线性回归

欧阳若飞

数据科学家

主要内容

- 一元线性回归
  - $y = wx + b$
- 多元线性回归
  - $y = w_0 + w_1 x_1 + w_2 x_2$
- 带正则项的线性回归
  - $y = w_0 + w_1 x_1 + w_2 x_2 + \lambda(w_1^2 + w_2^2)$
- 带核函数的线性回归
  - $y = w_0 + w_1 \phi(x_1) + w_2 \phi(x_2)$

课件代码

代码放在：
https://github.com/rfouyang/machine-learning

课件在课后助教会提供
我也会把最新版本和代码放在一起

机器学习

$$x \longrightarrow f(x) \longrightarrow y$$

以数学的名义，从数据中来，到数据中去

学习模型(训练)：已知数据(x, y)学习模型f

使用模型(预测)：已知数据的输入x,带入f 求得y

回归模型

$$x \longrightarrow f(x) \longrightarrow y$$

输入：数据特征

连续型 年龄，收入

离散型 性别(不可比较)

学历(可比较)

输出：数据标签

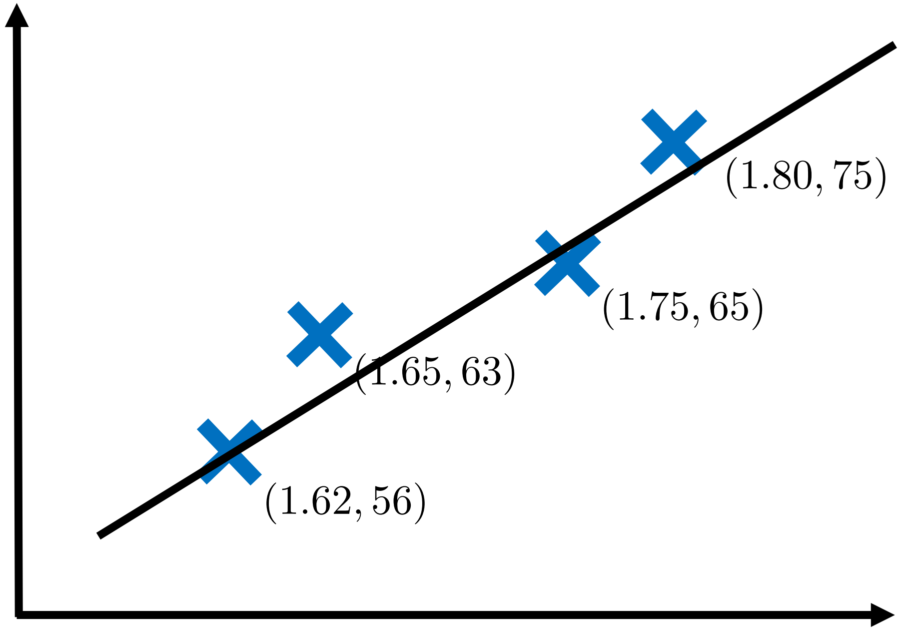
连续型 信用分数

输出为连续型变量就是回归

回归模型

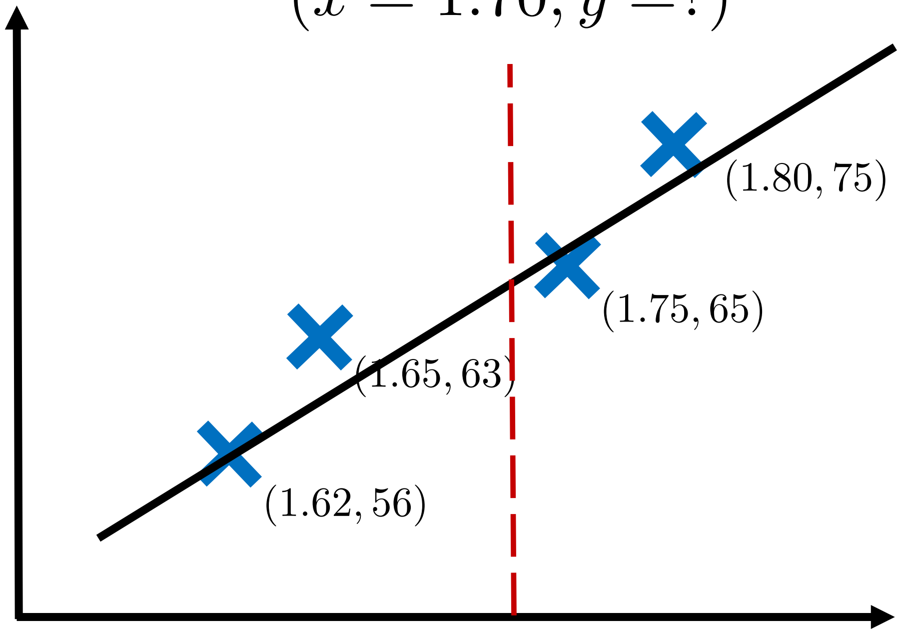
实例：根据身高预测体重

训练



预测

$$(x = 1.70, y = ?)$$



回归模型

问题 以下哪些是回归模型可以解决的问题？

- 根据食物的照片估算食物的热量
- 根据王者荣耀的排名预测玩家年龄
- 根据朋友圈内容判断妹子是否单身
- 根据历史歌单估计用户的星座

不管模型看似多复杂，输入是连续还是离散变量
只要输出为连续型变量就是回归

线性回归

线性回归是回归模型的一种，它有如下形式：

$$y = f(x) = w^\top x$$

$$y = wx + b = [x, 1]\begin{bmatrix} w \\ b \end{bmatrix} \checkmark$$

$$y = w_0 + w_1x_1 + w_2x_2 = [1, x_1, x_2]\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \checkmark$$

$$y = ax^2 + bx + c = [x^2, x, 1]\begin{bmatrix} a \\ b \\ c \end{bmatrix} \checkmark$$

$$y = a^2 + 2abx_1 + b^2x_2 = [1, x_1, x_2]\begin{bmatrix} a^2 \\ 2ab \\ b^2 \end{bmatrix} \times$$

