

回归

Regression

2019 年 1 月

第3章 回归

- 回归概念

- 线性回归

- 线性回归的概率解释

- 线性回归的扩展

- 线性回归实践

回归的定义

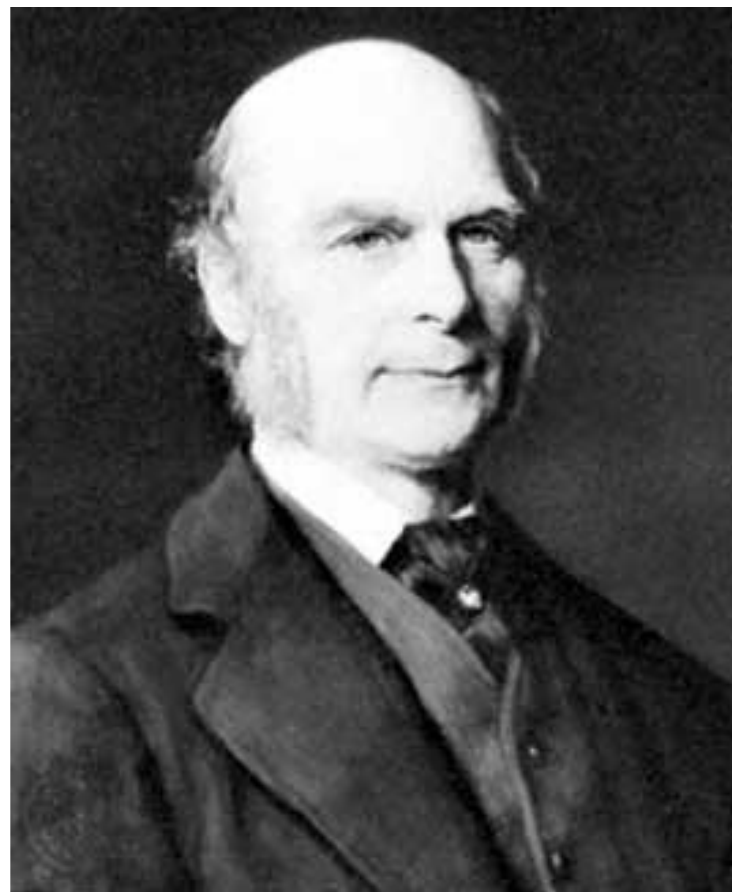
- 回归（Regression）分析可以对已知变量和预测变量之间的联系建模。
- 已知变量是描述样本的感兴趣的属性，一般已知变量的值是已知的，预测变量的值是我们要预测的。当已知变量和所有预测变量都是连续值时，回归分析是一个好的选择。
- 对具有相关关系(显著相关以上相关)的两个或两个以上的变量之间所具有的变化规律进行拟合，确立一个相应的数学表达式(经验公式)，通过一个或多个变量的变化去解释另一变量变化的方法，以便从定量的角度由已知量推测未知量，为估算预测或控制提供重要依据。
- 回归分析包括：线性回归、非线性回归以及逻辑回归等。

回归来源

弗兰西斯·高尔顿于1822年生于英格兰，与达尔文是表兄弟关系，他从小智力超常、聪颖过人，被誉为神童，是著名的优生学家、心理学家，差异心理学之父，也是心理测量学上生理计量法的创始人，享年89岁。

高尔顿一生在统计学方面贡献很多，首次引入了“Regression回归”一词，第一次使用了相关系数（correlation coefficient）的概念，并采用字母“ r ”来表示。

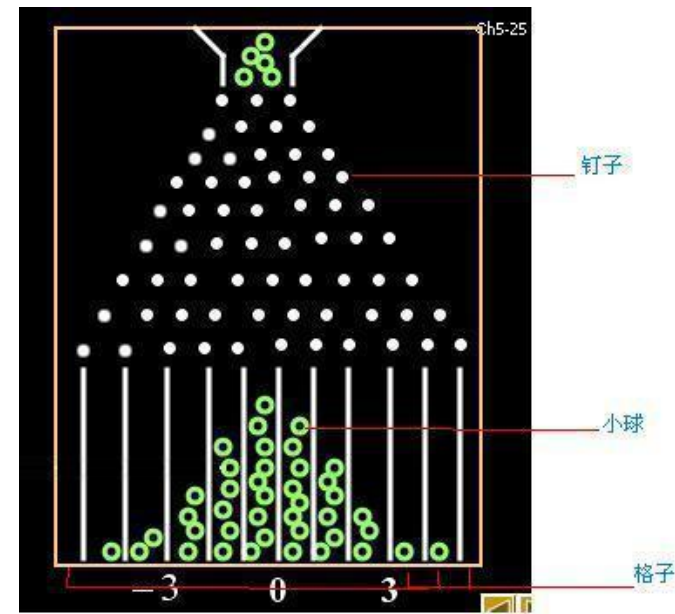
高尔顿设计的用来研究随机现象的高尔顿钉板模型，更是被广泛用来描述正态分布的经典例子。



高尔顿（Francis Galton）
1822—1911

钉板试验

自上端放入一小球，任其自由下落，在下落过程中当小球碰到钉子时，从左边落下与从右边落下的机会相等。碰到下一排钉子时又是如此，最后落入底板中的某一格子。因此，任意放入一球，则此球落入哪一个格子，预先难以确定。但是如果放入大量小球，则其最后所呈现的曲线，几乎总是一样的。



回归趋势

- “Regression(回归)”一词是由英国著名人类学家、气象学家和统计学家高尔顿于1885年在其《身高遗传中的平庸回归》一文中首次引入的,他在研究身高与遗传之间的联系时,观察了1078对夫妇的二人的平均身高 X 以及其一个成年后代的身高 Y ,从中发现在直角坐标系下,二者之间的关系近乎是一条直线,并且得到如下数学关系:

$$\hat{Y} = 33.73 + 0.516 X$$

结果解释

父辈平均身高每增加或减少一个单位,其成年后代的身高平均增加或减少0.516个单位。

从人类遗传上来说,父母个子高这一基因会遗传给他们的后代,导致产生高个子的下一代,但子代的身高并不会象其父辈,出现越来越高的现象,而是趋向于比他们父辈身高更加平均的水平。

