线性回归算法详解

- 欠拟合：拟合的函数和训练集误差较大，我们称这种情况为 欠拟合
  - 欠拟合问题，根本的原因是特征维度过少，导致拟合的函数无法满足训练集，误差较大。欠拟合问题可以通过增加特征维度来解决
- 合适拟合：拟合的函数和训练集误差较小，我们称这种情况为 合适拟合
- 过拟合：拟合的函数完美的匹配训练集数据，我们称这种情况为 过拟合
  - 过拟合问题，根本的原因则是特征维度过多，导致拟合的函数完美的经过训练集，但是对新数据的预测结果则较差。
  - 解决过拟合问题，则有2个途径：
    - 减少特征维度; 可以人工选择保留的特征，或者模型选择算法
    - 正则化; 保留所有的特征，通过降低参数θ的值，来影响模型（正则化（Regularization）包括L1、L2（L2 regularization也叫weight decay））
      - L2 regularization（权重衰减） --- 正则化就是在代价函数后面再加上一个正则化项(即所有权重w的绝对值的平方的和，乘以λ/2*n)
        - 更小的权值w，从某种意义上说，表示网络的复杂度更低，对数据的拟合刚刚好（这个法则也叫做奥卡姆剃刀）
      - L1 regularization 在原始的代价函数后面加上一个L1正则化项，即所有权重w的绝对值的和，乘以λ/n
        - 让w往0靠，使网络中的权重尽可能为0，也就相当于减小了网络复杂度，防止过拟合。
- 回归问题：主要用于预测数值型数据，典型的回归例子：数据拟合曲线 ，回归算法与分类算法都属于监督学习算法，不同的是，分类算法中标签是一些离散值，代表不同的分类，而回归算法中，标签是一些连续值，回归算法需要训练得到样本特征到这些连续标签之间的映射
- 1.什么是回归分析?
  - 回归分析是研究自变量和因变量之间关系的一种预测模型技术。这些技术应用于预测，时间序列模型和找到变量之间关系。例如可以通过回归去研究超速与交通事故发生次数的关系
- 2.我们为什么要使用回归分析?
  - 回归分析是研究自变量和因变量之间关系的一种预测模型技术。这些技术应用于预测，时间序列模型和找到变量之间关系。例如可以通过回归去研究超速与交通事故发生次数的关系
- 3.回归有哪些类型？
  - 三种方法分类：自变量的个数、因变量的类型和回归线的形状
  - 1.线性回归(Linear Regression)
    - 意义：线性回归是指全部由线性变量组成的回归模型
    - 重点：
      - 1.自变量与因变量之间必须要有线性关系。
      - 2.多重共线性、自相关和异方差对多元线性回归的影响很大。
      - 3.线性回归对异常值非常敏感，其能严重影响回归线，最终影响预测值。

        - 4.在多元的自变量中，我们可以通过前进法，后退法和逐步法去选择最显著的自变量。
    - 特点：
        - 1. 建模速度快，不需要很复杂的计算，在数据量大的情况下依然运行速度很快。
        - 2. 可以根据系数给出每个变量的理解和解释
        - 3. 对异常值很敏感
    - 参数:
        - 最小二乘法就是线性回归模型的损失函数，只要把损失函数做到最小时得出的参数，才是我们最需要的参数
- 2.逻辑回归(Logistic Regression)
    - 意义：逻辑回归是用来找到事件成功或事件失败的概率。当我们的因变量是二分类（0/1，True/False，Yes/No）时我们应该使用逻辑回归
    - 重点：
        - 1.在分类问题中使用的非常多。
        - 2.逻辑回归因其应用非线性log转换方法，使得其不需要自变量与因变量之间有线性关系。
        - 3.为防止过拟合和低拟合，我们应该确保每个变量是显著的。应该使用逐步回归方法去估计逻辑回归。
        - 4.逻辑回归需要大样本量，因为最大似然估计在低样本量的情况下表现不好。
        - 5.要求没有共线性。
        - 6.如果因变量是序数型的，则称为序数型逻辑回归。
        - 7.如果因变量有多个，则称为多项逻辑回归。
    - 优点：
        - 1）适合需要得到一个分类概率的场景。
        - 2）计算代价不高，容易理解实现。LR在时间和内存需求上相当高效。它可以应用于分布式数据，并且还有在线算法实现，用较少的资源处理大型数据。
        - 3）LR对于数据中小噪声的鲁棒性很好，并且不会受到轻微的多重共线性的特别影响。（严重的多重共线性则可以使用逻辑回归结合L2正则化来解决，但是若要得到一个简约模型，L2正则化并不是最好的选择，因为它建立的模型涵盖了全部的特征。）
    - 缺点：
        - 1）容易欠拟合，分类精度不高。
        - 2）数据特征有缺失或者特征空间很大时表现效果并不好。
- 3.多项式回归(Polynomial Regression)
    - 意义：线性回归适合于线性可分的数据，当我们处理非线性可分的数据时可以使用多项式回归
    - 重点：
        - 1.很多情况下，我们为了降低误差，经常会抵制不了使用多项式回归的诱惑，但事实是，我们经常会造成过拟合。所以要经常的把数据可视化，观察数据与模型的拟合程度。
        - 2.特别是要看曲线的结尾部分，看它的形状和趋势是否有意义。高的多项式往往会产生特别古怪的预测值。

