机器学习（本科生公选课）GEC6531

第2节 线性回归 Linear Regression

计算机科学与技术学院

张瑞 教授

邮箱：ruizhang6@hust.edu.cn

签到 & 思考

■ 微助教签到（学校要求）

1. 加入课堂：微信扫码或者通过微助教公众号

2. 微信扫码签到

在以下例子中，哪个适合用无监督学习算法来解决?

1.给定一个客户数据库，自动发现细分市场并将客户分到不同的细分市场中。

2.给定标记为垃圾邮件/非垃圾邮件的电子邮件，学习制作一个垃圾邮件过滤器。

3.给定一组在网页上找到的新闻文章，将它们分组，使得每一组都是关于同一个事件。

4.给定一个数据集，其中的患者被诊断为患有糖尿病或没有糖尿病，学习对新患者进行分类，判断他们是否患有糖尿病。

今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

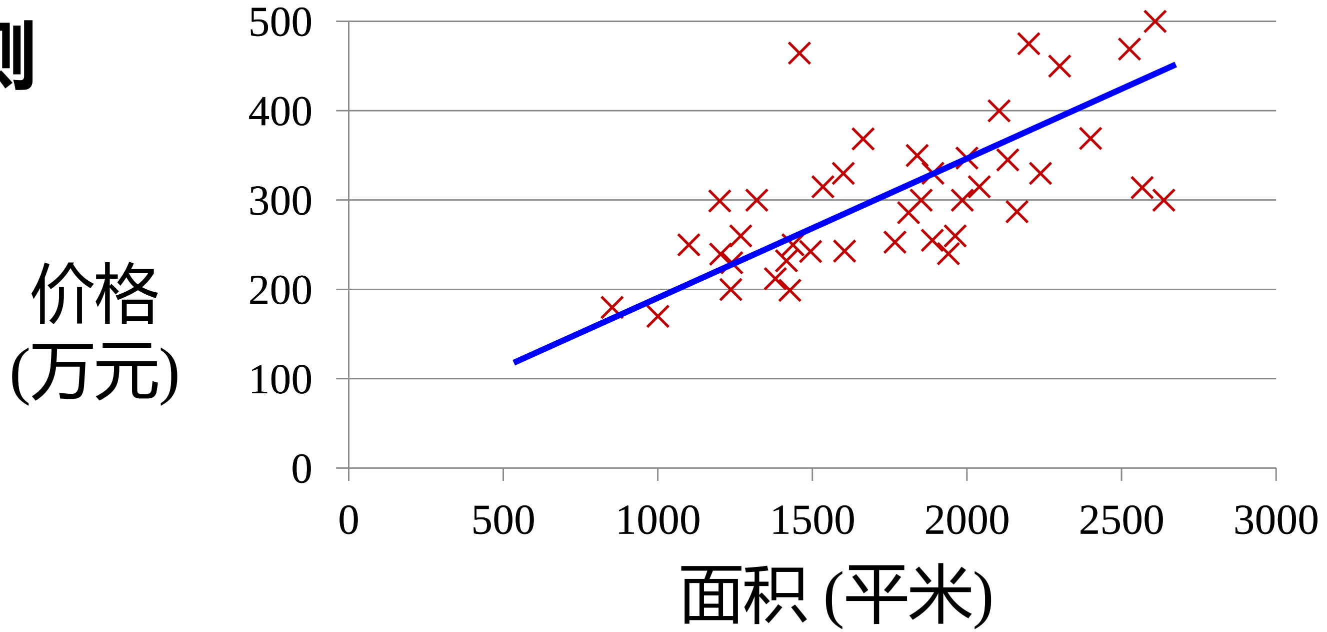
鸣谢：本课程的课件材料很多内容来自于其他材料、包括网络以及Andrew Ng、何琨等老师的课件

今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

一元线性回归 (Linear regression with one variable)

房价预测



<u>监督学习</u>

对每个房子根据面积预测价格

<u>回归问题</u>

预测一个值

一元线性回归 (Linear regression with one variable)

训练集

面积 (x)	房价 (y)
2104	460
1416	232
1534	315
852	178
…	…

$m = 47$

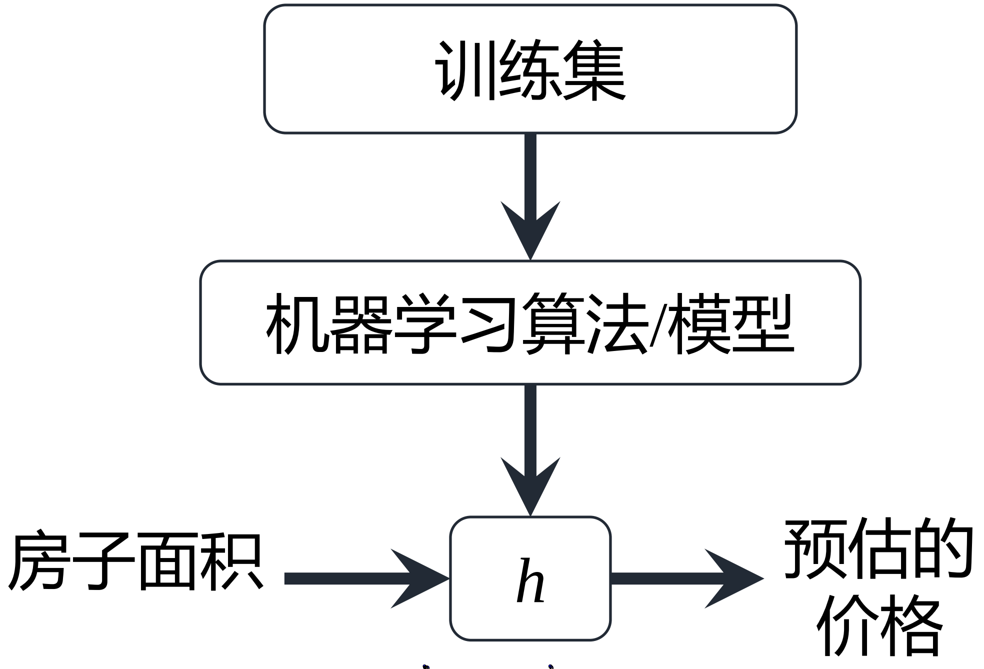
Notation:

m = 训练样本个数 Number of training examples
x’s = 输入变量/特征 “input” variable / features
y’s = 输出变量/目标变量 “output” variable / “target” variable

