# 线性回归

## 简单介绍线性回归

　是一种预测模型，利用各个特征的数值去预测目标值。线性回归的主要思想是给每一个特征分配一个权值，最终的预测结果是每个特征值与权值的乘机之和再加上偏置。所以训练的目标是找到各个特征的最佳权值和偏置，使得误差最小。线性回归的假设前提是噪声符合正态分布。

## 线性回归的损失函数 && 为什么使用

## 线性回归

# 求解方法

## 标准方程

$\theta = (X^T * X)^{-1} * X^T * y$

- 特点:
  - 要求X^T * X可逆
  - 适用于特征数量少的情况, 比如10000个一下
  - 时间复杂度约为O(n^3)

## 梯度下降

- 方法
  - 随机初始化θ_j的值
  - 针对损失函数J(θ), 分别计算θj的偏导数
  - 根据下式同时更新θ_j, 直到梯度下降为0, J(θ)收敛:

  $$\theta_j = \theta_j - a * \frac{\partial J(\theta)}{\partial \theta_j}$$

- 特点
  - 适用于特征数量比较大的训练集
  - 随机初始值的不同, 目标函数可能会收敛到不同的值, 即局部最优值(使用随机梯度下降可以解决这个问题).不过线性回归和逻辑回归的损失函数MSE都是凸函数, 只有一个全局最优值.
  - 超参数学习率a
    - 若太小, 则需要训练更长时间才能收敛
    - 若太大, 则可能左右横跳, 无法收敛
  - 需要对特征进行特征缩放
    - 不做特征缩放的话, 尽管也可以收敛, 但是会耗费更多的训练时间
    - 方法
      - 归一化(min-max normalization)

      特点: 对异常值敏感

- 标准化(Z-score Normalization)

- 种类
    - 批量梯度下降: 在每轮迭代计算梯度的时候需要使用全部的训练数据
        - 如果损失函数不规则, 不是凸函数, 有可能陷入局部最优解
    - 随机梯度下降: 在每轮迭代计算梯度的时候仅使用一个训练样本
        - 如果损失函数不规则, 不是凸函数, 可以跳出局部最优
        - 但是即使达到了全局最优点, 也无法收敛在那里, 因为它会一直跳动
        - 因此可以不断地改变学习率(步长)来进行优化, 这个过程叫做模拟退火
    - 小批量梯度下降:
        - 介于批量梯度下降和随机梯度下降之间, 即仅使用一部分数据进行梯度计算

# 多项式回归

遇到非线性数据的时候, 可以将每个特征的幂次方作为一个新的特征, 然后使用线性模型wx+b进行训练

- 特点
    - 可以处理非线性数据
    - 增加模型的复杂度
    - 特征的幂次方越大, 越有可能过拟合, 因此可以加入正则化罚项

# 正则线性模型

为了减少过拟合, 可以对模型进行正则化. 比如, 将多项式模型正则化的方简单方法就是降低多项式的阶数.

对线性模型而言, 正则化通常通过约束模型的权重来实现, 比如使高阶特征的权重约等于0, 即可使模型更简单.

## 岭回归

加入权重向量的L2范数的平方的一半

$$\lambda \sum_{i=1}^{n} \theta_i^2$$

- 特点
    - 超参数λ控制的是对模型进行正则化的程度.
    - 一般而言, 使用正则化的时候均需要进行特征缩放

# 套索回归(LASSO回归)

加入权重向量的L1范数

$$\lambda \sum_{i=1}^{n} |\theta_i|$$

- 特点
    - 比使用L2范数更能完全消除不重要的特征, 即特征权重置为0(L2范数只是让权重约等于0)

## 岭回归与套索回归的对比

岭回归是个不错的默认选择, 但是如果你认为时机用到的特征只有少数几个, 那就应该更倾向于套索回归或弹性网络, 因为它们会将无用特征的权重降为0.