高尔顿将人类这种遗传现象称为“回归”。人类也正是由于这种回归,才能生生不息的繁衍下去。

回归的数学描述

回归问题：根据给定的训练集，

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\}$$

其中

$$x_i \in X = R^n, y_i \in Y = R, i = 1, 2, \dots, l$$

要求寻找 X 上的决策函数

$$f(x): X \rightarrow Y$$

以便能用决策函数 $f(x)$ “较好地”推断任一模式 x 相对应的 y 值。

第3章 回归

- 回归概念

- 线性回归

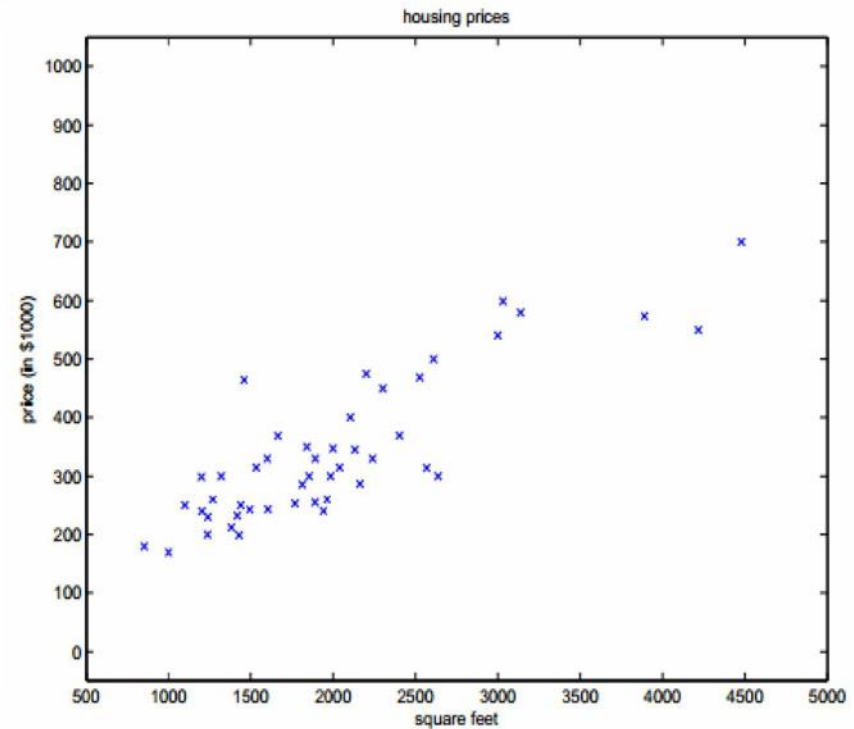
- 线性回归的概率解释

- 线性回归的扩展

- 线性回归实践

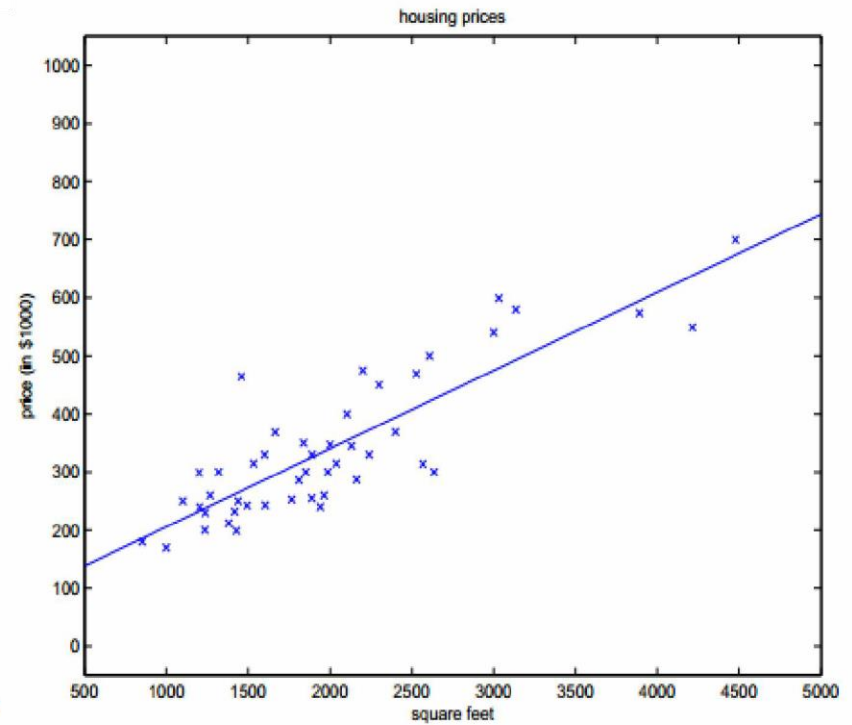
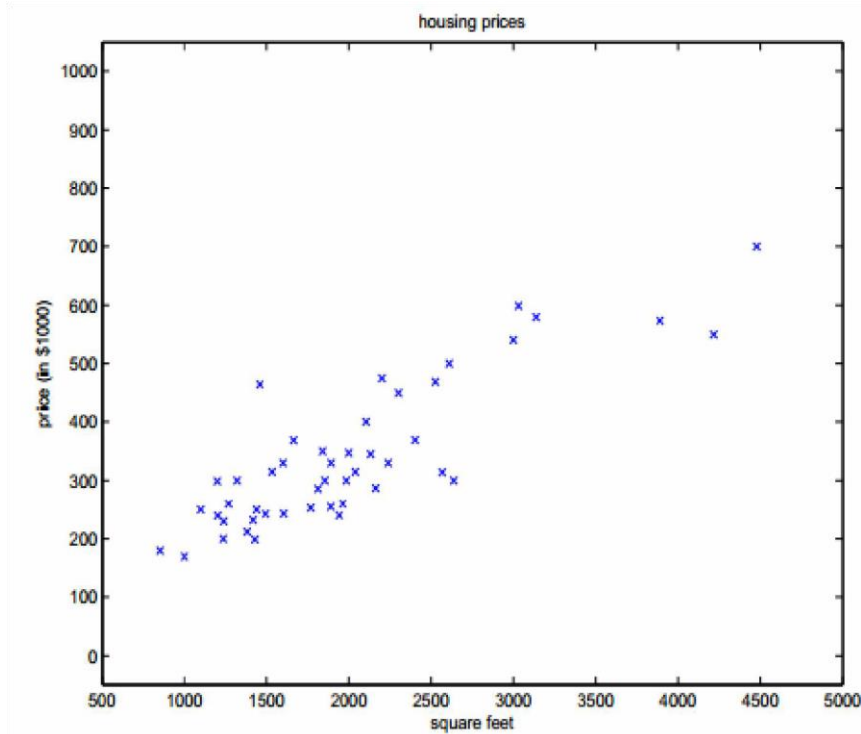
线性回归的例子

| Living area (feet ²) | Price (1000\$s) |
|----------------------------------|-----------------|
| 2104 | 400 |
| 1600 | 330 |
| 2400 | 369 |
| 1416 | 232 |
| 3000 | 540 |
| ⋮ | ⋮ |



线性回归的例子

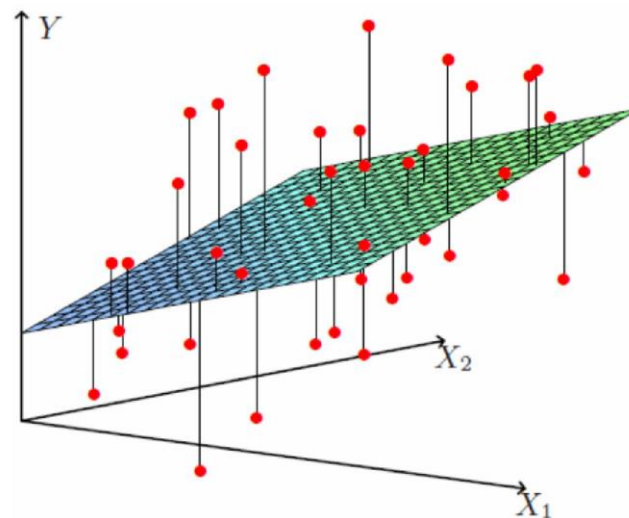
- $y=ax+b$



多个变量的情况

- 考虑两个变量

| Living area (feet ²) | #bedrooms | Price (1000\$) |
|----------------------------------|-----------|----------------|
| 2104 | 3 | 400 |
| 1600 | 3 | 330 |
| 2400 | 3 | 369 |
| 1416 | 2 | 232 |
| 3000 | 4 | 540 |
| ⋮ | ⋮ | ⋮ |



$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h(x) = \sum_{i=0}^n \theta_i x_i = \theta^T x$$

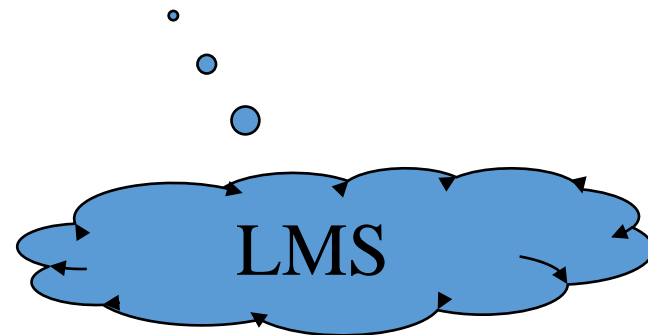
最小二乘方法（LMS）

- 如何确定模型 $h_{\theta}(x)$ 的参数 θ

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- 度量指标——代价函数(cost function)