(x, y) = 一条训练样本
$(x^{(i)}, y^{(i)})$ = 第 i 条训练样本

$x^{(1)} = 2104$
$x^{(2)} = 1416$
$y^{(1)} = 460$

一元线性回归 (Linear regression with one variable)



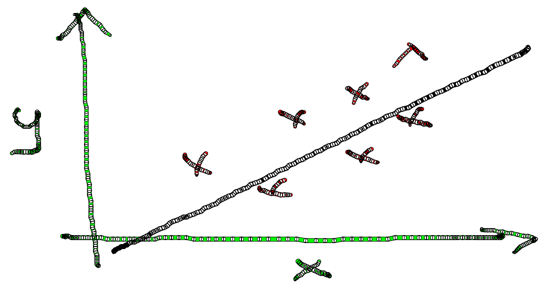
x　　hypothesis　　y

h maps from x's to y's.

如何表示 ***h*** ?

$h_\theta(x) = \theta_0 + \theta_1 x$

Shorthand: $h(x)$



$h(x) = \theta_0 + \theta_1 x$

θ_i : 参数 parameters

如何选参数?

一个变量的线性回归

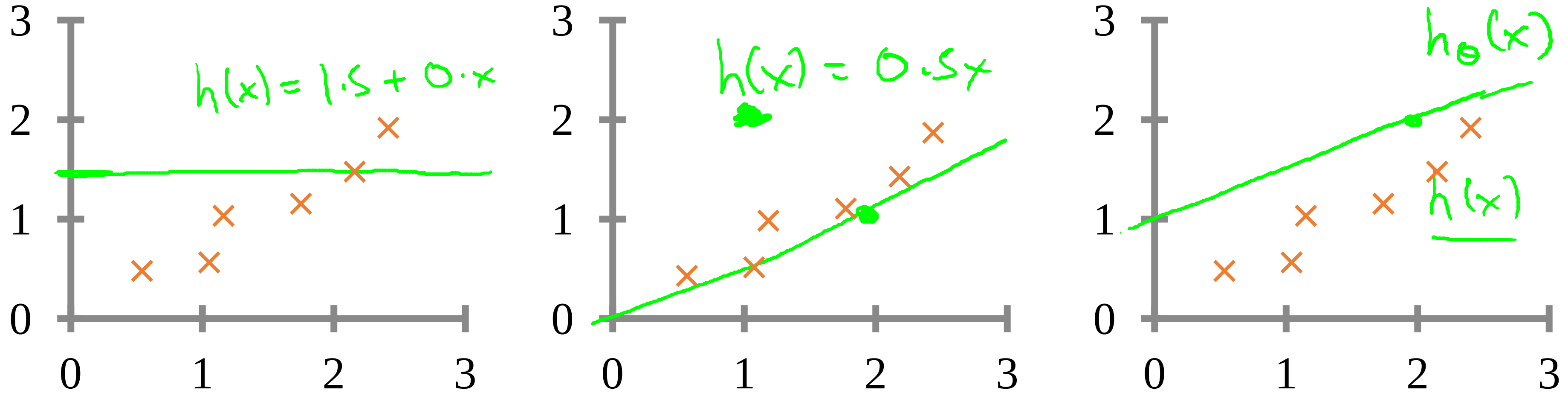
一元线性回归.

└ one variable

(x)

一元线性回归 (Linear regression with one variable)

$$h_\theta(x) = \theta_0 + \theta_1 x$$



$\theta_0 = 1.5$　　$\theta_0 = 0$　　$\theta_0 = 1$

$\theta_1 = 0$　　$\theta_1 = 0.5$　　$\theta_1 = 0.5$

选择 θ_0, θ_1 使得预测值 $h_\theta(x)$ 与实际值 y 在训练集上的平均误差最小

今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

损失函数 (Loss function / Cost function)

假设 Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

参数 Parameters: θ_0, θ_1

损失函数 Loss/Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

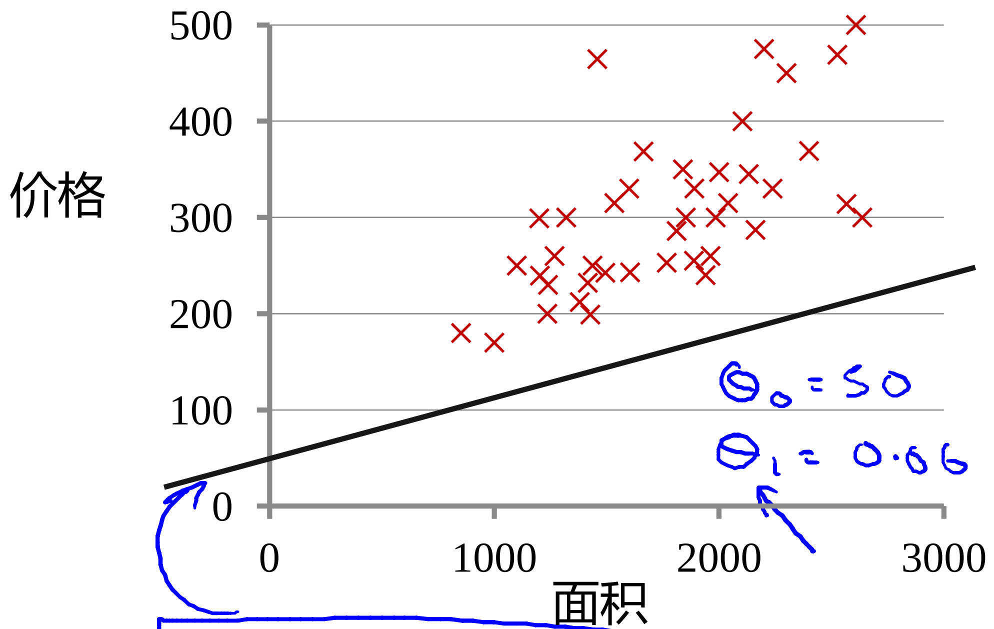
目标 Objective: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

