线性回归

线性回归（Linear Regression）

- 回归主要通过学习特征值与预计算值间的定量关系来求解业务需求。依据特征值与预计算值之间的表达式呈现线性还是非线性，回归分为线性回归与非线性回归两种。线性回归是用于确定两种或两种以上特征间依赖的定量关系的一种统计分析方法。

- 线性回归算法是使用线性方程对数据集进行拟合的算法，是一个非常常见的回归算法。

- 线性回归能够用一条直线较为精确地描述数据之间的关系。当出现新数据的时候给出预测值。
  - 只包括一个自变量和一个因变量则成为一元线性回归分析，也叫简单回归
  - 包括两个或者两个以上的自变量，则成为多元线性回归分析

线性模型基本形式

- 线性回归拟合的目标是要得到输出向量fx和输入特征x之间的线性关系，
  - 即求解ω，截距b的值，尽量达到实际值与拟合值之间距离最小。

- 线性模型一般形式：

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

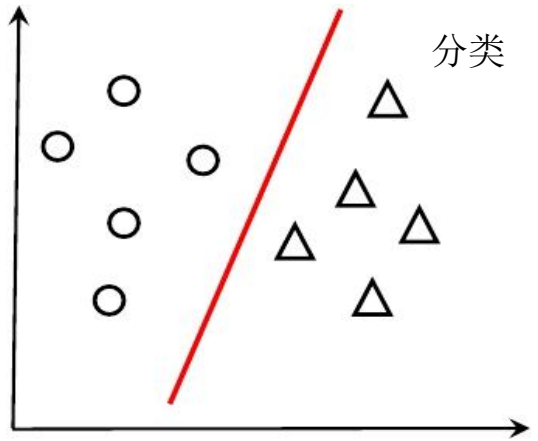
$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$ 是由属性描述的示例，其中 x_i 是 $\boldsymbol{x}$ 在第 i 个属性上的取值

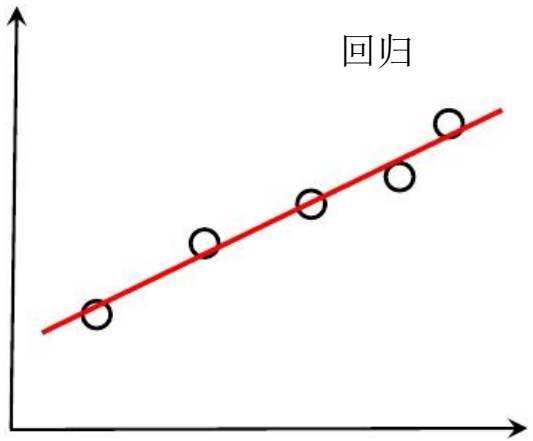
- 向量形式：

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$$

其中

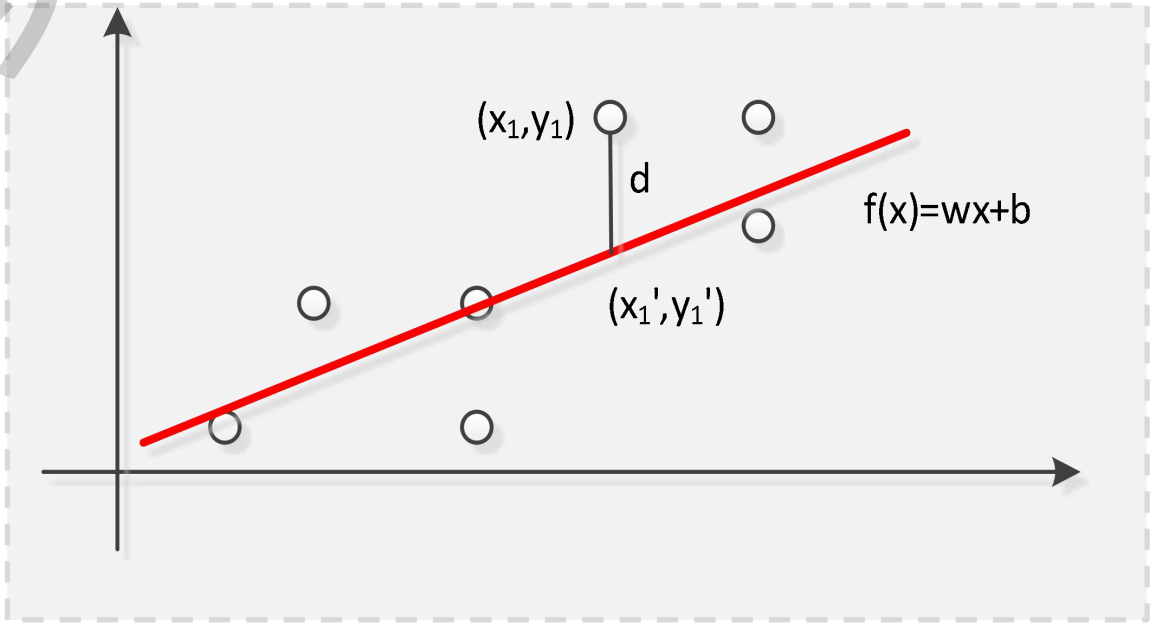
$$\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$$