模型参数需要为一次形式(线性)，输入可以是高次

一元线性回归

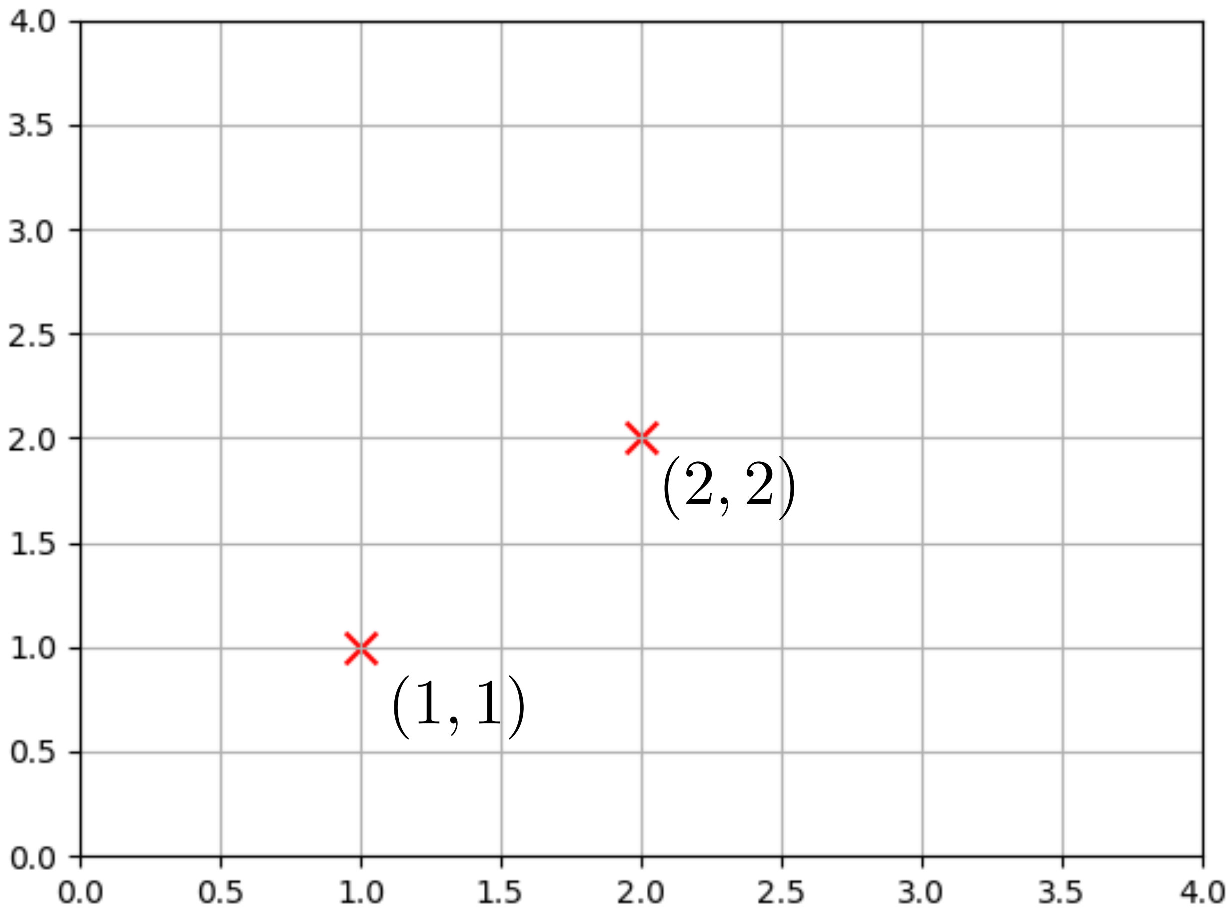
如果数据只有两个点

$$(1,1) \quad (2,2)$$

线性回归模型为：

$$y = wx + b$$

$$\begin{aligned} 1 &= w + b \\ 2 &= 2w + b \end{aligned} \quad \text{解得：} \quad \begin{aligned} w &= 1 \\ b &= 0 \end{aligned}$$