L_p norm (p 范数) $\quad ||x_p|| = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \quad \textit{where } p > 0$

直观理解 (Intuition)

$h_\theta(x)$

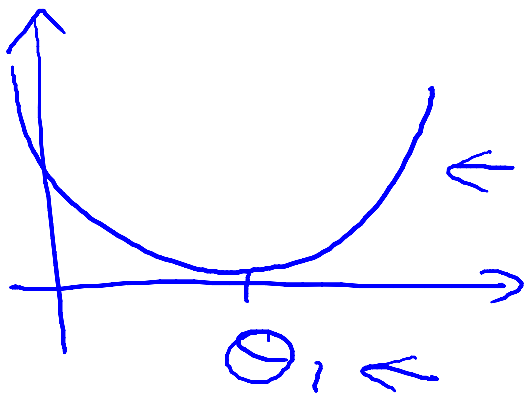
(for fixed θ_0, θ_1 , this is a function of x)



$$\boxed{h_\theta(x) = 50 + 0.06x}$$

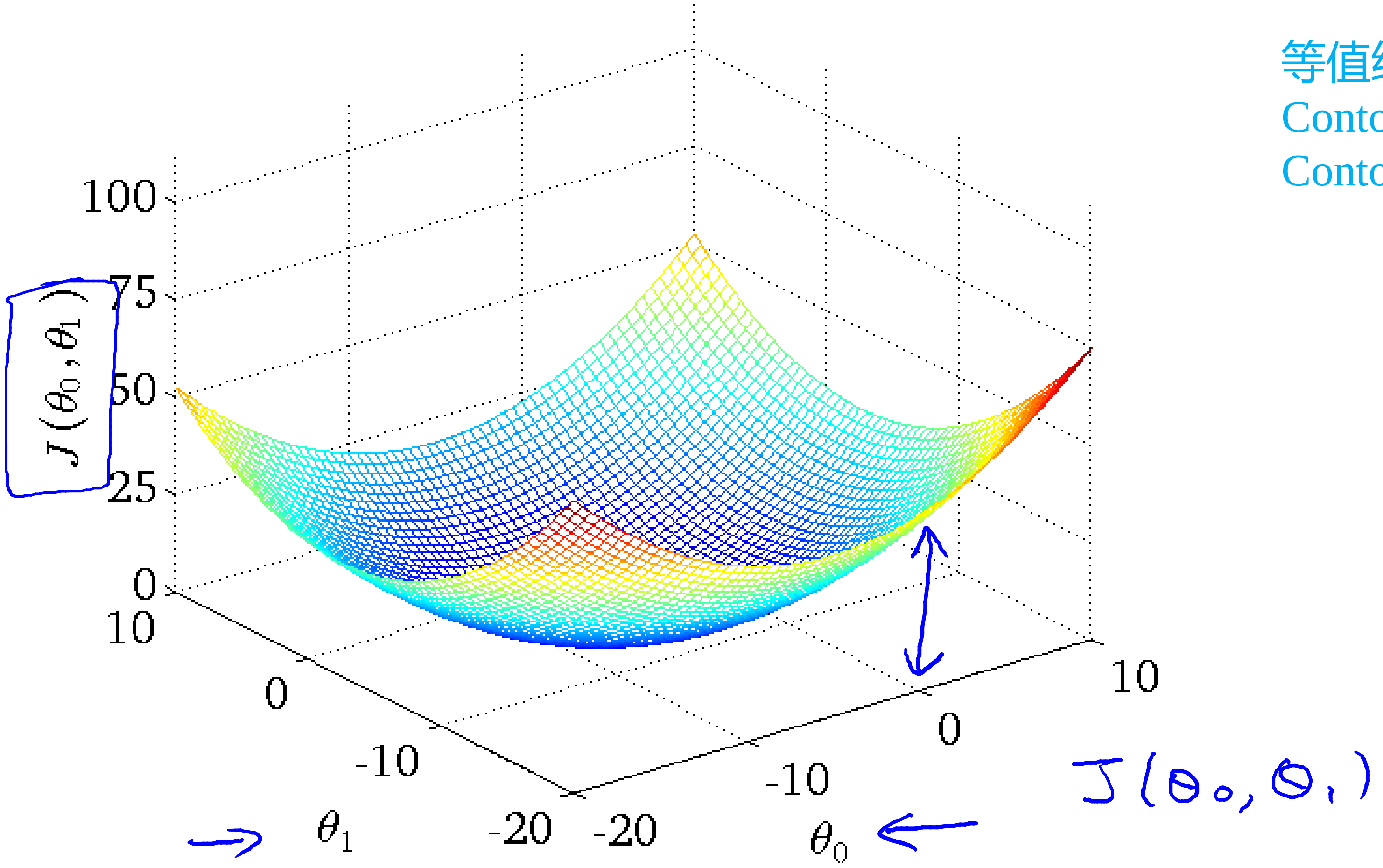
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



$$\boxed{\theta_0 , \theta_1}$$

直观理解 (Intuition)

