**第二章 线性回归**

回归（Regression）的目的就是预测数值型的目标值。

比如根据父母的身高来预测孩子的身高，

式中，表示孩子身高，和分别表示父母的身高，为参数。这就是所谓的回归方程（Regression Equation）。

一旦有了参数，给定输入和，做预测就非常容易了。

回归和预测一点关系都没有，那为什么称为回归呢？

回归是由达尔文（Charles Darwin）的表兄弟Francis Galton发明的。Galton于1877年完成了第一次回归预测，目的是根据上一代豌豆种子（双亲）的尺寸来预测下一代豌豆种子（孩子）的尺寸。Galton在大量对象上应用了回归分析，甚至包括人的身高。他注意到：如果双亲的高度比平均值高，他们的子女也倾向于比平均值高，但尚不及双亲。孩子的高度向平均高度回退（回归）。Galton在多项研究上都注意到这个现象。所以尽管英文回归这个单词和数值预测没有任何关系，但这种研究方法仍被称作回归（Ian Ayres， 2008）。

Ian Ayres, Super Crunchers, Bantam Books, 2008, 24.

线性回归模型虽然很简单，但是它是最重要的数学模型之一。

有一个笑话：

1. 数学家觉得自己受够了数学，于是他去消防队去宣布他想去当消防员。

消防队长说：“您看上去不错，可是我得先给您一个测试。”消防队长带数学家到消防队后院小巷，巷子里有一个货栈、一只消防栓和一卷软管。

消防队长问：“假设货栈起火，您怎么办？”

数学家回答：“我把消防栓接到软管上，打开水龙头，把火浇灭。”

消防队长说：“完全正确。最后一个问题：假如您走进小巷，而货栈没有起火，您怎么办？“

数学家疑惑地思索了半天，终于答道：“我就把货栈点着。“消防队长大叫起来：”太可怕了！您为什么要把货栈点着？“

数学家说：“这样我就把问题化简为一个我已经解决过的问题了。“

线性回归的应用：宏观经济领域的IS-LM模型。

IS模型，即投资-储蓄模型（Investment-Saving），它研究的是国民储蓄和银行利率的关系。LM模型，即流动性偏好-货币供给模型（Liquidity Preference-Money Supply），它研究的是国民收入与货币需求量之间的关系。

这两个模型都涉及到两个变量：表示国民收入，可以简单理解为GDP；表示利率。

IS模型：。

对于投资储蓄行为，假设在短期内，国民收入增加，这意味着人们用于储蓄的资金增加，市场上的利率会下降。

LM模型：。

对于货币供给，假设在短期内，央行的货币供应量是不变的，而且国民收入增加。由于收入增加，人们对货币的需求是增加的，而货币量并没有增加，这会抬高市场上的利率。

将以上两个模型结合起来，就可以得到市场上的利率及对应的国民收入。

**线性回归模型**：

式中，：因变量（Dependent Variables），响应变量（Response Variables），标签（Label）；

：自变量（Independent Variables），解释变量(Explanatory Variables)，预测变量（Predictive Variable），特征（Features）；

为未知参数，回归系数（Regression Coefficient）。

个独立的观测数据，

式中，。

上式可以写成矩阵形式：

**参数学习**：

平方损失函数(Quadratic Loss Function)为：

对的偏导数是：

令，即。

为的最优解。以上求解线性回归的方法称为普通最小二乘法（Ordinary Least Square）估计。上述公式中，需要计算的逆矩阵，因此需要判断矩阵是否可逆。

为的无偏估计，即

**梯度下降法（Gradient Descent）**：

当不可逆时，也可以通过梯度下降法来求解。

梯度下降法的基本思想是：要找到某函数的最小值，最好的办法是沿着该函数的梯度反方向探寻。相反，要找到某函数的最大值，最好的办法是沿着该函数的梯度方向探寻（即梯度上升法）。