一元线性回归

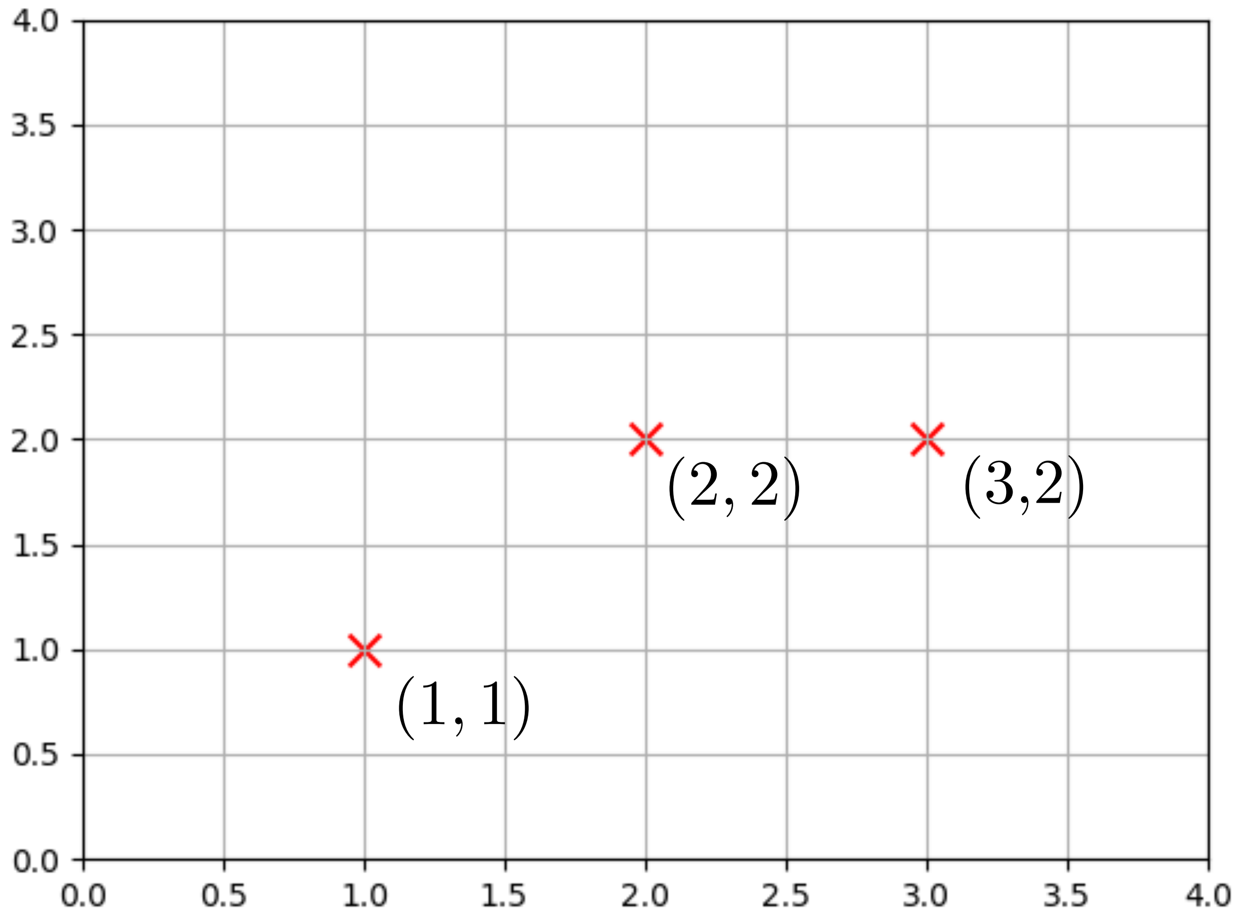
那数据量多于两个呢?

$$(1,1) \quad (2,2) \quad (3,2)$$

$$1 = w + b$$
$$2 = 2w + b$$
$$2 = 3w + b$$

三个方程两个未知数 不一定有解

我们实际的数据可能有一万个。。。



最小二乘法(Least Square)