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

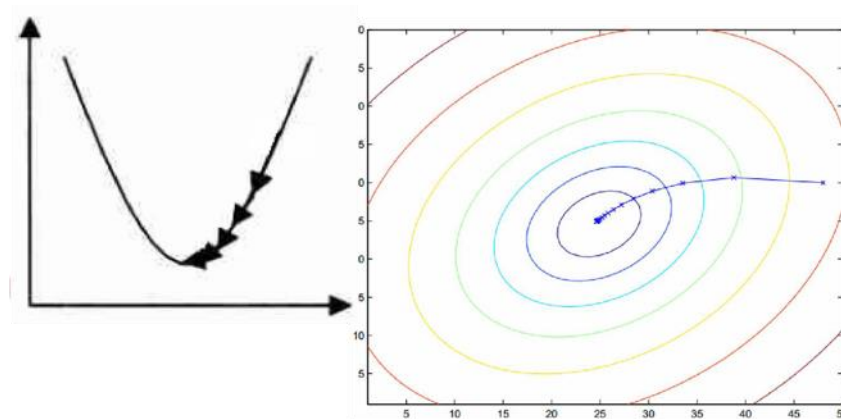


梯度下降法

- 初始化 θ (随机初始化)
- 迭代, 新的 θ 能够使得 $J(\theta)$ 更小
- 如果 $J(\theta)$ 能够继续减少, 返回(2)

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

- α 称为学习率/步长



梯度方向

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_{\theta}(x) - y)^2 \\&= 2 \cdot \frac{1}{2} (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_{\theta}(x) - y) \\&= (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^n \theta_i x_i - y \right) \\&= (h_{\theta}(x) - y) x_j\end{aligned}$$

批处理梯度下降法

Repeat until convergence {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^m (y^{(i)} - h_{\theta}(x^{(i)})) x_j^{(i)}$$

}

gradient descent. Note that, while gradient descent can be susceptible to local minima in general, the optimization problem we have posed here for linear regression has only one global, and no other local, optima; thus gradient descent always converges (assuming the learning rate α is not too large) to the global minimum. Indeed, J is a convex quadratic function.

随机梯度下降法

```
Loop {  
    for i=1 to m, {  
         $\theta_j := \theta_j + \alpha (y^{(i)} - h_{\theta}(x^{(i)})) x_j^{(i)}$   
    }  
}
```

This algorithm is called **stochastic gradient descent** (also **incremental gradient descent**). Whereas batch gradient descent has to scan through the entire training set before taking a single step—a costly operation if m is large—stochastic gradient descent can start making progress right away, and continues to make progress with each example it looks at. Often, stochastic gradient descent gets θ “close” to the minimum much faster than batch gradient descent. (Note however that it may never “converge” to the minimum, and the parameters θ will keep oscillating around the minimum of $J(\theta)$; but in practice most of the values near the minimum will be reasonably good approximations to the true minimum.²) For these reasons, particularly when the training set is large, stochastic gradient descent is often preferred over batch gradient descent.

mini-batch

- 如果不是每拿到一个样本即更改梯度，而是若干个样本的平均梯度作为更新方向，则是 **mini-batch** 梯度下降算法。

Repeat until convergence {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^m (y^{(i)} - h_{\theta}(x^{(i)})) x_j^{(i)}$$

}

Loop {

for i=1 to m, {

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_{\theta}(x^{(i)})) x_j^{(i)}$$

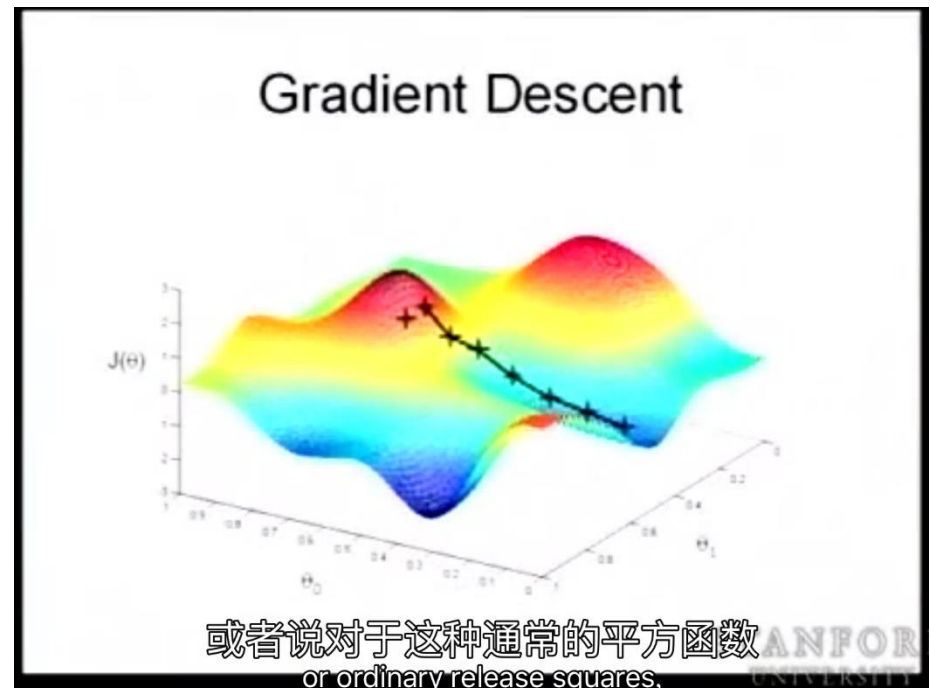
}

}

全局极小值 vs. 局部极小值

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

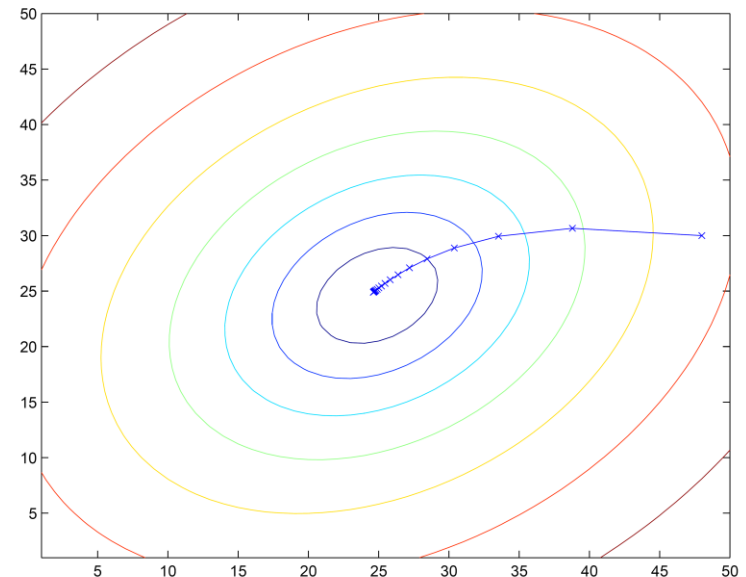
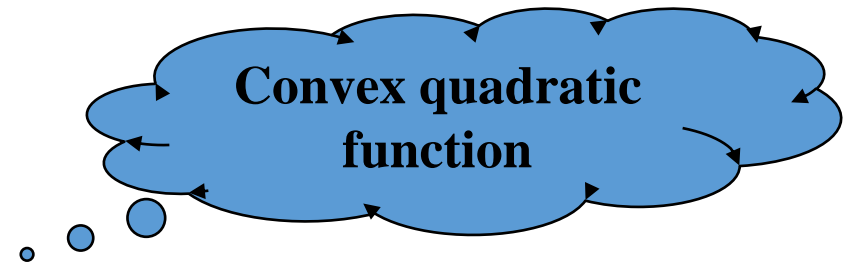
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$