梯度下降法作为机器学习中较常使用的优化算法，有三种不同的形式：批量梯度下降（Batch Gradient Descent， BGD）、随机梯度下降（Stochastic Gradient Descent， SGD）和小批量梯度下降（Mini-Batch Gradient Descent，MBGD，有时候也被称为SGD）。其中小批量梯度下降法也常用在深度学习中进行模型的训练。

**批量梯度下降**：

平方损失函数(Quadratic Loss Function)为：

对的偏导数是：

由梯度下降法，我们有

1. 初始化；
2. 沿着负梯度方向迭代，即

其中，是学习率。

优点：全局最优解，能保证每一次更新权值，都能降低损失函数；易于并行实现。

缺点：当样本数目很大时，每迭代一步都需要对所有样本计算，训练过程会很慢。

**随机梯度下降法**：

梯度下降算法在每次更新回归系数时都需要遍历整个数据集，该方法在处理100个左右的数据集时尚可，但如果有数十亿和成千上万的特征，那么该算法的计算复杂度就太高了。一种改进方法是一次仅用一个样本点来更新回归系数。该方法称为随机梯度下降法。

优点：训练速度快（由于不是在全部训练数据上的损失函数，而是在每轮迭代中，随机优化某一条训练数据上的损失函数，这样每一轮参数的更新速度大大加快）。

缺点：准确度下降，并不是全局最优；不易于并行实现。

**小批量梯度下降法**：

其中，。

是前两者的折中，但是参数（batch size）的不当选择可能会带来问题。

**极大似然估计方法（Maximum Likelihood Estimate，MLE）**：

极大似然估计是建立在极大似然原理的基础上的一个统计方法。极大似然估计提供了一种给定观察数据来评估模型参数的方法，即：“模型已定，参数未知”。通过若干次试验，观察其结果，利用试验结果得到某个参数值能够使样本出现的概率为最大，则称为极大似然估计。

极大似然估计的一般步骤：

（1） 写出似然函数；

（2） 求对数似然函数；

（3） 对对数似然函数求导；

（4） 解似然方程 。

回归模型：

我们得到。

我们有个独立的观测数据。

（1）写出似然函数：

（2）对取对数得到：

（3）对对数似然函数求偏导：

（4）令，我们可以求得的极大似然估计是

**岭回归（Ridge Regression）**：

如果数据的特征比样本点还多（即），这时候矩阵不可逆，怎么办？为了解决这个问题，统计学家们引入了岭回归的概念。

岭回归就是在矩阵上加一个从而使得矩阵非奇异，进而能对矩阵求逆，其中，矩阵是一个单位阵，即对角线的元素全为1，其余元素全为0。

岭回归的解可以看作是

的最小二乘解。

上式中的对的偏导数是：

即可求得岭回归的解。

矩阵半正定。。

矩阵正定。。

。所以，矩阵可逆。

岭回归的岭是什么？

岭回归使用了单位矩阵乘以常量，我们观察其中的单位矩阵，可以看到值1贯穿整个对角线，其余元素全是0。形象地，在0构成的平面上有一条1组成的“岭”，这就是岭回归的”岭”的由来。

**多项式回归**：

线性回归的局限性在于其只能解决存在线性关系的数据的拟合问题。如果数据之间存在非线性关系，虽然仍可以用线性回归来解决问题，但是效果将会很差。这时候就需要对线性回归模型进行改进，使之能够更好地拟合非线性数据。

多项式回归（Polynomial Regression）是研究一个因变量与一个或多个自变量间多项式的回归分析方法。

多项式回归在回归分析中很重要，因为任意一个连续函数至少在一个紧集内都可以用多项式任意逼近。

线性回归：

二阶多项式回归：

三阶多项式回归：

令，我们就可以得到线性回归模型

**评估模型效果**：

1. 定义均方差（Mean of Squared Error， MSE）

越小，模型的效果越好。

1. 模型的决定系数（Coefficient of Determination），通常记为，可衡量回归方程整体拟合度的好坏。

残差（Residual）：实际观察值和估计值之间的差，。

离差（Deviation）：一个特定数值对于其平均值的偏离，。

我们有：

其中称为已解释离差(Explained deviation)，因为它是由倚的回归关系决定的。称为未解释离差(Unexplained deviation)，因为它随机变化，与倚的回归关系无关。