既然无解，我们妥协一下：

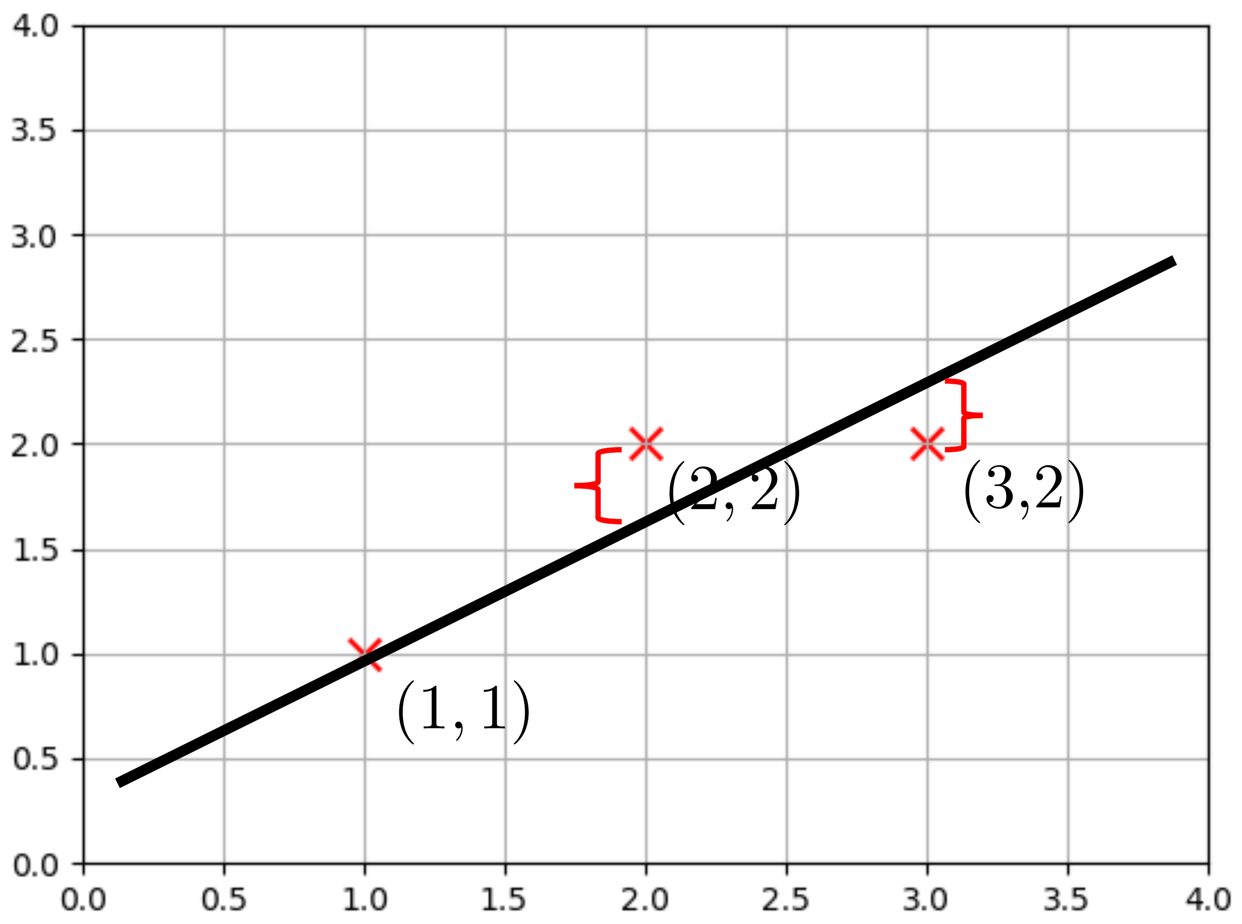
拟合的直线离所有点误差最小

$$y = f(x) + \underbrace{\varepsilon}_{\text{噪音}}$$

$$\min_{w,b} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Least

Square



最小二乘法(Least Square)

把原来的解方程组的问题转化为一个目标函数优化问题

目标函数: $L(w,b)=\min_{w,b}\sum_{i=1}^{n}(y_i-\widehat{y_i})^2$

如何求一个函数的最小值呢?

如果函数为凸函数，那函数的极小值就是唯一最小值

对函数求导取零(解析解法)

对函数求导，根据梯度下降方法迭代，直到导数为零(数值解法)

注意:

目标函数的输入是参数，输出是误差值，此时数据x,y是已知的常数

模型函数的输入是数据特征，输出是数据标签

最小二乘法(Least Square)

$$\min \sum_{i=1}^{n}(y_i - \widehat{w}^\top x_i - \widehat{b})^2 \qquad \widehat{y}_i = \widehat{w}^\top x_i + \widehat{b}$$

求导:

$$2\sum_{i=1}^{n}(y_i - \widehat{w}^\top x_i - \widehat{b})(-x_i) = 0$$

$$2\sum_{i=1}^{n}(y_i - \widehat{w}^\top x_i - \widehat{b})(-1) = 0$$

解得:

$$\widehat{w} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\widehat{b} = \bar{y} - \widehat{w}^\top \bar{x}$$

数据集的中点一定在直线上

一元线性回归

整体流程：

给定一个数据集 $\{(x_i, y_i)\}_{i=1}^{n}$

- 假设：

经过分析判断可以使用一元线性回归模型 $y = w^\top x + b + \varepsilon$ 求解

- 训练：

使用刚才所学的最小二乘法求得模型参数 $\widehat{w}, \widehat{b}$

- 预测：

对新的数据输入 x'_j 预测数据输出 $\widehat{y'}_j = \widehat{w} x'_j + \widehat{b}$

一元线性回归

回归模型评价标准

mean square error: $\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2$

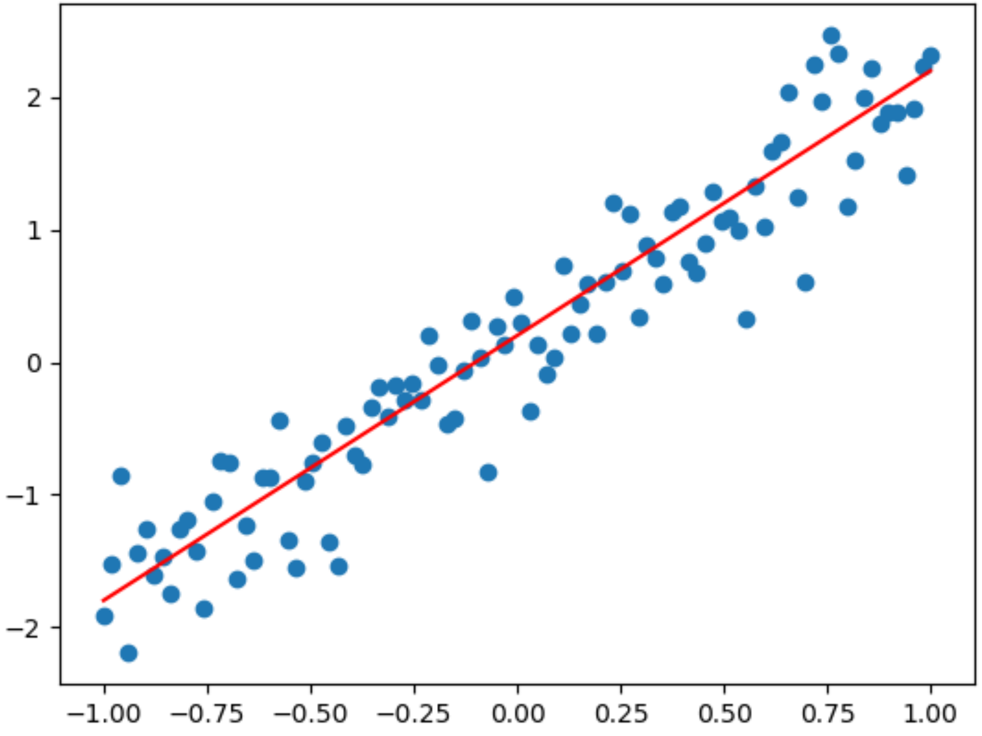
R^2 score: $1-\frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