线性回归的全局极小值保证

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$h(x) = \sum_{i=0}^n \theta_i x_i = \theta^T x$$



线性回归

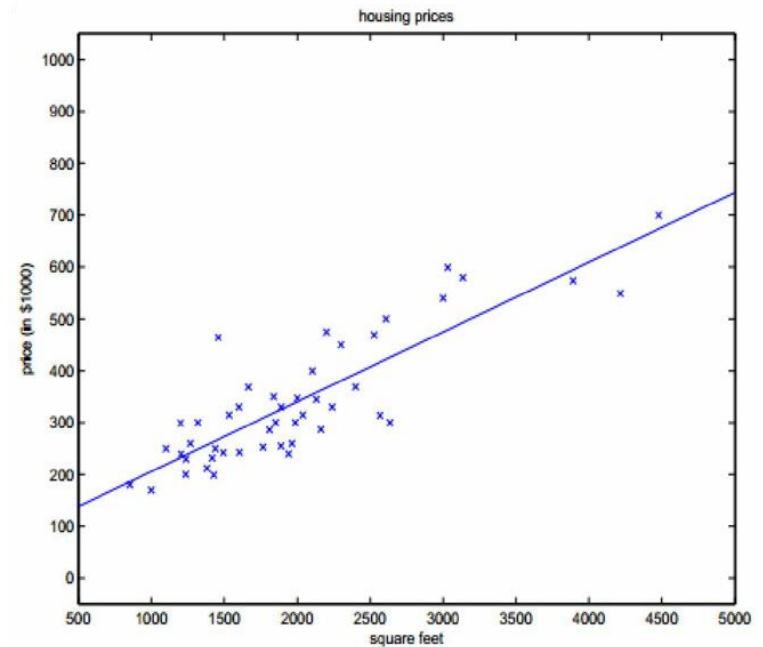
单变量线性回归实例

| Living area (feet ²) | Price (1000\$s) |
|----------------------------------|-----------------|
| 2104 | 400 |
| 1600 | 330 |
| 2400 | 369 |
| 1416 | 232 |
| 3000 | 540 |
| ⋮ | ⋮ |

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1$$

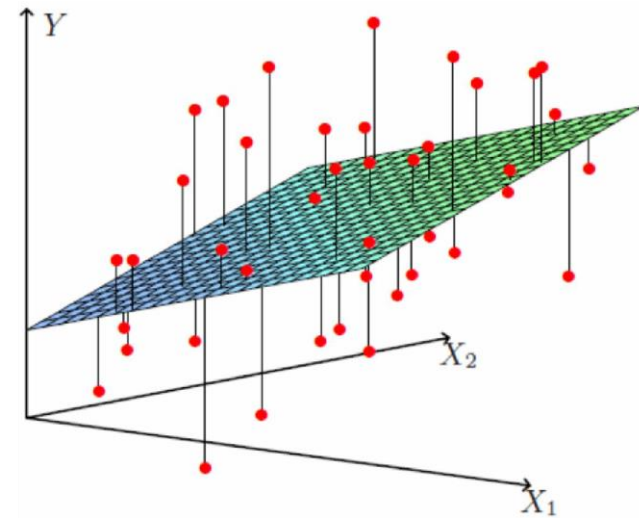
$$\theta_0 = 71.27,$$
$$\theta_1 = 0.1345.$$

By batch
gradient descent



多变量线性回归示例

| Living area (feet ²) | #bedrooms | Price (1000\$s) |
|----------------------------------|-----------|-----------------|
| 2104 | 3 | 400 |
| 1600 | 3 | 330 |
| 2400 | 3 | 369 |
| 1416 | 2 | 232 |
| 3000 | 4 | 540 |
| ⋮ | ⋮ | ⋮ |



$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

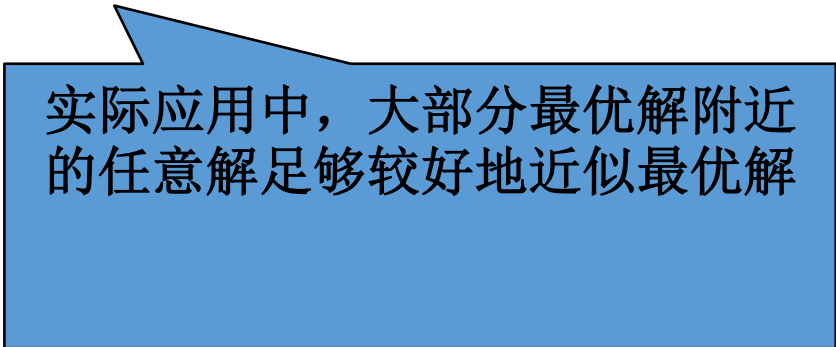
$$\theta_0 = 89.60, \theta_1 = 0.1392, \theta_2 = -8.738$$

不同梯度下降策略比较

- Batch VS. stochastic gradient descent (GD)
 - Batch GD 扫描整个训练集后再更新参数
 - Stochastic GD 遇到一个样本后立即更新参数
 - 对于大样本问题，BGD收敛较慢
 - 但SGD有可能发生震荡，而无法收敛到极小值



对于大样本问题，
推荐SGD



实际应用中，大部分最优解附近的
任意解足够较好地近似最优解

第3章 回归

- 回归概念
- 线性回归
- 线性回归的概率解释
- 线性回归的扩展
- 线性回归实践

线性回归的概率解释

最小二乘回归为何是非常自然的算法

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$

- 误差 $\varepsilon^{(i)}$ ($1 \leq i \leq m$) 是独立同分布的，服从均值为0，方差为某定值 σ^2 的**高斯分布**。
 - 原因：**中心极限定理**
 - 实际问题中，很多随机现象可以看做众多因素的独立影响的综合反应，往往近似服从正态分布。

线性回归的概率解释

误差的似然函数

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$p(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\begin{aligned} L(\theta) &= \prod_{i=1}^m p(y^{(i)} | x^{(i)}; \theta) \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

线性回归的概率解释

极大似然估计

- 似然函数(likelihood function):

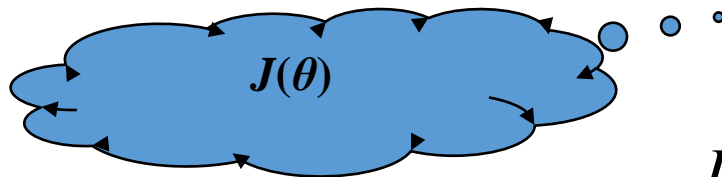
$$\begin{aligned} L(\theta) &= L(\theta; X, \vec{y}) = p(\vec{y}|X; \theta) \\ &= \prod_{i=1}^m p(y^{(i)} | x^{(i)}; \theta) \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

- 选择 θ 最大化似然函数 $L(\theta)$

线性回归的概率解释

- 最大化对数似然函数: $\ell(\theta)$

$$\begin{aligned}\ell(\theta) &= \log L(\theta) \\ &= \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right) \\ &= \sum_{i=1}^m \log \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right) \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \left[\frac{1}{2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 \right]\end{aligned}$$



$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

线性回归的概率解释

最大化对数似然函数 $\ell(\theta) \leftrightarrow$ 最小化 $J(\theta)$

$$l(\theta) = \log L(\theta) \leftrightarrow J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

最小二乘回归对应着关于参数 θ 的最大似然估计

Note however that the probabilistic assumptions are *by no means necessary* for least-squares to *be a perfectly good and rational procedure*, and there may—and indeed there are—other natural assumptions that can also be used to justify it.