, Total sum of squares about mean

, Regression sum of squares

, Residual (error) sum of squares

即，两边除以得到

决定系数等于因变量的方差可由自变量解释的比例。

, 说明模型能够很好地拟合数据；

, 说明模型不能够较好地拟合数据。

**Python的应用**

制作提供信息的可视化（有时称为绘图）是数据分析中的最重要任务之一。

可视化是探索性过程的一部分，例如可以帮助识别异常值或所需的数据转换，或者为建模提供一些想法。

1. **箱形图**

箱形图统计学知识

分位数（Quantile），亦称分位点。四分位数，Quartile。

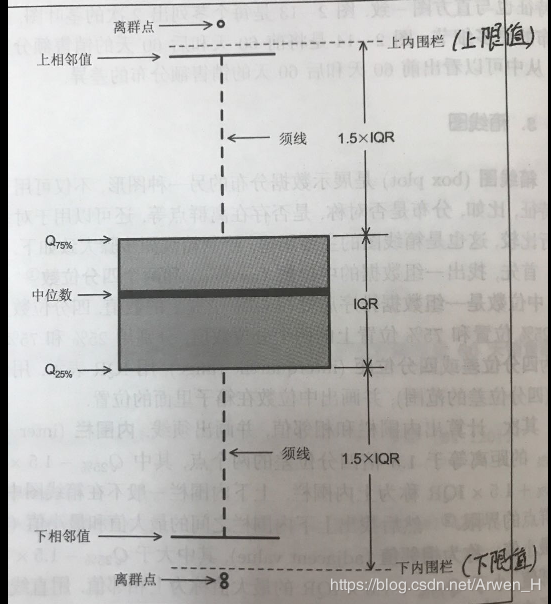
分位数指的就是连续分布函数中的一个点，这个点对应概率。若概率，随机变量或它的概率分布的分位数，是指满足条件的实数

1）第一四分位数(Quartile，Q1)，又称“较小四分位数”，等于该样本中所有数值由小到大排列后第25%的数字；

2）第二四分位数(Q2)，又称“中位数”，等于该样本中所有数值由小到大排列后第50%的数字；

3）第三四分位数(Q3)，又称“较大四分位数”，等于该样本中所有数值由小到大排列后第75%的数字。

第三四分位数与第一四分位数的差距又称四分位距（Interquartile Range， IQR）。



上限值：Q1-1.5×IQR

上相邻值：距离上限值最近的值

须线：上下分位数各自与上下相邻值的距离

上四分位数（Q1)：一组数据按顺序排列，从小至大第25%位置的数值

中位数：一组数据按顺序排列，从小至大第50%位置的数值

四分位距（Interquartile Range, IQR）：Q3-Q1上四分位数至下四分位数的距离

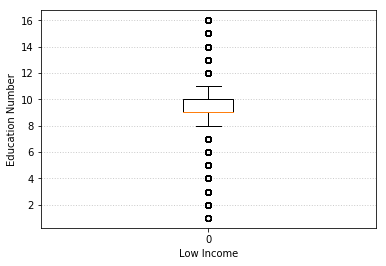
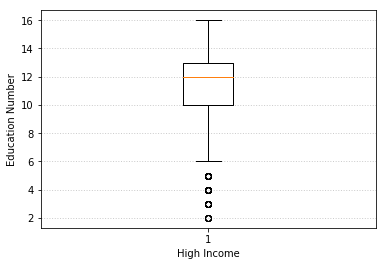
下四分位数（Q3）：一组数据按顺序排列，从小至大第75%位置的数值