模型怎么都要比用平均值猜来的好，$R^2<0$说明线性模型无效或错误

最差得和用平均值猜一样， $R^2=0$

最好没有误差， $R^2=1$

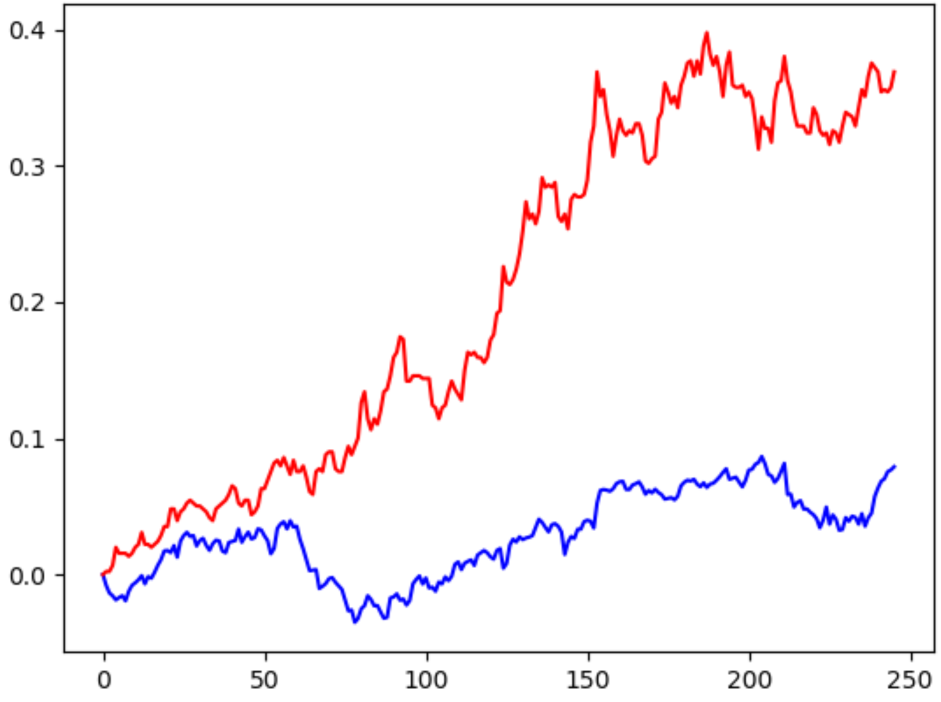
一元线性回归实例

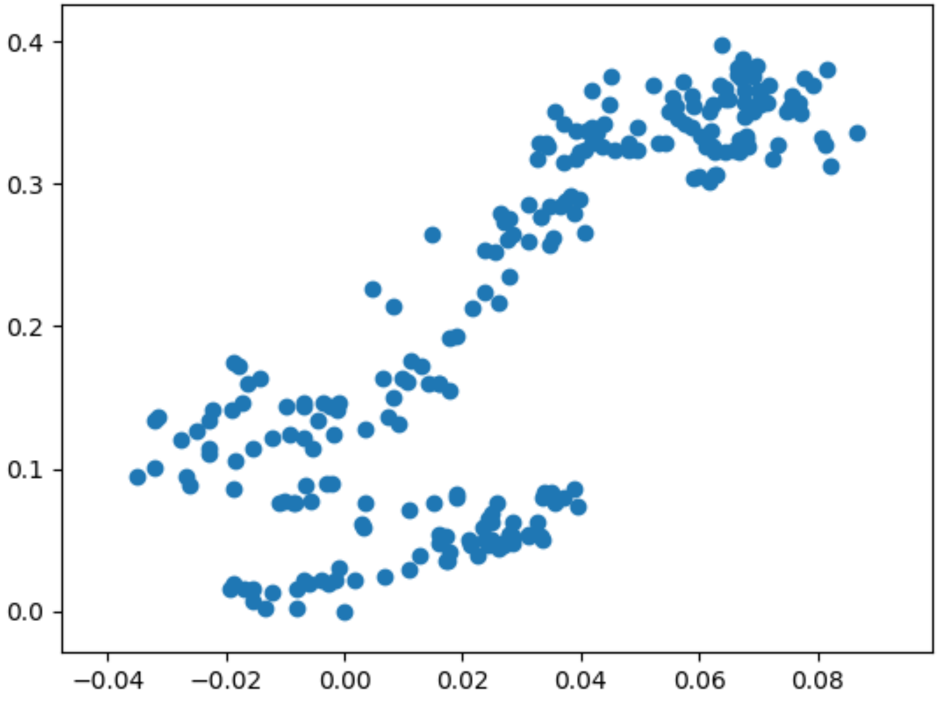


我们使用两个python库实现
sklearn
tensorflow

一元线性回归实例

股票收益 = 市场收益 + 超额收益　$r_s = \beta r_m + \alpha$





多元线性回归

刚才我们看到的问题，输入只是单一的特征

- 身高预测体重
- 大盘指数预测个股价格

但这种简单的特征无法反应数据的分布，我们需要更加丰富的特征！

- (身高，性别)预测体重
- (大盘指数，版块强弱，概念热度)预测个股价格

当数据输入为多个特征时，此时为多元线性回归

多元线性回归

多元线性回归：

$$y = f(x) + \varepsilon = w^\top x + \varepsilon$$

$$\begin{aligned} y &= f(x) + \varepsilon \\ &= b + w_1 x_1 + \ldots + w_d x_d + \varepsilon \\ &= w_0 x_0 + w_1 x_1 + \ldots + w_d x_d + \varepsilon \\ &= w^\top x + \varepsilon \end{aligned}$$

为了推导方便，把 b 改写成 w_0

$x_0 = 1$

$Y = Xw + \varepsilon$ $\quad X_{n\times(d+1)}$ Design Matrix: n行d+1列的数据输入矩阵

不同书的记法不一样,区别是带不带转置
我自己按深度学习的习惯来记

多元线性回归

最小二乘法：