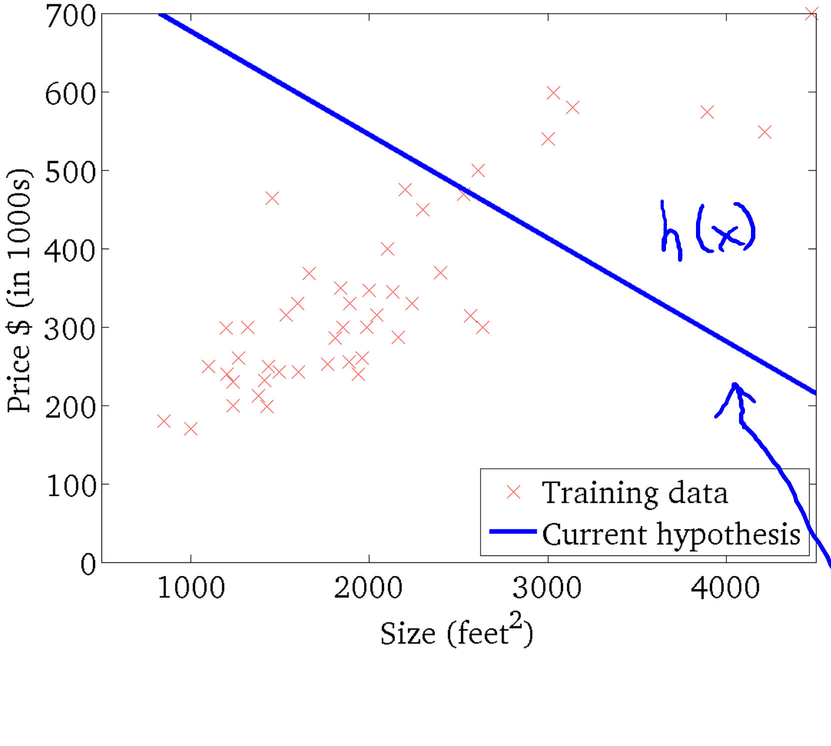


等值线图/等高线图
Contour plot
Contour figure

直观理解 (Intuition)

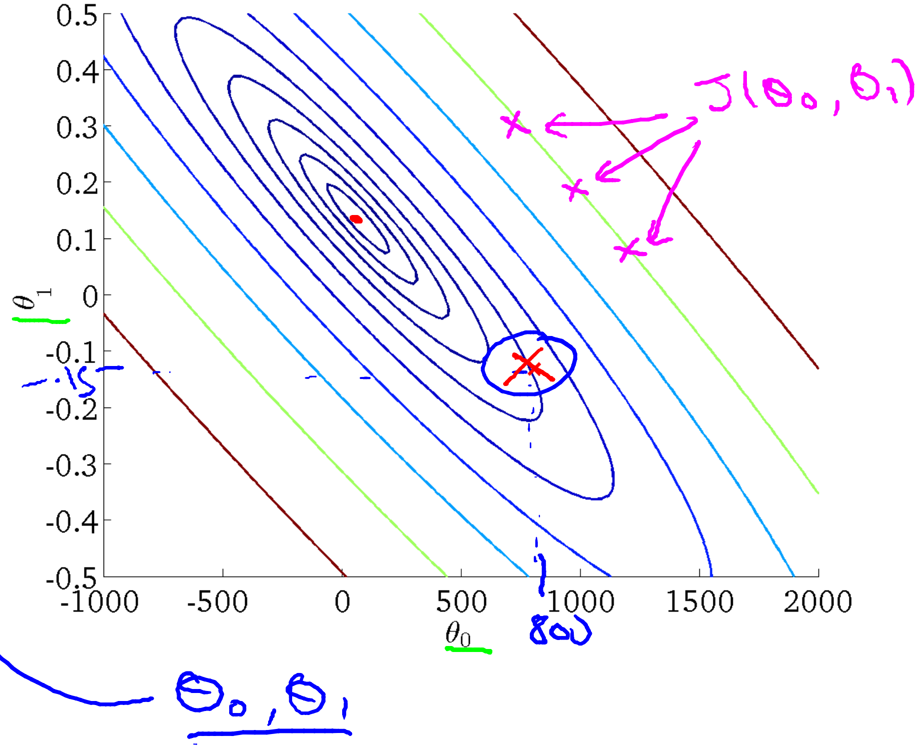
$$h_\theta(x)$$

(for fixed θ_0, θ_1 , this is a function of x)



$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



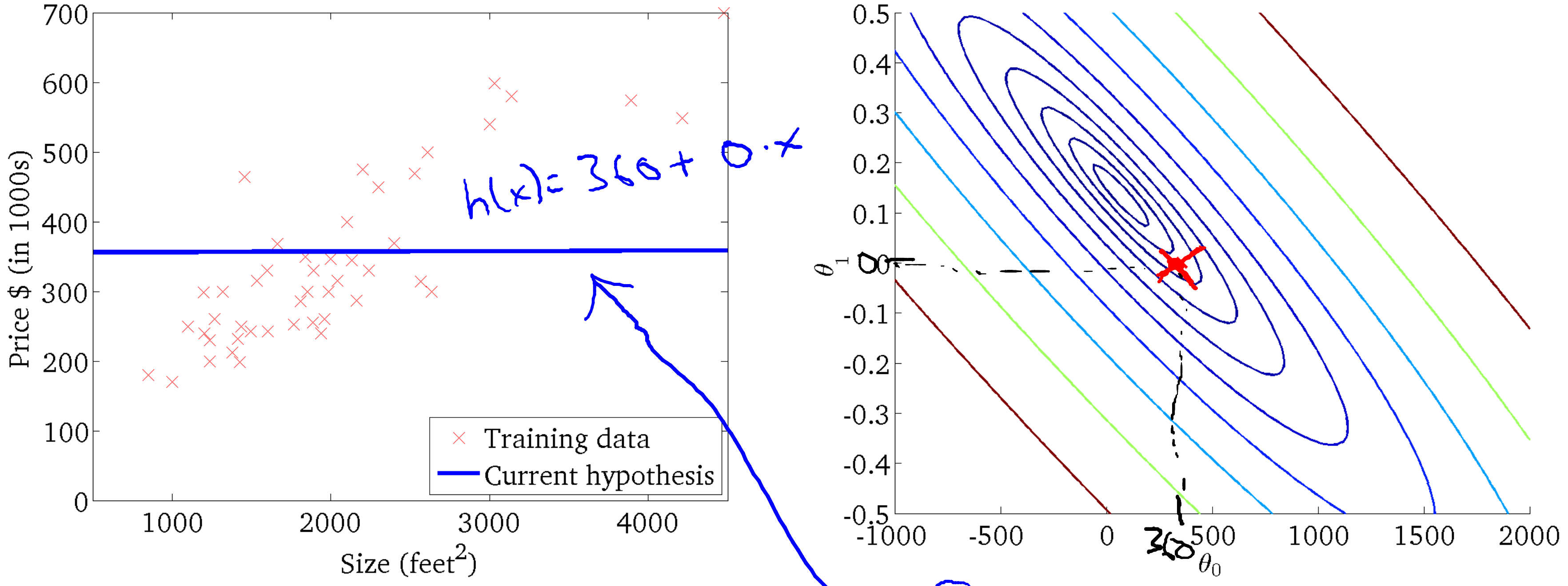
直观理解 (Intuition)