线性回归的概率解释

正则项与防止过拟合

$$\begin{cases} \lambda > 0 \\ \rho \in [0,1] \end{cases}$$

- L2-norm: $J(\vec{\theta}) = \frac{1}{2} \sum_{i=1}^m (h_{\vec{\theta}}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2$

- L1-norm: $J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\vec{\theta}}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n |\theta_j|$

- Elastic Net:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\vec{\theta}}(x^{(i)}) - y^{(i)})^2 + \lambda \left(\rho \cdot \sum_{j=1}^n |\theta_j| + (1 - \rho) \cdot \sum_{j=1}^n \theta_j^2 \right)$$

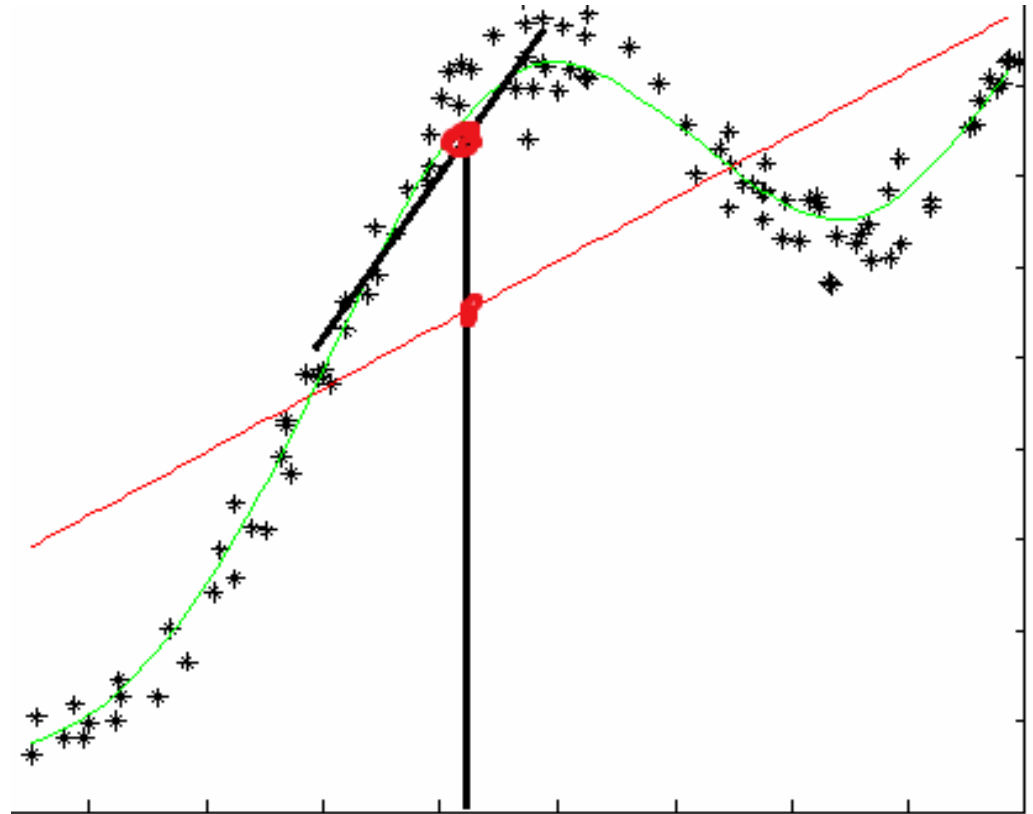
第3章 回归

- 回归概念
- 线性回归
- 线性回归的概率解释
- 线性回归的扩展
- 线性回归实践

线性回归的扩展

局部加权回归(LWR)

- 黑色是样本点
- 红色是线性回归曲线
- 绿色是局部加权回归曲线



局部加权回归(LWR)

- 线性回归方法的步骤:

1. Fit θ to minimize $\sum_i (y^{(i)} - \theta^T x^{(i)})^2$
2. Output $\theta^T x$.

- 局部加权线性回归方法的步骤:

1. Fit θ to minimize $\sum_i w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2$
2. Output $\theta^T x$.

- 权值的作用
 - 放大邻近点的贡献
 - 缩小甚至忽略远距离点的贡献

局部加权回归(LWR)

- ω 的一种可能的选择方式(高斯核函数):

$$w^{(i)} = \exp \left(-\frac{(x^{(i)} - x)^2}{2\tau^2} \right)$$

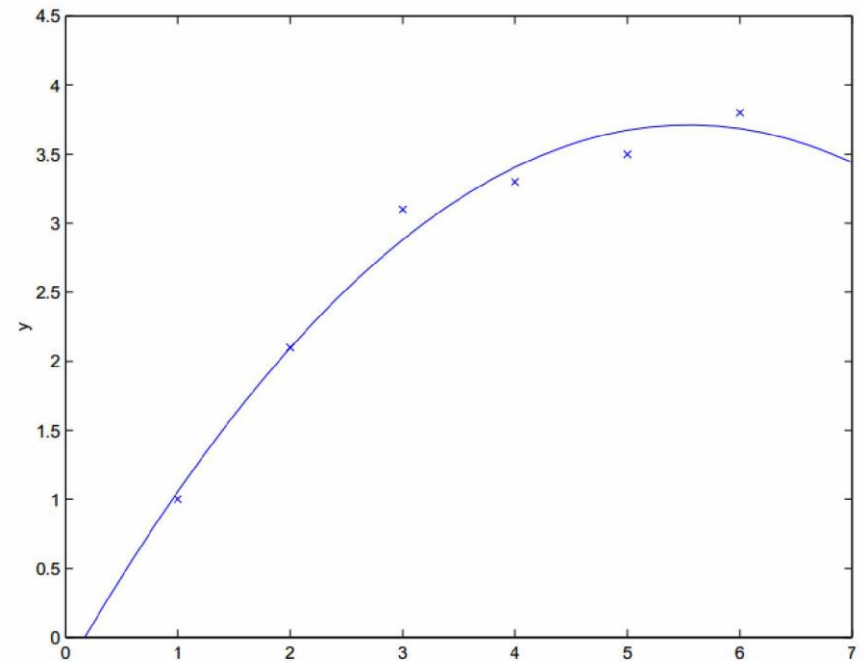
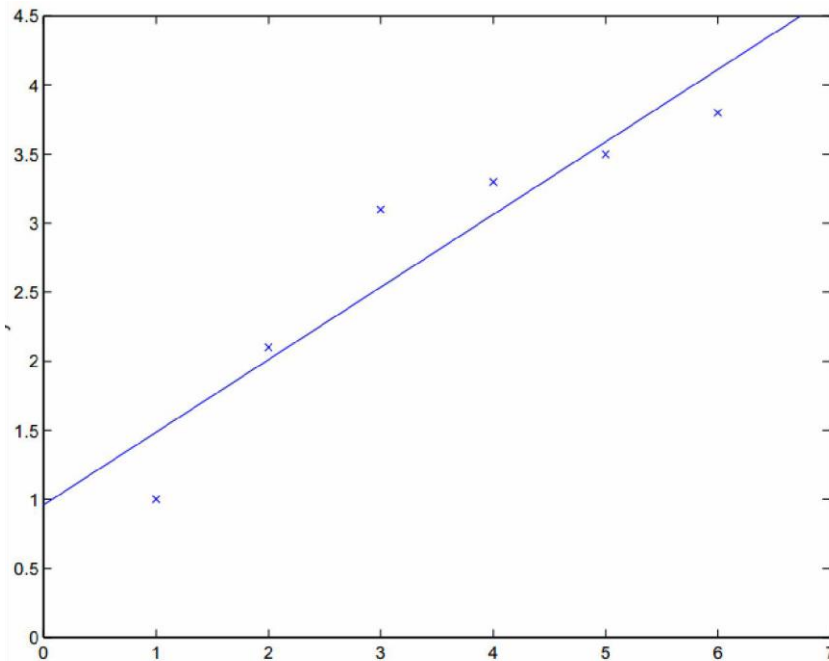
- τ 称为带宽，它控制着训练样本随着与 $\mathbf{x}^{(i)}$ 距离的衰减速率。
- 多项式核函数

$$\kappa(\mathbf{x}_1, \mathbf{x}_2) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + R)^d$$