$$\begin{aligned}\| e \|^2 &= e^\top e \\ &= (Y-\widehat{Y})^\top (Y-\widehat{Y}) \\ &= (Y - X\widehat{w})^\top (Y - X\widehat{w}) \\ &= (Y^\top - \widehat{w}^\top X^\top)(Y - X\widehat{w}) \\ &= Y^\top Y - Y^\top X\widehat{w} - \widehat{w}^\top X^\top Y + \widehat{w}^\top X^\top X\widehat{w}\end{aligned}$$

$$(A+B)^\top = A^\top + B^\top \quad (AB)^\top = B^\top A^\top$$

求导：

$$0 - Y^\top X - (X^\top Y)^\top + 2\widehat{w}^\top X^\top X = 0$$

$$(Ax)' = A \quad (xA)' = A^\top \quad (x^\top Ax)' = 2x^\top A$$

$$\widehat{w}^\top X^\top X = Y^\top X \longrightarrow \widehat{w} = (X^\top X)^{-1} X^\top Y \qquad \mathcal{O}(d^2 n)$$

多元线性回归

参数解析解： $\widehat{w} = (X^\top X)^{-1} X^\top Y \qquad \mathcal{O}(d^2 n)$

低维度求解没问题

高维度求解过慢　　例如：文本分类 特征向量维度为词典大小

推荐系统 用户听过的音乐轻松过万

所以一般我们使用梯度下降方法求解参数

多元线性回归实例

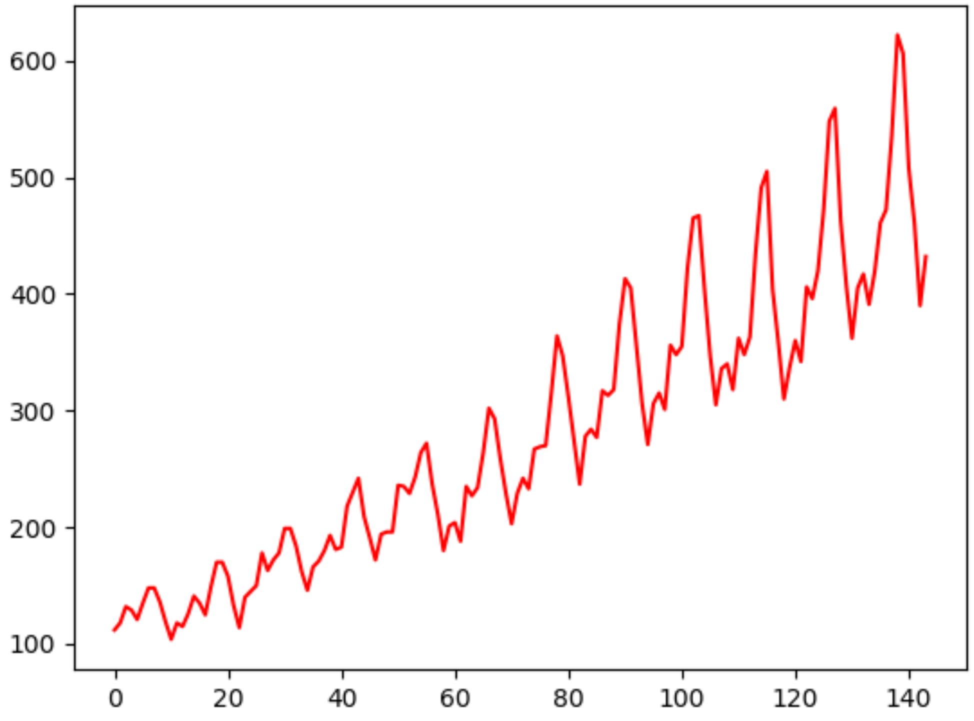
波士顿房价预测

输入: CRIM, ZN, INDUS, CHAS, …

输出: 房价

多元线性回归实例

飞机乘客流量预测



这怎么是线性回归?
多元怎么体现?

时间序列

自回归模型： $$x_{t+1} = w_1 x_t + w_2 x_{t-1} + w_3 x_{t-2} + w_4$$

自己和自己回归，输出和输入是一个意义下的数据

拓展模型：
- 马尔科夫过程 隐马尔科夫过程
- 卡尔曼滤波器
- 粒子滤波器
- ARMA-GARCH

过度拟合(Overfitting)

训练时误差越小越好吗?