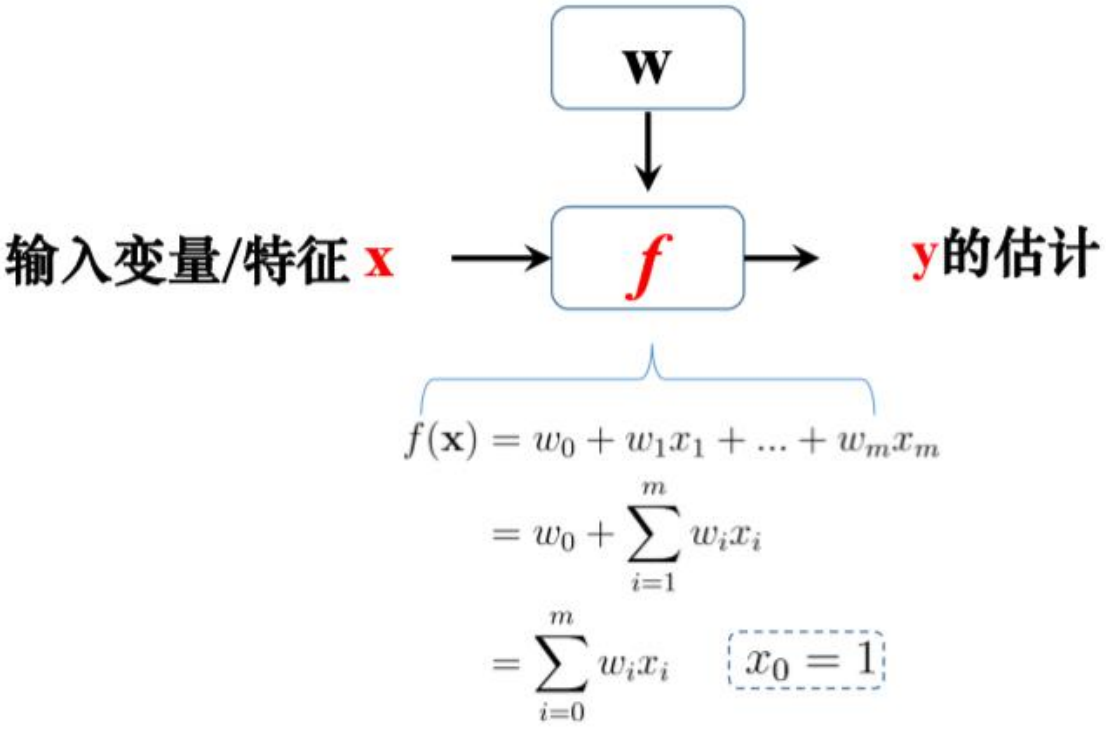


线性模型

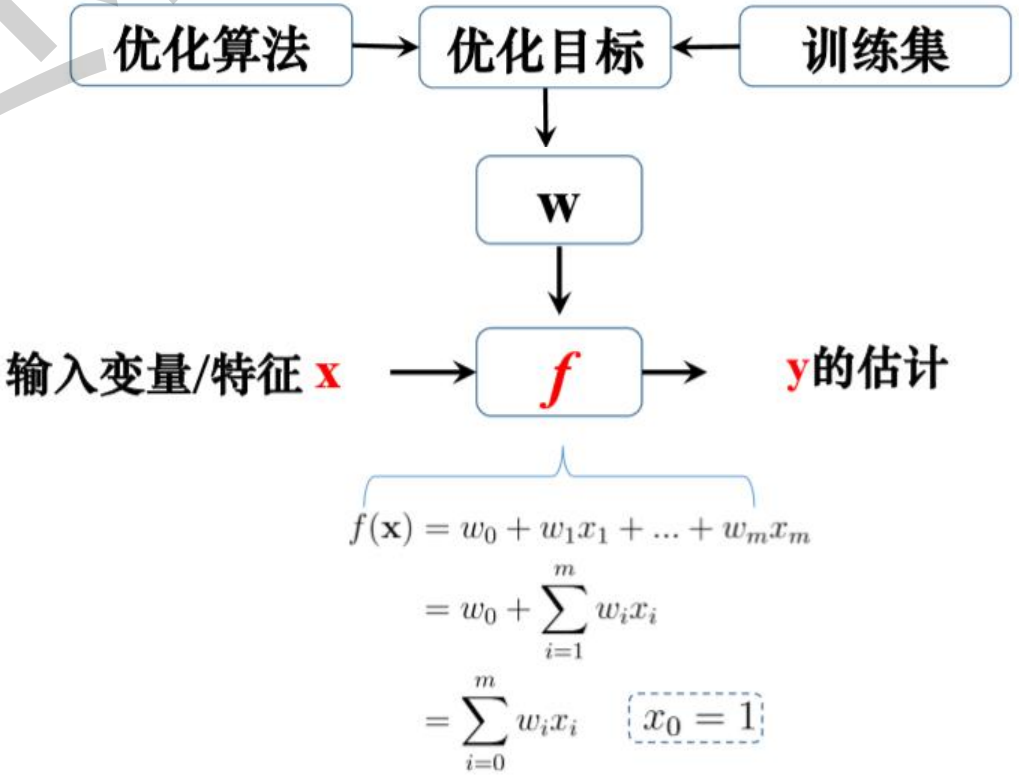
- 如图，演示了只有一个输入特征，即单变量线性回归模型。
  - 〇表示实际的值，
  - 直线表示拟合出的能够说明实际值趋势的直线fx=ωx+b，
  - d示意了实际值x1,x2与拟合出来的曲线对应的拟合值x1’,x2’之间的差
    - **如果每一个实际值与拟合值对应的距离最小**，即是求解出来的回归方程，即 fx=ωx+b。



- 这是线性模型的应用过程
- 其中，f(x)为训练得到的线形模型，
- 待预测数据是函数f(x)的参数，
- 计算f(x)将得到一个输出值y，
- 该输出值就是预测结果。



- 对于确定的输入数据，最终的预测结果取决于模型f(x)中的w和b
- 那么我们的学习过程就是设计优化目标，不断优化算法，以学习得到最好的w和b的过程。



了解线性回归代码

百度智能云

```
import numpy as np
from sklearn.linear_model import LinearRegression

if __name__ == "__main__":
  x_data = np.array([1,2,3,5,7])  # x轴原始数据
  y_data = np.array([4,8,9,10,19])  # y轴原始数据

  # 转一下维度，sklearn框架才能识别
  x_data = x_data[:, np.newaxis]
  y_data = y_data[:, np.newaxis]

  model = LinearRegression()  # 建立线性回归模型
  model.fit(x_data, y_data)  # 开始训练

  # 求解f(x)=wx+b模型
  print("w = ", model.coef_[0]," b= ", model.intercept_)
```

运行结果

w = [2.15517241] b= [2.24137931]

