线性回归

优达学城

**什么是线性方程（Linear Equation）？**

描述直线的方程：，这里是直线的斜率，是与轴的交点。请注意这里方程的幂次是1。当我们在更高维度上讨论线性方程时我们都是在说一次方程。例如：。请注意，这是一个线性方程并且是一个描述平面的等式，而**不是一条线**。

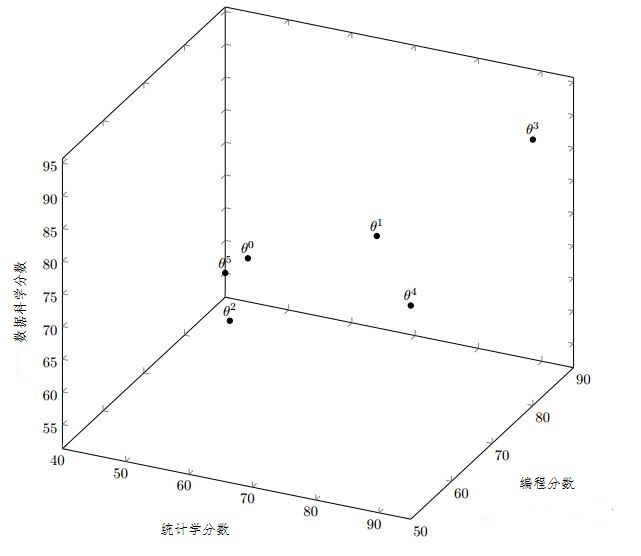
**什么是回归分析（Regression Analysis）？**

回归分析是统计学中的概念。这里的思路是通过观察数据来建立一个方程，从而让我们能够对缺失的数据或未来的数据做出预测。

让我们来看一下下面这些展示了统计学、编程以及数据科学等课程的测验分数（从100个数据选出）的数据。

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 统计学 | 编程 | 数据科学 |
| A | 50 | 80 | 65 |
| B | 80 | 65 | 83 |
| C | 60 | 60 | 69 |
| D | 95 | 80 | 92 |
| E | 95 | 50 | 84 |
| F | 40 | 90 | 55 |

我们试着来给这个数据建模。那样的话，下一次我们知道学生的统计学和编程分数的时候，我们就可以预测他们的数据科学分数。这也将会对那些想要知道在每个主题上应该投入多少来学好数据科学的学生有所脾益。



**什么是线性回归（Linear Regression）？**

线性回归是试图为因变量（Dependent Variable,（））和自变量（Independent Variable，（））之间的线性关系建模。

换句话说，我们想要写出这种类型的方程：

这里是输出变量（output variable），是输入变量（input variables），而被称作模型的参数（parameters）或权重（weights）。

所以在上述的分数数据集中，是数据科学分数（我们想要预测的数值）。是统计学分数，而是编程分数。等式就变成这样：

科学分数

数据科学分数

常数

常数

常数

统计学

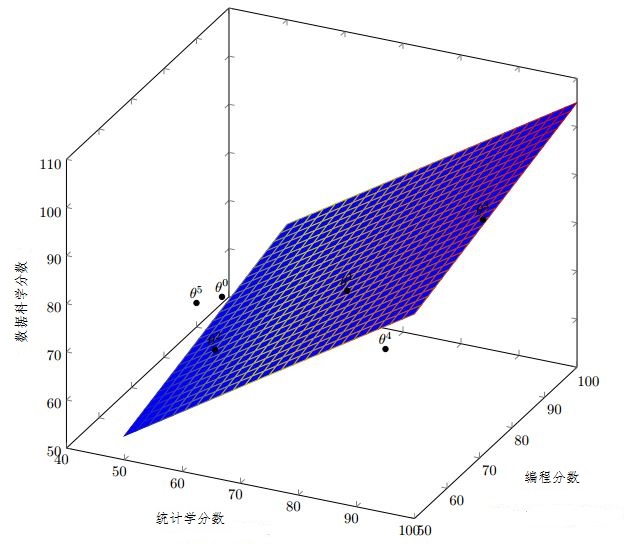
编程

**为什么把这些称作权重？**

简单来说是因为这些告诉了我们在预测输出的过程中相对应的有多重要。这就意味着，如果一个特定的值相比于其他的值来说非常小，相对应的在预测输出的角色中就非常小。

例如，假设这个分数方程是。请注意统计学的权重是而编程学的权重是。这意味着学生如果想在数据科学中考的好，就得在统计学上多下点功夫。

注意：我们一会儿将学习如何寻找这样的等式。



**什么是误差（Error）？**

如果我们能精确地预测输出，那就实在是太棒了。然而，在现实世界中，事情并不会总是这样。我们的数据可能不能完美地对应于我们的线性模型。这给我们的模型引入了误差。

让我们设是预测的输出，也就是我们用以上线性方程得到的输出。

计算误差的方法很多，两个最标准的方法是：

**绝对误差和**

**方差和** （我们乘以是为了计算方便，在稍后的梯度下降算法中我们将要求出这个表达式的导数，第2部分会消去分母中的2）

**为什么梯度下降对于方差和有效？**请记住，梯度下降算法利用函数的导数来进行最小化。将误差平方化以后能使这个误差函数变得可微。我们能很容易地求出这个函数的导数。

让我们再次思考分数的例子。假设我们选择处理的模型方程是。以下是关于预测值和观察值的表格。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 统计学 | 编程 | 数据科学 | 数据科学观察值 | 预测值 |
| A | 50 | 80 | 65 | 65 | 62.5 |
| B | 80 | 65 | 83 | 83 | 81.25 |
| C | 60 | 60 | 69 | 69 | 65 |
| D | 95 | 80 | 92 | 92 | 96.25 |
| E | 95 | 50 | 84 | 84 | 88.75 |
| F | 40 | 90 | 55 | 55 | 57.5 |

那么，绝对误差和=19.75，方差和=72.1875。

**我们用误差做什么？**

我们的目标是最小化误差。换句话说，我们希望找到能给出最小可能误差的值。这是线性模型中方差和的形式。预测值=

方差和

再次让我们回到我们的学生分数这个例子看一看当时方差和会变成什么。

方差和

我们的目标是找到使上面的方差和最小化的值。常用的一个办法是使用微积分。这就是能够利用梯度下降法的地方。同时请注意，对这个误差函数求导是非常容易的。所以请看一看梯度下降法的文档以后再回到这里找出能拟合我们的数据的线性方程。