线性回归的扩展

非线性模型的多元回归

- 可以对样本是非线性的，只要对参数 θ 线性



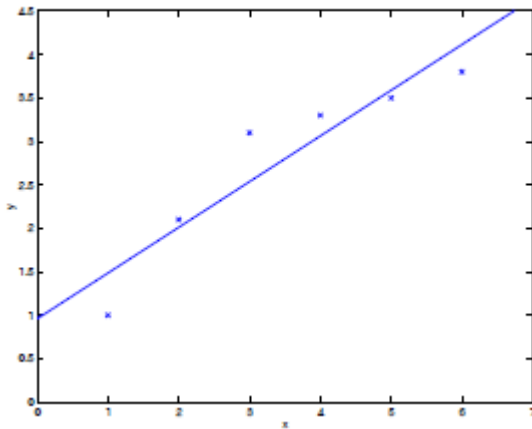
$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

线性回归的扩展

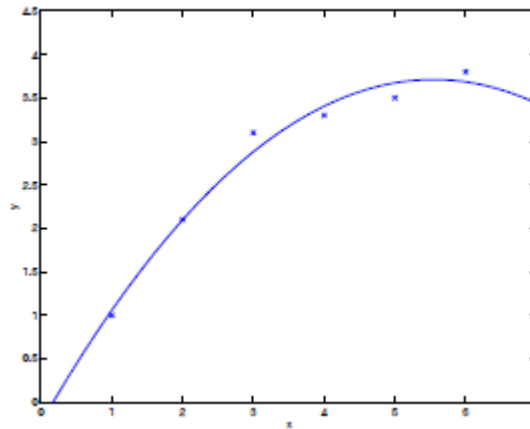
非线性模型的多元回归

- 实际数据可能并不适用线性预测模型

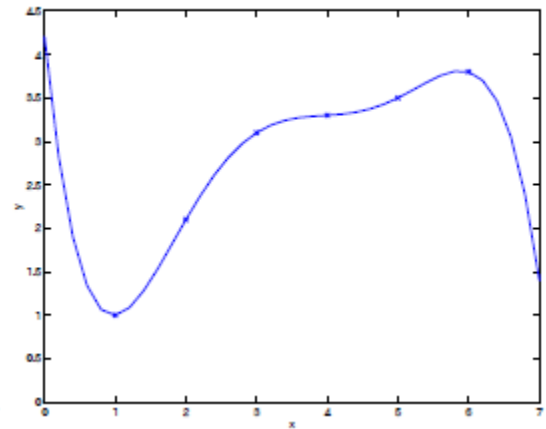
$$\theta_0 + \theta_1 x$$



$$\theta_0 + \theta_1 x + \theta_2 x^2$$



$$\sum_{j=0}^5 \theta_j x^j$$



- 换个角度考虑：引入新特征

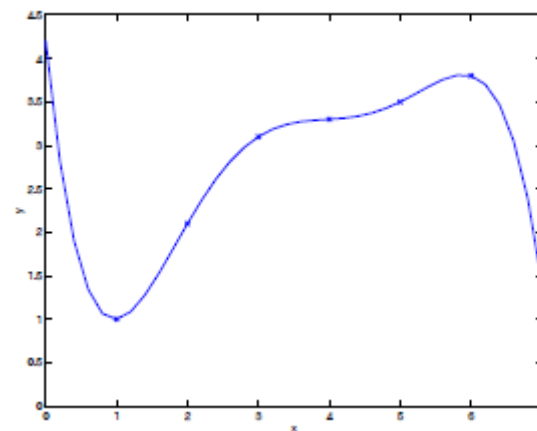
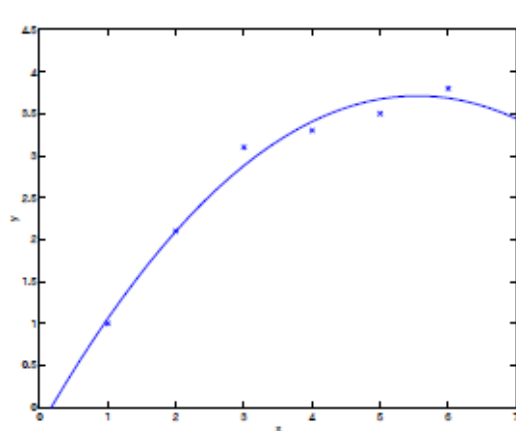
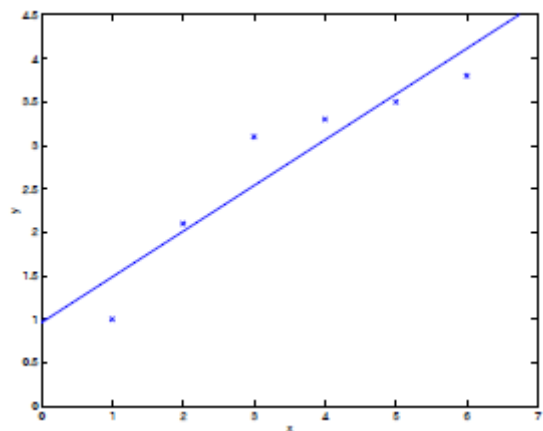
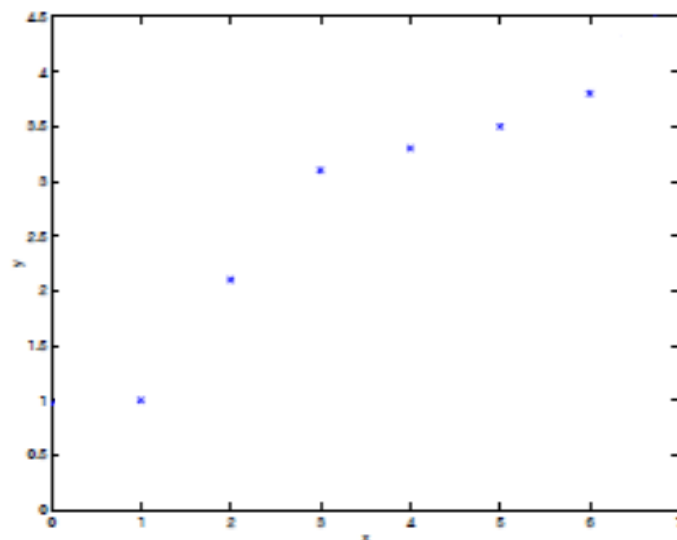
$$\theta_0 + \theta_1 x_1 + \theta_2 x_2$$