查看代码：线性回归方法朴素示例.ipynb

回归方程的评估

- 针对衡量线性回归方程的评价指标，常用的有:
  - 残差
    - 在数理统计中是指所有拟合数据与原始数据（拟合值）之间的差的和，蕴含了回归方程中的误差信息。当用来考察模型假设的合理性及数据的可靠性时，称为残差分析
  - MSE（均方误差）
    - 是拟合数据和原始数据对应点误差的平方和的均值，即SSE/n，蕴含知识与SSE没有太大的区别
  - RMSE（均方根误差）
    - 也称回归系统的拟合标准差，是MSE的平方根
  - SSR（回归平方和）
    - 是拟合数据与原始数据均值之差的平方和
  - SSE(和方差)
    - 拟合数据和原始数据对应点的误差的平方和。SSE越接近于0，说明模型选择和拟合更好，数据预测也越成功
  - SST（总平方和）
    - 是原始数据和均值之差的平方和，即SST=SSE+SSR
  - R^2（确定系数）
    - 是通过数据的变化来表征一个拟合的好坏。由公式可以看出，R^2(确定系数)的正常取值范围为[0,1]，越接近1，表明方程变量的解释能力越强，模型对数据的拟合程度也越好。

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

了解模型评估方法

```
# 模型评估
  mse = np.average((predict - np.array(y_data)) ** 2)  #均方误差
  rmse = np.sqrt(mse)  #均方根误差
  r2 = model.score(x_data, y_data)  #确定系数
  print('MSE = ', mse, end=' ')
  print('RMSE = ', rmse)
  print('R2 = ',r2)
```

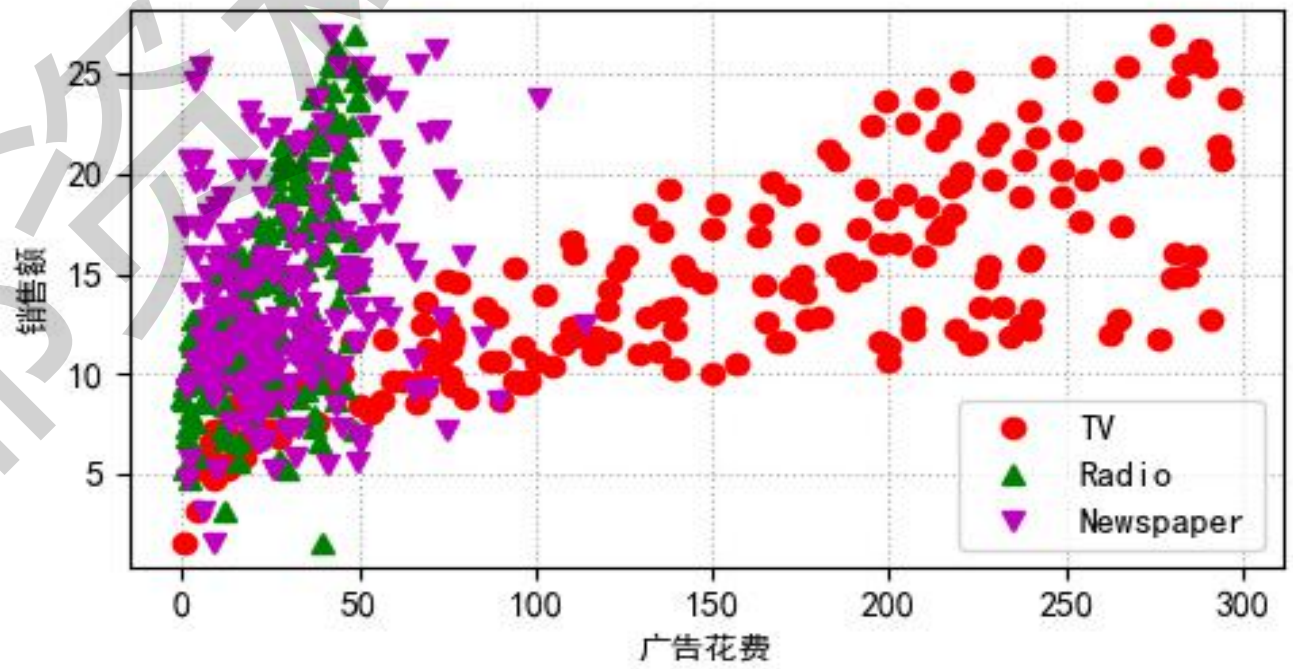
```
MSE =  2.8482758620689643
RMSE =  1.6876835787756437
R2 =  0.8832673827020916
```