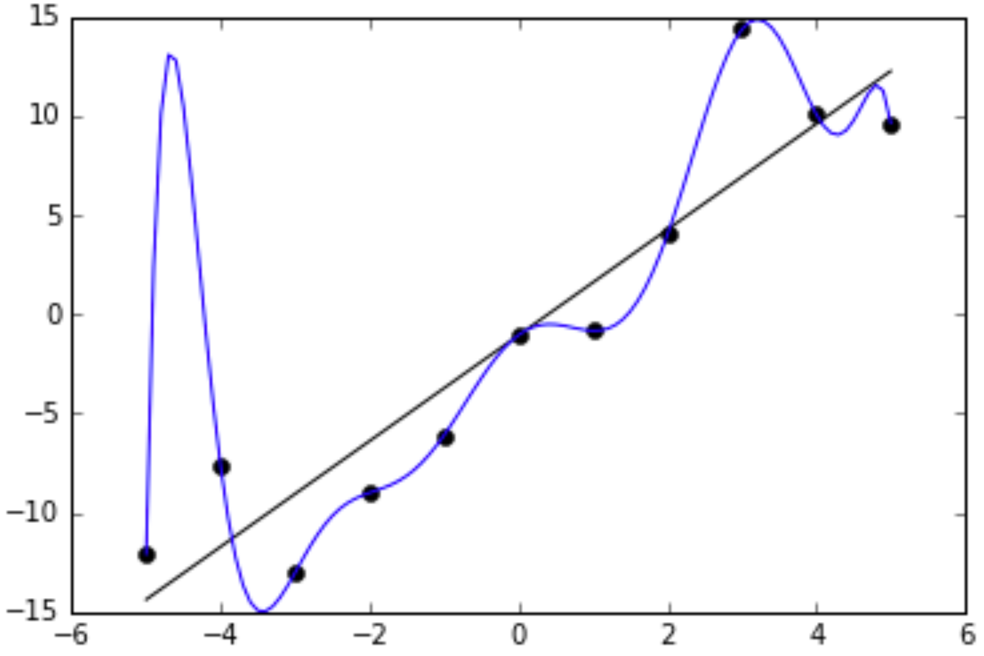
训练的目的是为了精确的预测

原有数据的信息需要泛化到新数据上

解决方法：加入正则项

$$L(w) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \lambda \parallel w \parallel_2^2$$

$$L(w) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \lambda \parallel w \parallel_1$$



L2正则项

$$
\begin{aligned}
& e^\top e+\lambda \widehat{w}^\top \widehat{w} \\
= & (Y-X\widehat{w})^\top (Y-X\widehat{w})+\lambda \widehat{w}^\top \widehat{w} \\
= & Y^\top Y-Y^\top X\widehat{w}-\widehat{w}^\top X^\top Y+\widehat{w}^\top X^\top X\widehat{w}+\lambda \widehat{w}^\top \widehat{w}
\end{aligned}
$$

$$0-Y^\top X-(X^\top Y)^\top+2\widehat{w}^\top X^\top X+2\widehat{w}^\top \lambda I=0$$

$$0-Y^\top X-(X^\top Y)^\top+2\widehat{w}^\top \underline{(X^\top X+\lambda I)}=0$$

$$\widehat{w}=(X^\top X+\lambda I)^{-1}X^\top Y$$

L2正则项

原有参数解析解： $\widehat{w} = (X^\top X)^{-1} X^\top Y$

带L2正则的参数解析解： $\widehat{w} = (X^\top X + \lambda I)^{-1} X^\top Y$

线性回归中输入特征相关性过强的话，$X^\top X$ 可能不正定，不能求逆

$X^\top X + \lambda I$ 增加了对角线的数值，能保持矩阵正定

小tirck:

如果很懒，不想做去除输入特征的相关性，暴力加L2正则项吧

L1正则项

$$
\begin{aligned}
& e^{\top} e + \lambda \parallel \widehat{w} \parallel \\
&= (Y - X\widehat{w})^{\top}(Y - X\widehat{w}) + \lambda \parallel \widehat{w} \parallel \\
&= Y^{\top}Y - Y^{\top}X\widehat{w} - \widehat{w}^{\top}X^{\top}Y + \widehat{w}^{\top}X^{\top}X\widehat{w} + \lambda \parallel \widehat{w} \parallel
\end{aligned}
$$

$$
\begin{aligned}
& 0 - Y^{\top}X - (X^{\top}Y)^{\top} + 2\widehat{w}^{\top}X^{\top}X + \lambda = 0 \qquad \widehat{w} \geq 0 \\
& 0 - Y^{\top}X - (X^{\top}Y)^{\top} + 2\widehat{w}^{\top}X^{\top}X - \lambda = 0 \qquad \widehat{w} < 0
\end{aligned}
$$

$$
\begin{aligned}
& \widehat{w} = (X^{\top}X)^{-1}(X^{\top}Y - \frac{1}{2}\lambda) \qquad \widehat{w} \geq 0 \\
& \widehat{w} = (X^{\top}X)^{-1}(X^{\top}Y + \frac{1}{2}\lambda) \qquad \widehat{w} < 0
\end{aligned}
$$