x^2

单一指标的
不同形式！

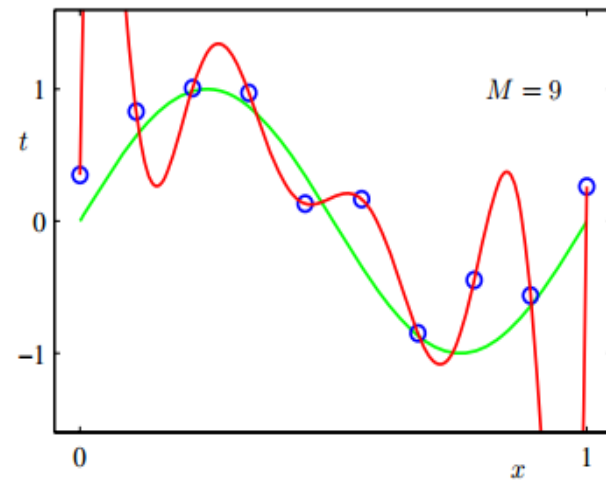
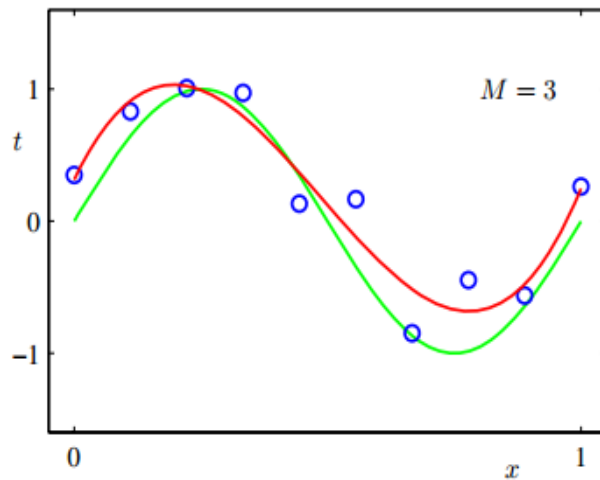
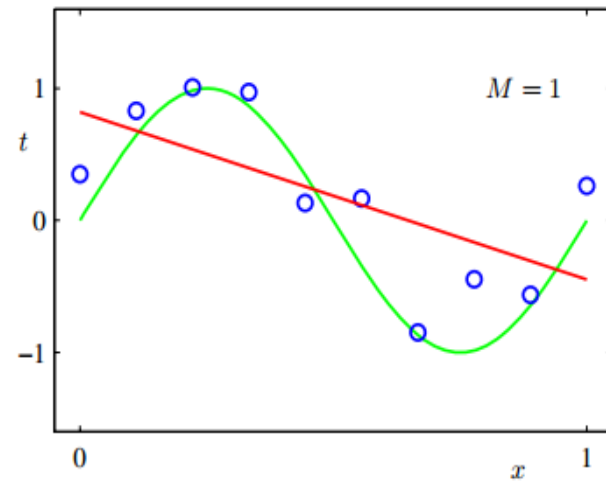
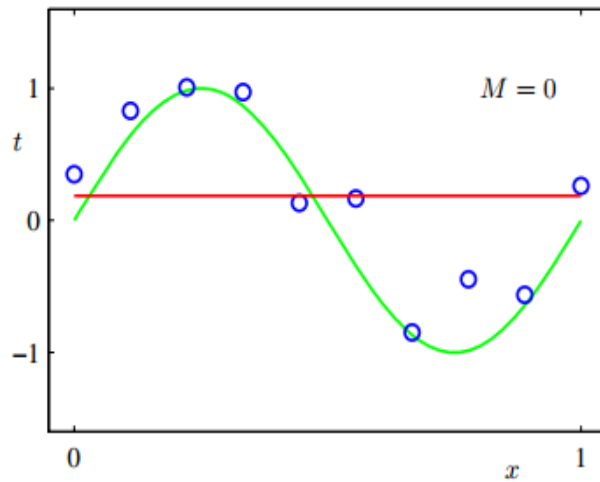
线性回归的扩展

欠拟合/过拟合



线性回归的扩展

欠拟合/过拟合



线性回归的扩展

模型训练和预测

训练数据 $\rightarrow \theta$

训练数据 $\rightarrow \theta$

测试数据

训练数据 $\rightarrow \theta$

验证数据 $\rightarrow \lambda$

测试数据

- 交叉验证
 - 如：十折交叉验证

结果评价

均方误差 (MSE)

$$\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

均方根误差 (RMSE)

$$\sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2}$$

R Squared

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \quad SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2, \quad SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

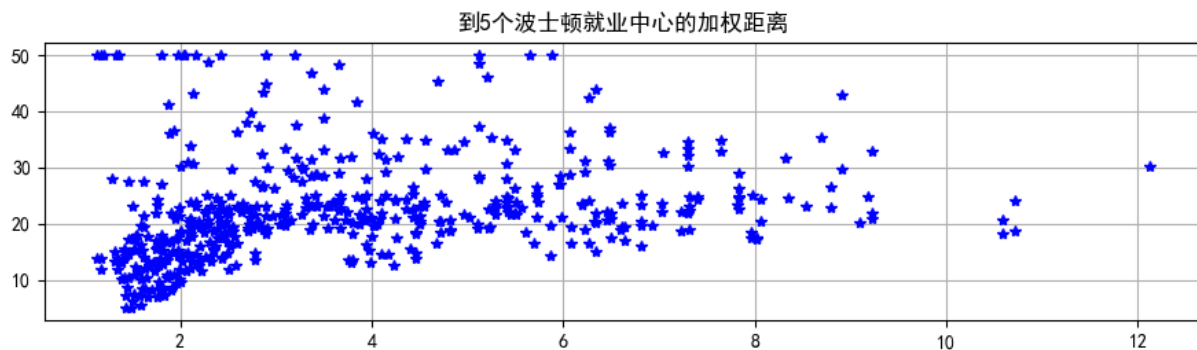
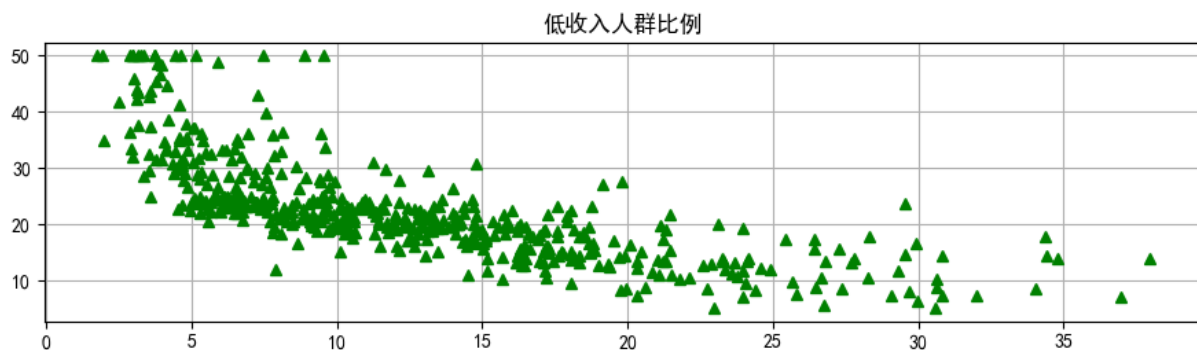
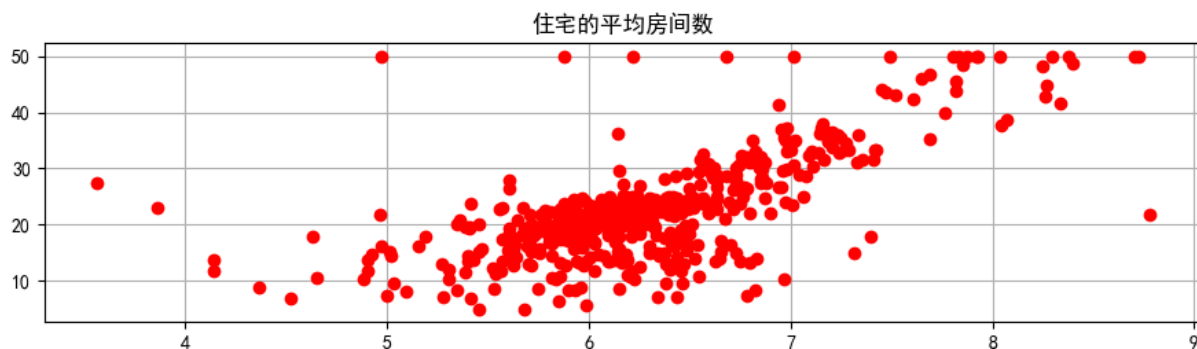
第3章 回归

