数据分析（基本方法）

- 相关性分析和回归分析
  - 相关性分析
    - 反映变量之间是否存在关联。例如，我们可以用相关分析来探究人的身高和体重之间是否有关系，或者气温和冰淇淋销量之间是否有关系。
    - 相关分析包括皮尔逊分析、斯皮尔曼分析、肯德尔分析。
  - 回归分析
    - 研究一个变量（因变量）如何受到一个或多个变量（自变量）影响的方法。例如，我们可以用回归分析来预测人的身高（因变量）和年龄（自变量）之间的关系，或者气温（因变量）和季节（自变量）之间的关系。回归分析的结果通常用回归方程来表示，回归方程是一个描述因变量和自变量之间关系的数学公式，可以用来估计或预测因变量的值。
    - 回归分析包括线性回归分析、多元回归和非线性回归。
  - 区别和联系
    - 相关分析只能反映变量之间是否存在关联，以及关联的方向和程度，但不能说明变量之间的因果关系。回归分析则可以反映变量之间的因果关系，以及因变量如何受到自变量的影响。
    - 相关分析是回归分析的基础和前提，回归分析则是相关分析的深入和继续。只有当变量之间存在高度相关时，进行回归分析寻求其相关的具体形式才有意义。如果在没有对变量之间是否相关以及相关方向和程度做出正确判断之前，就进行回归分析，很容易造成“虚假回归”。
- 一般回归分析方法区别

  估计回归方程参数的方法：最小二乘法或者是极大似然估计
  - （多元）线性回归

    即Y=a+b*X + e，一般利用最小二乘法来确定（a和b的值），用来刻画回归方程的性能。
    - 使用该方法需要注意的要点
      - 线性回归对异常值非常敏感。它会严重影响回归线，最终影响预测值。所以前期的数据预处理需要做好
      - 自变量与因变量之间必须有线性关系
      - 多元回归存在多重共线性，自相关性和异方差性。（？）
      - 在多个自变量的情况下，我们可以使用向前选择法，向后剔除法和逐步筛选法来选择最重要的自变量。
  - Logistic 回归/逻辑回归

    逻辑回归是用来计算“事件=Success”和“事件=Failure”的概率。当因变量的类型属于二元（1 / 0，真/假，是/否）变量时，我们就应该使用逻辑回归。
    - 使用该方法需要注意的要点

        - 它广泛的用于分类问题。
        - 逻辑回归不要求自变量和因变量是线性关系。它可以处理各种类型的关系，因为它对预测的相对风险指数OR使用了一个非线性的log转换。
        - 为了避免过拟合和欠拟合，我们应该包括所有重要的变量。有一个很好的方法来确保这种情况，就是使用逐步筛选方法来估计逻辑回归。
        - 它需要大的样本量，因为在样本数量较少的情况下，极大似然估计的效果比普通的最小二乘法差。
        - 自变量不应该相互关联的，即不具有多重共线性。然而，在分析和建模中，我们可以选择包含分类变量相互作用的影响。
        - 如果因变量的值是定序变量，则称它为序逻辑回归。
        - 如果因变量是多类的话，则称它为多元逻辑回归。
- 多项式回归

    对于一个回归方程，如果自变量的指数大于1，那么它就是多项式回归方程。例如：y=a+b*x^2

    - 使用该方法需要注意的要点
        - 易导致过拟合，注意判断选择
- Ridge岭回归

    岭回归分析是一种用于存在多重共线性（自变量高度相关）数据的技术。在多重共线性情况下，尽管最小二乘法（OLS）对每个变量很公平，但它们的差异很大，使得观测值偏移并远离真实值。岭回归通过给回归估计上增加一个偏差度，来降低标准误差。

    - 介绍
        - 在线性回归方程y=a+ b*x中增加误差项e，即：y=a+b*x+e，在一个线性方程中，预测误差e可以分解为偏差和方差。岭回归通过收缩参数λ解决多重共线性问题。

            $$= \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

            在这个公式中，有两个组成部分。第一个是最小二乘项，另一个是β2（β-平方）的λ倍，其中β是相关系数。为了收缩参数把它添加到最小二乘项中以得到一个非常低的方差
    - 使用该方法需要注意的要点
        - 除常数项以外，这种回归的假设与最小二乘回归类似；
        - 它收缩了相关系数的值，但没有达到零，这表明它没有特征选择功能
        - 这是一个正则化方法，并且使用的是L2正则化。
- 回归正则化方法（Lasso，Ridge和ElasticNet）在高维和数据集变量之间多重共线性情况下运行良好。
- 正则化是用来防止模型过拟合的一种手段
- 正则化详细解释见：[一篇文章完全搞懂正则化（Regularization）-CSDN博客]
- 回归拟合程度的一个讨论

    - 欠拟合原因：
        - 训练样本数量少
        - 模型复杂度过低
        - 参数还未收敛就停止循环
    - 欠拟合的解决办法：
        - 增加样本数量
        - 增加模型参数，提高模型复杂度
        - 增加循环次数
        - 查看是否是学习率过高导致模型无法收敛
    - 过拟合原因：
        - 数据噪声太大
        - 特征太多
        - 模型太复杂
    - 过拟合的解决办法：
        - 清洗数据
        - 减少模型参数，降低模型复杂度
        - 增加惩罚因子（正则化），保留所有的特征，但是减少参数的大小（magnitude）。
- 聚类分析

  利用多个数据项之间的相似性对数据对象进行聚类

    - 聚类的应用场景
        - 聚类：将数据集中的对象划分为多个组，使得同一组内的对象之间的相似性高，而与其他组的对象相似性低。
        - 聚类质量：用于评估聚类结果的标准，如内部距离、间距等。
        - 聚类算法：用于实现聚类分析的方法，如K均值聚类、DBSCAN、AGNES等。
    - 如何评定聚类的质量
        - 可以通过内部距离、间距、隶属度等指标进行评估。
    - 聚类于其他数据挖掘技术的联系
        - 聚类分析与分类相比，前者没有先前的标签信息，而后者有标签信息。（特征值）
        - 聚类分析与主成分分析(PCA)相比，前者关注数据之间的相似性，后者关注数据的线性组合。
        - 聚类分析与倾向分析相比，前者关注数据的分组，后者关注数据的特征值。
- 主成分分析

  分析数据的主要成分达到**数据降维**的效果，也就是抓住主要矛盾，相对保留信息达到最多

    - PCA主要概念
        - 方差即信息，方差越大表示保留的信息越多。诸如将二维空间中的离散的点向一维空间中投影，不同的投影方式点的投影位置不同，有些投影的角度非常积聚，有些则非常离

> 散，越离散表示保留的信息越多，也就是方差越大。

- 学习参考：学习笔记 | 主成分分析(PCA)及其若干应用_主成分分析算法应用-CSDN博客