- 特点:
  - 1. 能够拟合非线性可分的数据，更加灵活的处理复杂的关系
  - 2. 因为需要设置变量的指数，所以它是完全控制要素变量的建模
  - 3. 需要一些数据的先验知识才能选择最佳指数
  - 4. 如果指数选择不当容易出现过拟合

- 4.逐步回归
  - 意义：
    - 逐步回归是一种线性回归模型自变量选择方法，其基本思想是将变量一个一个引入，引入的条件是其偏回归平方和经验是显著的。同时，每引入一个新变量后，对已入选回归模型的老变量逐个进行检验，将经检验认为不显著的变量删除，以保证所得自变量子集中每一个变量都是显著的。此过程经过若干步直到不能再引入新变量为止。这时回归模型中所有变量对因变量都是显著的。
  - 总结变量选择
    - 1）子集选择 这是传统的方法，包括逐步回归和最优子集法等，对可能的部分子集拟合线性模型，利用判别准则 （如AIC,BIC,Cp,调整R2 等）决定最优的模型
    - 2）收缩方法（shrinkage method） 收缩方法又称为正则化（regularization）。主要是岭回归（ridge regression）和lasso回归。通过对最小二乘估计加入罚约束，使某些系数的估计为0。（岭回归：消除共线性；模的平方处理；Lasso回归：压缩变量，起降维作用；模处理）
    - (3)维数缩减 主成分回归（PCR）和偏最小二乘回归（PLS）的方法。把p个预测变量投影到m维空间（m<p），利用投影得到的不相关的组合建立线性模型。
  - 子集选择
    - 1.最优子集选择（best-subset selection）
    - 2.前向/后向逐步选择（forward/backwards-stepwise selection）
    - 3.前向分段回归（Forward-Stagewise Regression）

- 5.岭回归(Ridge Regression)
  - 意义： 岭回归(ridge regression, Tikhonov regularization)是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法
  - 岭回归用于处理下面两类问题：
    - 1.数据点少于变量个数
    - 2.变量间存在共线性
      - 变量间存在共线性是，最小二乘回归得到的系数不稳定，方差很大，这是因为系数矩阵x与它的转置矩阵相乘得到的矩阵不能求逆
  - 原理 ：在平方误差的基础上增加正则项 即所有权重w的绝对值的平方的和，乘以λ
    - 通过确定的值可以使得在方差和偏差之间达到平衡：随着的增大，模型方差减小而偏差增大 对求导，结果为令其为0，可求得W 的值
  - 特点:
    - 1. 领回归的假设和最小平方回归相同，但是在最小平方回归的时候我们假设数据服从高斯分布使用的是极大似然估计(MLE)，在领回归的时候由于添加了

偏差因子，即w的先验信息，使用的是极大后验估计(MAP)来得到最终的参数
    - 2. 没有特征选择功能
  - 重点：
    - 1.岭回归的假设与最小二乘法回归的假设相同除了假设正态性。
    - 2.它把系数的值收缩了，但是不会为0.
    - 3.正则化方法是使用了l2正则.
  - 优点：
    - 可对变量之间共线性比较严重或病态数据偏多的数据类型作回归分析，对这类数据作回归得到的回归系数更符合实际，更可靠。另外，岭回归能让估计参数的波动范围变小，变的更稳定。
  - 缺点：
    - 对系数的估计时，会损失部分信息、降低精度。同时岭回归方程的R平方值会稍低于普通的回归方法。
    - 通常岭回归方程的R平方值会稍低于普通回归分析，但回归系数的显著性往往明显高于普通回归，在存在共线性问题和病态数据偏多的研究中有较大的实用价值。
- 6.套索回归算法 Lasso回归
  - 意义：Lasso回归有时也叫做线性回归的L1正则化，和Ridge回归的主要区别就是在正则化项，Ridge回归用的是L2正则化，而Lasso回归用的是L1正则化
  - 原理： Lesso与岭回归非常相似，都是在回归优化函数中增加了一个偏置项以减少共线性的影响，从而减少模型方程。不同的是Lasso回归中使用了绝对值偏差作为正则化项
  - 重点：
    - 1.其假设与最小二乘回归相同除了正态性。
    - 2.其能把系数收缩到0，使得其能帮助特征选择。
    - 3.这个正则化方法为l1正则化。
    - 4.如果一组变量是高度相关的，lasso会选择其中的一个，然后把其他的都变为0.
- 7.弹性网络回归(ElasticNet Regression)
  - 意义： 弹性网络回归算法是 套索回归算法和岭回归算法的混合体, 在训练模型时，综合使用L1 和 L2 两种正则化方法
  - 优点：
    - 1. 鼓励在高度相关变量的情况下的群体效应，而不像Lasso那样将其中一些置为0.当多个特征和另一个特征相关的时候弹性网络非常有用。Lasso倾向于随机选择其中一个，而弹性网络倾向于选择两个。
    - 2. 对所选变量的数量没有限制。
  - 重点：
    - 1.在选择变量的数量上没有限制
    - 2.双重收缩对其有影响

幕布 - 思维概要整理工具