭梯度法、BFGS(拟牛顿法)、L-BFGS

使用fminunc()方法计算参数，而不需要参数。

多类别分类：一对多模式

思路：选择其中个类别A，然后其他类别的都归为一类，目标是把类别A分出，然后以此类推，找出每个类的分类器。在进行预测的时候，将特征变量输入分类器中，选择最大值作为预测结果，即对应的类别。

过度拟合（Overfitting）

样本数据的拟合分为：欠拟合、拟合、过拟合三种情况。欠拟合是指代价函数过于简单平滑，使得样本数据不能更好的拟合。过拟合是指代价函数过于复杂曲折，特征变量过于复杂，使得已知样本数据准确拟合，而不能泛化到新的样本数据，不具有一般性。过度拟合问题的出现一般是由样本数据少、代价函数复杂造成，为了避免过度拟合问题的发生，可以采用减少特征变量数量和正规化两种方法。

正规化（Regularization）

正规化的思路：对于代价函数过于复杂的情况，通过添加正规化项，由于需要代价函数最小化，使得对应的参数变量接近0，从而减小过拟合的情况。参数变量越小，代价函数拟合越光滑，就不易出现过拟合的情况。对于特征变量多的情况，每个特征变量都有用，但由于事先不知道需要对那些特征参数进行缩小，所以对代价函数进行整体参数缩小（除外）。

正规化后的代价函数（Cost Fuction）如下：



注意：是正规化项（i从1开始），是正规化参数（是用来平衡欠拟合和过拟合，过大和过小都不利于拟合数据）。

正规化后的代价函数第一项是用来拟合数据的，第二项是用来保持特征变量较小。

正规化线性回归

线性回归有梯度下降和正规方程两种方法，正规化是对两种方法的实现过程进行调整。如下：

1. 梯度下降

由于改变，所有在每次迭代求时也发生了改变。如下：



化简后得：



1. 正规方程

正规化后得：



正规方程方法引入正规化，能够解决不可逆的问题。

正规化逻辑回归

代价函数：

（正规化前）

（正规化后）

为正规化项，为正规化参数。

学习总结：

这周主要学习了逻辑回归、梯度下降、正规化方法等。知识与前两周的知识联系，代价函数和梯度下降方法是接触比较多的内容。逻辑回归和线性回归在处理方法上也是有相同之处的，学习理解相对容易一点。学习正规化，使得线性回归和逻辑回归更加准确的拟合样本数据，通过添加正规化项来避免欠拟合和过拟合两种情况。下周学习进度暂停，首先是回顾前三周的知识，编程作业重新回顾，没做出来的再思考；然后对前三周知识进行总结，解决课程中遇到的疑问；下周毕业设计开题答辩和考试。