$h_\theta(x)$

(for fixed θ_0, θ_1 , this is a function of x)

$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



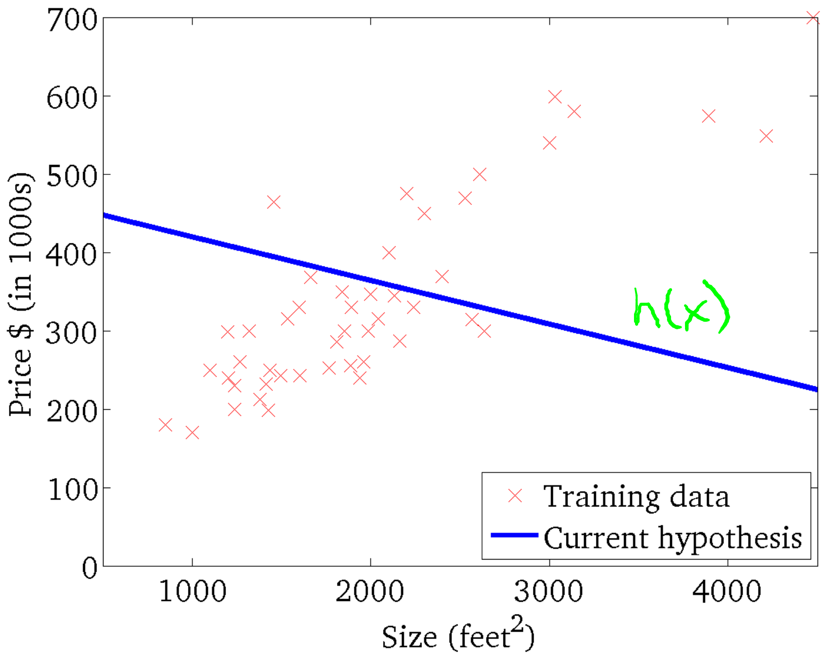
$\theta_0 = 360$

$\theta_1 = 0$

直观理解 (Intuition)

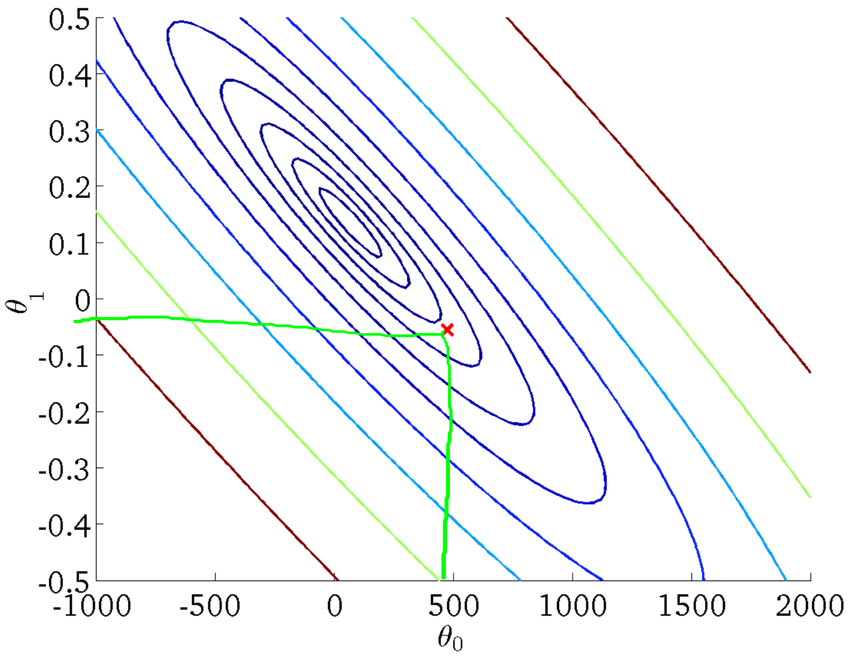
$h_\theta(x)$

(for fixed θ_0, θ_1 , this is a function of x)