L1正则项

原有参数解析解： $\widehat{w} = (X^\top X)^{-1} X^\top Y$

带L2正则的参数解析解： $\widehat{w} = (X^\top X + \lambda I)^{-1} X^\top Y$ 怎么折腾都不为零

带L1正则的参数解析解：

$$\widehat{w} = (X^\top X)^{-1} (X^\top Y - \frac{1}{2}\lambda) \quad \widehat{w} \geq 0$$

$$\widehat{w} = (X^\top X)^{-1} (X^\top Y + \frac{1}{2}\lambda) \quad \widehat{w} < 0$$

可以为零！

- 右倾保守的L2：不希望特征权重差别过大
- 左倾激进的L1：挑选出出类拔萃的特征

小trick: L1可作为特征选择依据

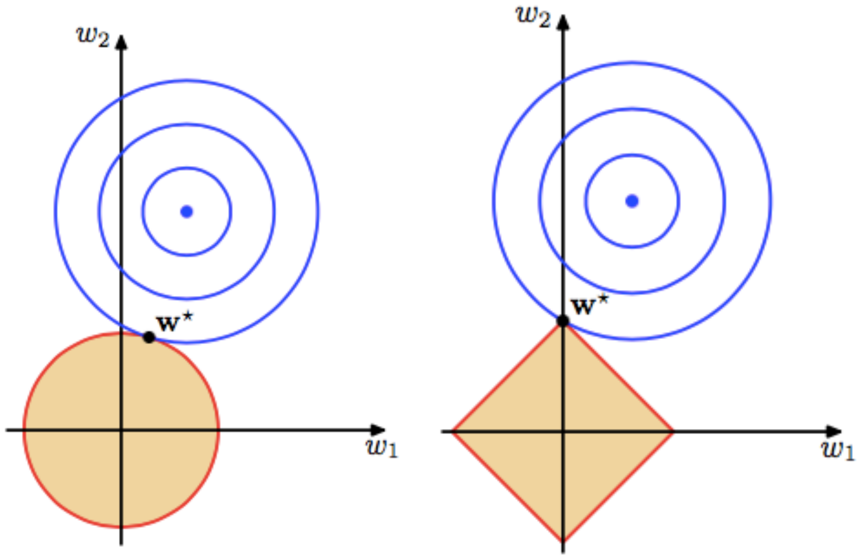
拉格朗日乘数法

现在来解释正则项前面的系数λ的由来

$$\begin{aligned} &\min f(x) \\ s.t.\quad &g(x) \le \eta \end{aligned} \qquad \min f(x) + \lambda g(x)$$

$$\begin{aligned} &\min e^\top e \\ s.t.\quad &\| w \|^2 \le \eta \end{aligned} \qquad \min e^\top e + \lambda \| w \|^2$$

$$\begin{aligned} &\min e^\top e \\ s.t.\quad &\| w \| \le \eta \end{aligned} \qquad \min e^\top e + \lambda \| w \|$$



正则项超参数调节

正则项的 λ 和 w 不同，它是超参数

参数V.S.超参数

- 参数是基于给定数据，由目标函数优化计算得到
- 超参数在模型优化前预先设定，对模型效果敏感性不如参数本身那么大

超参数一般通过在训练数据上的交叉验证进行调节

1	2	3	4	5

给定 $\lambda \in [0.05, 0.1, 1, 5]$

for $i \in [1, 2, 3, 4, 5]$ 训练 X_{-i} 预测 X_i

取综合预测效果最好的 λ

带核函数的线性回归

$$
\begin{aligned}
y &= ax^2 + bx + c \\
&= [c, b, a] \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \\
&= w^\top \phi(x)
\end{aligned}
$$

线性回归不仅可以画直线也可以画曲线?

$$\phi(x_d) = x^d$$

多项式回归是带核函数的线性回归的一个特例

$$y = w_0 + w_1 x + w_2 x^2 + \ldots + w_d x^d$$

带核函数的线性回归

$$y = f(x) + \varepsilon = w^\top \phi(x) + \varepsilon$$

回归的对象不是一组数据点，而是一组函数

核函数的选择:

- Radial basis function: 高斯过程，支持向量机
- Fourier basis function: 傅里叶变换
- Wavelet basis function: 小波变换
- Laplace basis function: 拉普拉斯变换

这些都是线性回归，简单来说就是找核函数的线性系数!

带核函数的线性回归

$$\widehat{w} = (X^\top X + \lambda I)^{-1} X^\top Y$$

把 x 装进 ϕ 里去

$$\widehat{w} = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y \;= \Phi^\top (\Phi\Phi^\top + \lambda I)^{-1} Y$$

$$(PQ + I)^{-1} P = P(QP + I)^{-1}$$

预测：

$$y = X_* \widehat{w} = \phi(x_*) \Phi^\top (\Phi\Phi^\top + \lambda I)^{-1} Y = k(K + \lambda I)^{-1} Y$$

核函数选择： $k(x, x') = \exp\{-\frac{1}{2}(x - x')^\top (x - x')\}$

核函数方法会在后续详细讲解，现在感受下就好

小结

- 一元线性回归
  - 回归模型训练预测流程
  - 最小二乘法
  - 模型评价标准
- 多元线性回归
  - 自回归和时间序列
- 带正则项的线性回归
  - L1L2正则项的作用
  - 拉格朗日乘数法
  - 十字交叉法超参数设定
- 带核函数的线性回归
  - 多项式回归

数学推导并不重要，理解原理就好

在不同标准数据集上比较效果

工作中，处理数据远比调试模型重要

谢谢大家

Q&A