- 回归概念
- 线性回归
- 线性回归的概率解释
- 线性回归的扩展
- 线性回归实践

波士顿住房价格数据

1. CRIM: 人均犯罪率
2. ZN: 25,000平方英尺以上民用土地的比例
3. INDUS: 城镇非零售业商用土地比例
4. CHAS: 是否邻近查尔斯河, 1是邻近, 0是不邻近
5. NOX: 一氧化氮浓度 (千万分之一)
6. RM: 住宅的平均房间数
7. AGE: 自住且建于1940年前的房屋比例
8. DIS: 到5个波士顿就业中心的加权距离
9. RAD: 到高速公路的便捷度指数
10. TAX: 每万元的房产税率
11. PTRATIO: 城镇学生教师比例
12. B: $1000(B_k - 0.63)^2$ 其中 B_k 是城镇中黑人比例
13. LSTAT: 低收入人群比例
14. MEDV: 自住房中位数价格, 单位是千元

实验数据



实验过程

- 构建训练集与测试集

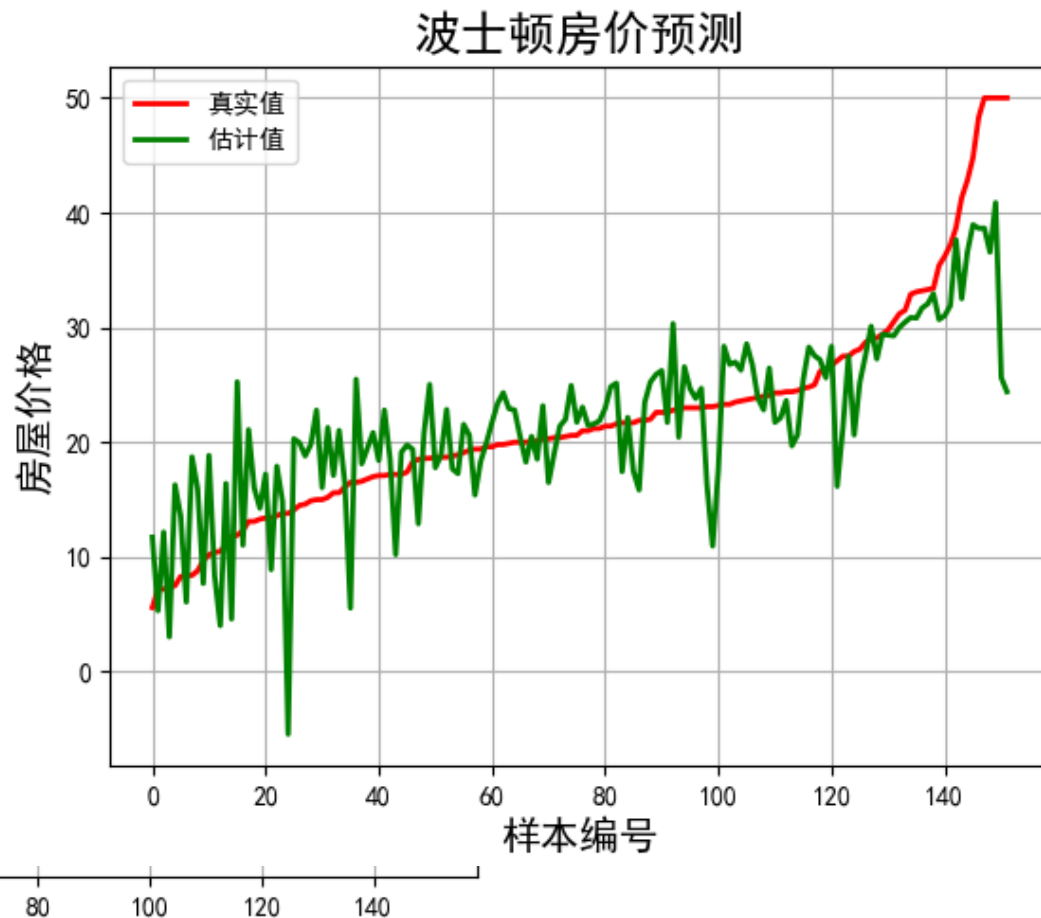
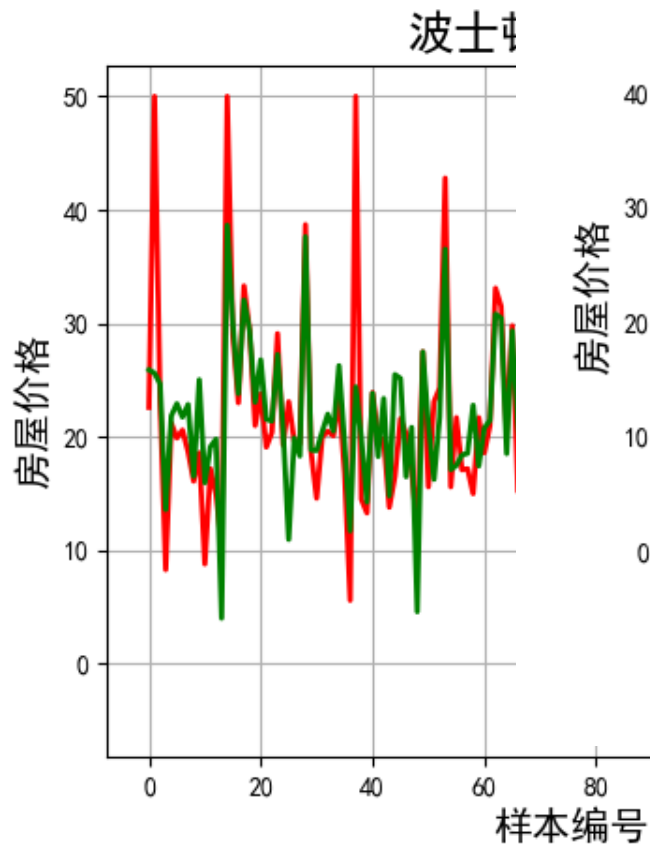
```
from sklearn.model_selection import train_test_split  
x_train, x_test, y_train, y_test = train_test_split(x, y,  
train_size=0.7, random_state=1)
```

- **sklearn** 的线性回归

```
from sklearn.linear_model import LinearRegression  
  
linreg = LinearRegression()  
model = linreg.fit(X_train, y_train)  
linreg.coef_, linreg.intercept_
```

由此得到各项系数， $y=2.1283+4.9827*RM-0.4667*DIS-0.7282*LSTAT$

实验过程



实验过程

- 回归问题的评价测度

(1)平均绝对误差(Mean Absolute Error, MAE)

(2)均方误差(Mean Squared Error, MSE)

(3)均方根误差(Root Mean Squared Error, RMSE)

Mean Squared Error

mse = np.average((y_hat - np.array(y_test)) ** 2)

Root Mean Squared Error

rmse = np.sqrt(mse)

Acknowledgement

《大数据挖掘》 吉林大学 韩霄松

《机器学习》 吉林大学 吴春国、李瑛

《机器学习》 小象学院 邹博

Thanks