评估结果表明原始值与预估值的值误差并不大，确定系统达到大于88.3%的准确率，能确定拟合的回归方程基本描述了原始数据蕴含的信息。

使用线性模型完成广告分析

使用线性模型完成广告分析

- 实现电视（TV）、收音机（Radio）和报纸（Newspaper）广告投入与销售额预测回归分析



查看代码：使用线性模型完成广告分析.ipynb

多重共线问题

多重共线性的问题

- 多重共线性（Multicollinearity）是指线性回归模型中的自变量之间由于存在高度相关关系而使模型的权重参数估计失真或难以估计准确的一种情形
  - 多重指一个自变量与多个其他自变量之间存在相关关系。
- 回顾广告实验
  - 在只有收音机广告投入的数据时
    - 线性回归的确定系统：35.714%
    - 均方根误差：3.45
  - 加入电视广告投入时
    - 线性回归的确定系统达到：89.473%
    - 均方根误差： 1.398

但当加入新闻报纸的特征因素时

- 确定系统：89.276%，有小幅度下降
- 均方根误差： 1.411 ，有小幅度上升

使拟合结果呈不好的趋势。

这是由于新加入的特征因素与之前因素之间存在多重共线性关系的结果

多重共线性的问题

百度智能云

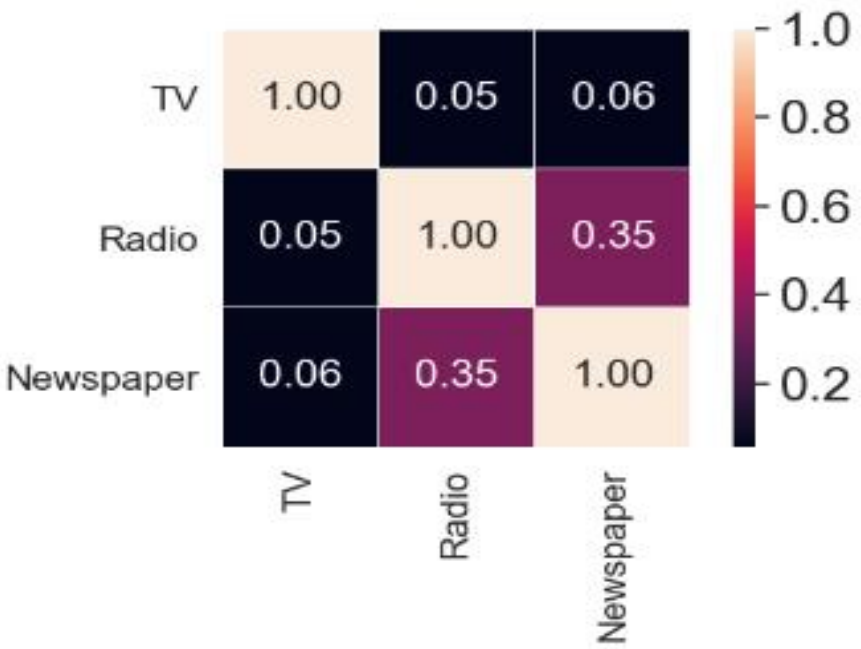
- 求解电视（TV）、收音机（Radio）和报纸（Newspaper）广告投入上述三列特征的相关系数。
  - 进行三个特征因素之间的相关关系的计算
  - 其中相关系数，是指对于一般的矩阵 X，执行 A = corrcoef(X) 后，A 中每个值的所在行和列，反应的是原矩阵 X 中相应的列向量间的相似程度。