$J(\theta_0, \theta_1)$

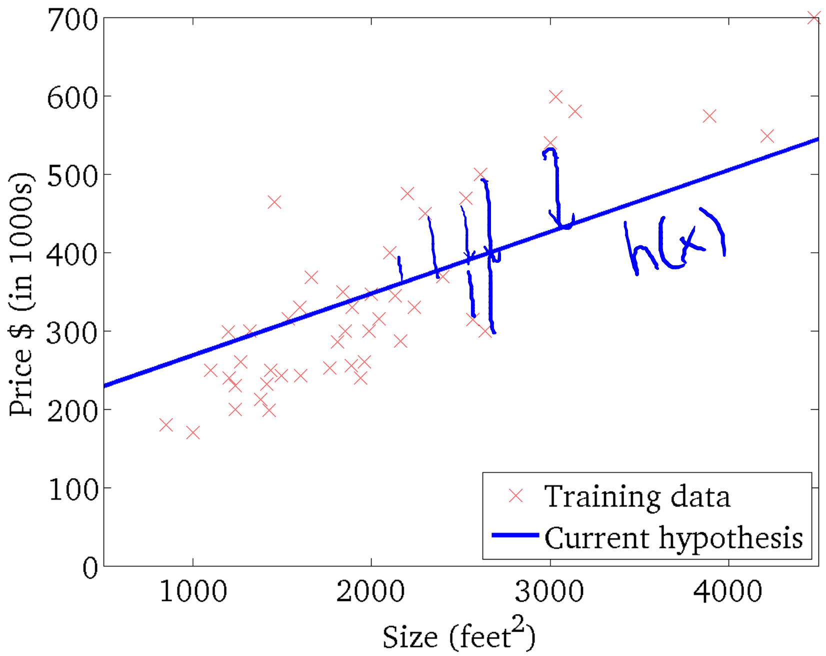
(function of the parameters θ_0, θ_1)



直观理解 (Intuition)

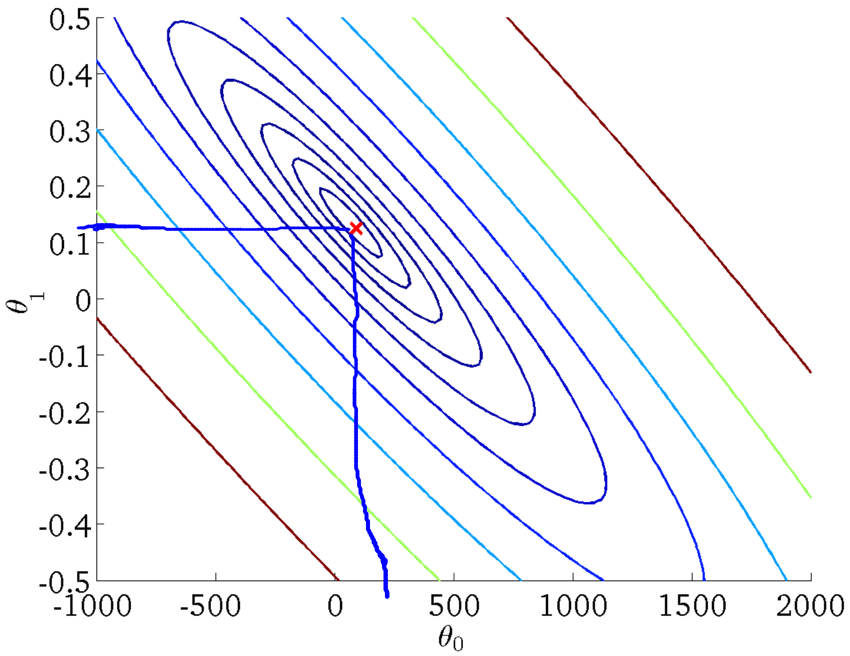
$h_\theta(x)$

(for fixed θ_0, θ_1 , this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

梯度下降预览 (Gradient Descent)