下相邻值：距离下限值最近的值

下限值：Q3+1.5×IQR

离群值（异常值）：一组数据中超过上下限的真实值

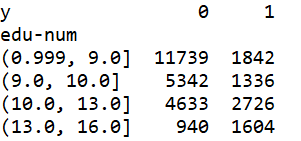
原文链接：<https://blog.csdn.net/Arwen_H/article/details/84855825>

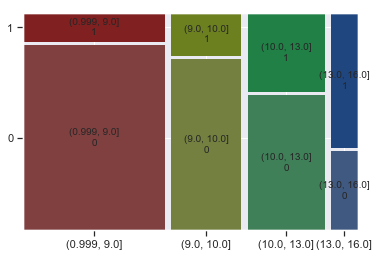
不同收入水平的受教育年限箱形图

**2. 交叉报表**

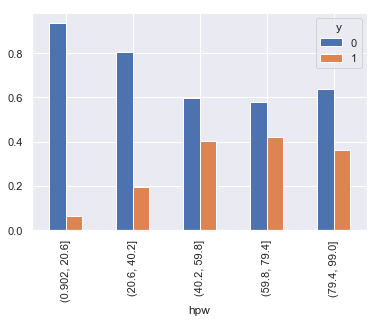
交叉报表是一种常用的数据分析方法。它可以描述两个变量之间的关系。



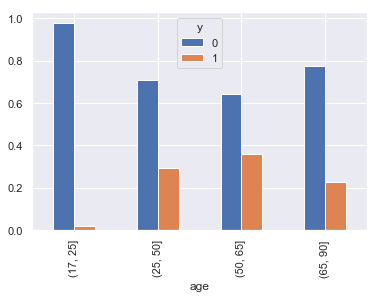
受教育年限和收入水平的交叉报表



受教育年限和收入水平的交叉报表



每周工作时间和收入水平的交叉报表



年龄和收入水平的交叉报表

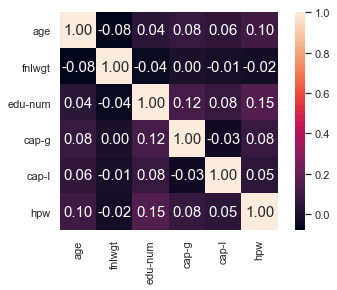
**3. 热力图**

相关系数（Correlation coefficient）：研究变量之间线性相关程度的量。

相关系数为1：表示两个随机变量完全线性正相关；

相关系数为−1：表示两个随机变量完全线性负相关。

利用热力图可以看数据表里多个特征两两的相似度。



相关系数的热力图

Adult数据集简要说明

|  |  |  |
| --- | --- | --- |
| 变量名 | 变量类型 | 说明 |
| age | 数值型变量 | 年龄 |
| workclass: wc | 类别型变量（共6类） | 工作类型，如公务员、私企职工等 |
| fnlwgt | 数值型变量 | 抽样权重（普查时使用的变量，与建模无关） |
| education: edu | 类别型变量（共16类） | 学历，如本科、研究生等 |
| education-num: edu-num | 数值型变量 | 受教育年限 |
| martial-status: ms | 类别型变量（共7类） | 婚姻状况 |
| occupation: occ | 类别型变量（共14类） | 所在行业 |
| relationship: rs | 类别型变量（共6类） | 家庭加色，比如妻子、丈夫 |
| race | 类别型变量（共5类） | 种族 |
| sex | 类别型变量（共2类） | 性别 |
| capital-gain: cap-g | 数值型变量 | 年度投资收益 |
| capital-loss: cap-l | 数值型变量 | 年度投资损失 |
| hours-per-week: hpw | 数值型变量 | 每星期工作时间 |
| native-country: native | 类别型变量（共42类） | 出生国家 |
| label: y | 类别型变量（共2类） | 年收入分类，“”和“” |

建立一个模型来预测一个人的收入水平（高、低收入），并分析哪些因素对一个人的收入起决定性的多用。