查看代码：多重共线性的问题.ipynb

多重共线性的问题

- 存在多重共线性的问题
  - 从图中可以看到，Radio 与 Newspaper 具有 35% 的相关系数。这个模式表现还不够明显，在模型引入 Newspaper 特征因素后，虽然影响了模型质量，但是影响并不大。
  - 在有些业务中常常会存在更多的因素，因素间的相关系数有的达到70%以上，此种情况，是需要对这些特征因素进行处理后再加入模型的。

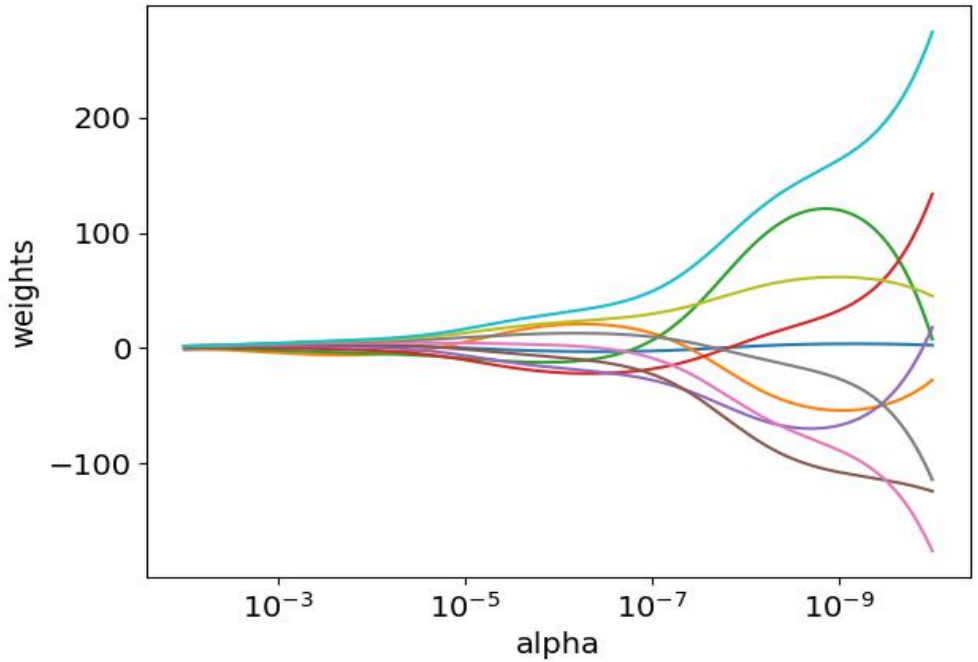
- 解决共线性的方法有很多
  - 排除引起共线性的变量
  - 将原模型变换为差分模型
  - 使用主成分分析法进行降维
  - 借用算法模型
    - 可减小参数估计量方差的**岭回归法**



岭回归

- 岭回归是一种专用于共线性数据分析的有偏估计回归方法
- 岭回归实质上是一种改良的线性回归法
  - 放弃了的无偏性
    - 无偏：在反复抽样的情况下，样本均值的集合的期望等于总体均值
  - 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
  - 对共线性问题和病态数据的拟合要强于之前的方法
  - 常用于多维问题与不适定问题

- 如图展示了岭回归模型的10个分量随正则化参数 alpha 变化而变化的趋势
  - 图中每一种颜色代表了不同的相关系数向量特征
  - 随着传入的正则化参数 Alpha 的变化而变化
  - 由于变化曲线呈现“脊”的形状，岭回归又称为脊回归



- 这个例子展示了岭回归处理病态矩阵（ ill-conditioned matrices）的优势。在病态矩阵里每一个目标变量微小的变动都会产生巨大的方差，对于这种情况就需要设置一个比较合适的alpha值来减少离差（噪声）。

- 当Alpha非常大的时候，正则化的影响支配了二乘法函数，相关系数趋近于0。在路径的结尾，当正则参数alpha趋近于0的时候，结果解趋近于了普通最小二乘法，系数表现出了很大的震荡。在实践中要不断的调节正则参数Alpha，以求在应用过程中寻求一种平衡。
- 岭回归方法的计算的复杂度与普通最小二乘相同。

百度智能云

THANK YOU

CLOUD.BAIDU.COM ABCXUEYUAN.BAIDU.COM