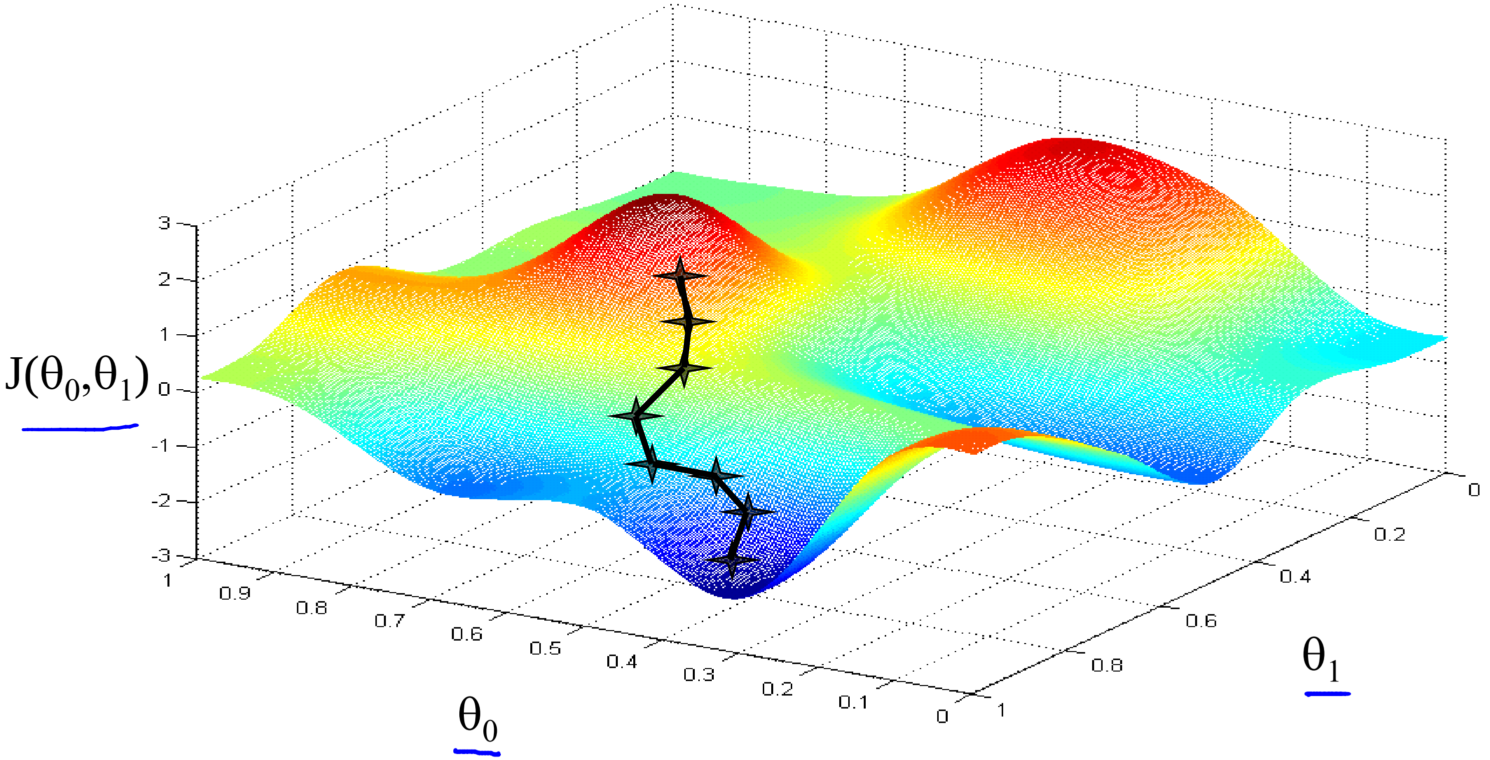
损失函数 $J(\theta_0, \theta_1)$

目标 $\min\limits_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

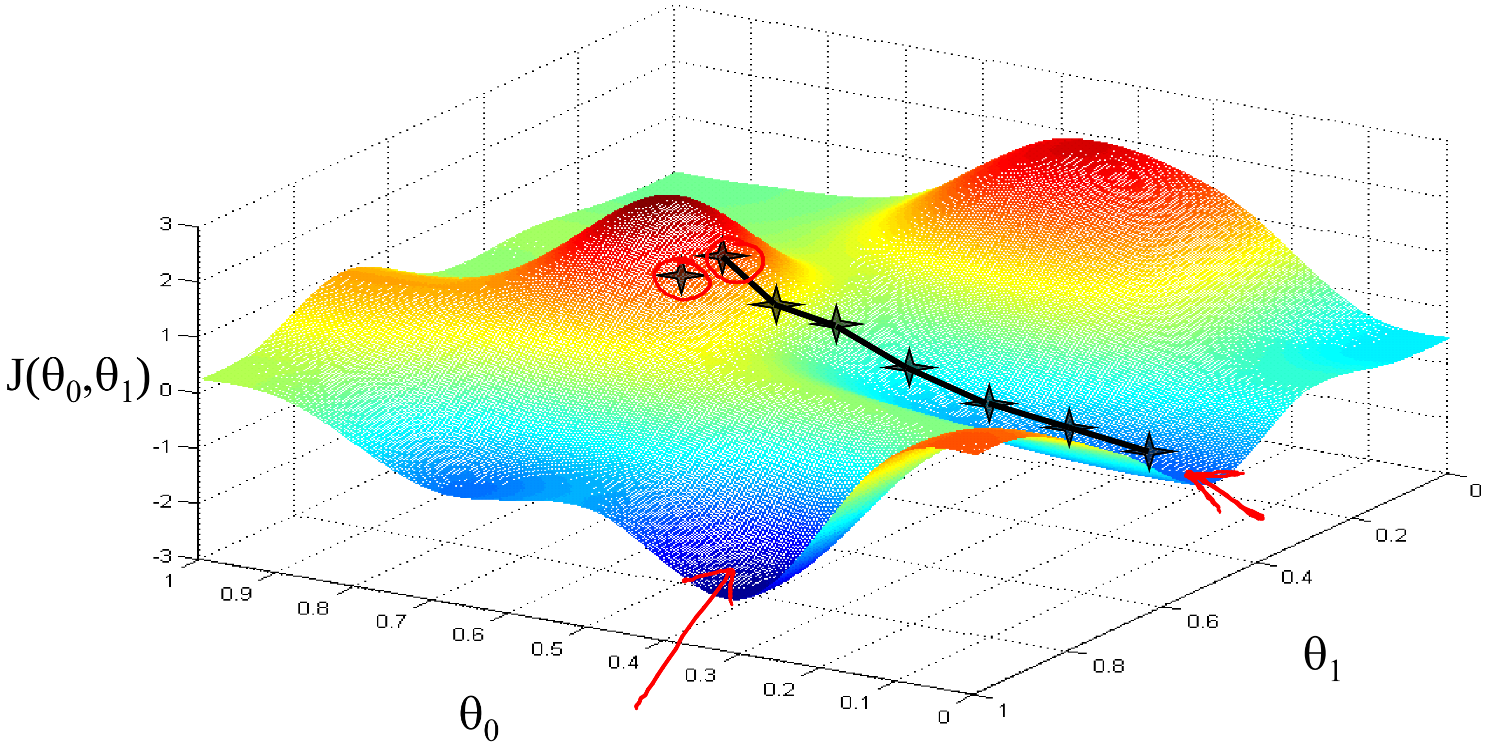
概览:

- 以某个值开始 θ_0, θ_1
- 不停更新 θ_0, θ_1 来减小 $J(\theta_0, \theta_1)$
- 直到我们认为达到了最小值

梯度下降预览 (Gradient Descent)



梯度下降预览 (Gradient Descent)



今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

用最大似然估计（MLE）求参

Maximum-likelihood estimation (MLE)

$$
\begin{aligned}
\mathbf{w} &= \underset{\mathbf{w}}{\operatorname{argmax}}\, P(y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n | \mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i, \mathbf{x}_i | \mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i | \mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left[P(y_i | \mathbf{x}_i, \mathbf{w}) \right] \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \left(e^{-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}} \right) \right]
\end{aligned}
$$

用最大似然估计（MLE）求参

$$= \underset{\mathbf{w}}{\operatorname{argmax}} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{w} - y_i)^2$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{w} - y_i)^2$$

- 我们最小化损失函数，$\ell(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{w} - y_i)^2$。这个特殊的损失函数也被称为均方损失或最小二乘损失 (Ordinary Least Squares)。最小二乘损失可以用梯度下降法、牛顿法或闭式解进行优化。

  闭式解形式 $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$，其中 $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ 且 $\mathbf{y} = [y_1, \ldots, y_n]^\top$.

用最大似然估计（MLE）求参

$$\ell(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

$$= (X^T w - y)^T (X^T w - y)$$

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n], \mathbf{y} = [y_1, \ldots, y_n]^\top$$

$$\Rightarrow \hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y} \in \mathbb{R}^{d\times 1}, \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}, \mathbf{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n\times 1}.$$

可利用公式：

$$\frac{\partial Ax}{\partial x} = A^T$$

$$\frac{\partial x^T A}{\partial x} = A$$

$$\frac{\partial A^T x B}{\partial x} = AB^T$$

$$\frac{\partial A^T x^T B}{\partial x} = BA^T$$

今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

用最大后验估计（MAP）求参

Max a posteriori (MAP)

$$
\begin{aligned}
\mathbf{w} &= \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{w}|y_1, \mathbf{x}_1, ..., y_n, \mathbf{x}_n) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}}\, \frac{P(y_1, \mathbf{x}_1, ..., y_n, \mathbf{x}_n|\mathbf{w})P(\mathbf{w})}{P(y_1, \mathbf{x}_1, ..., y_n, \mathbf{x}_n)} \\
&= \underset{\mathbf{w}}{\operatorname{argmax}}\, P(y_1, \mathbf{x}_1, ..., y_n, \mathbf{x}_n|\mathbf{w})P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \left[\prod_{i=1}^{n} P(y_i, \mathbf{x}_i|\mathbf{w})\right] P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})P(\mathbf{x}_i|\mathbf{w})\right] P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})P(\mathbf{x}_i)\right] P(\mathbf{w})
\end{aligned}
$$

附加模型假设: $P(\mathbf{w}) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\mathbf{w}^\top \mathbf{w}}{2\tau^2}}$

用最大后验估计（MAP）求参

$$
\begin{aligned}
&= \underset{\mathbf{w}}{\operatorname{argmax}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) \right] P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i, \mathbf{w}) + \log P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \frac{1}{2\tau^2} \mathbf{w}^\top \mathbf{w} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \lambda ||\mathbf{w}||_2^2 \qquad \lambda = \frac{\sigma^2}{n\tau^2}
\end{aligned}
$$

- 这个目标被称为岭回归。它有一个闭式解：$\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1}\mathbf{X}\mathbf{y}$，其中 $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ 且 $\mathbf{y} = [y_1, \ldots, y_n]^\top$.

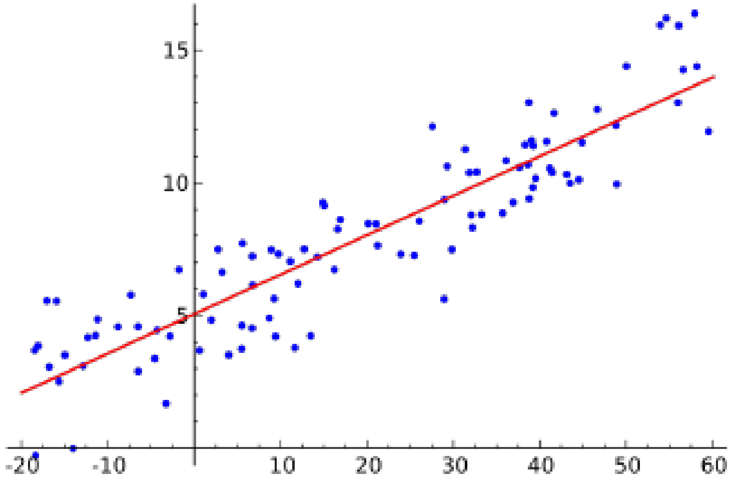
今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

简单（一元）和多重线性回归

简单和多重线性回归

单一标量自变量 x 和单一标量因变量 y 的最简单情况被称为简单线性回归。
扩展到多个和/或向量值预测变量（用大写 X 表示）被称为多重线性回归，也称为多变量线性回归。
几乎所有现实世界的回归模型都涉及多个预测因子，线性回归的基本描述通常是用多重回归模型来描述的。



广义线性模型

广义线性模型

广义线性模型（Generalized linear models，GLMs）是一种为有界或离散的因变量建模的框架。例如:

- 当对变化幅度较大的正值变量（例如价格或人口）建模时，这些正值量能被偏态分布（skewed distribution，频数分布的高峰位于一侧，尾部向另一侧延伸的分布）描述，如对数正态分布或泊松分布（尽管 GLMs 不用于对数正态数据，但是因变量利用对数函数进行了简单的转换）;
- 对分类（可枚举）数据建模时，例如在选举中对给定候选人的选择（对于二元选择使用伯努利分布/二项分布可更好地描述，对于多元选择使用分类分布/多项分布），其中这些数量固定的选择不能被有意义地排序;
- 当对有序数据建模时，例如从 0 到 5 的评分，不同的结果可以排序，但数量本身可能没有任何绝对意义（例如评分 4 可能不是评分 2 在客观意义上的“两倍好”，而只是表明它比 2 或 3 好，但不如 5）。

层次线性模型

层次线性模型

层次线性模型（或多层回归）将数据组织成一个回归的层次结构（Hierarchy），例如 A 在 B 上回归，B 在 C 上回归。

它通常用于变量具有自然的层次结构，例如在教育统计中，学生嵌套在教室中，教室嵌套在学校中，学校嵌套在一些学区中。

因变量可以是学生成绩的衡量标准，如考试分数，不同的协变量将在教室、学校和学区级别收集。

今天的目录

- **一元线性回归**
  - 简介
  - 损失函数
  - 梯度下降预览
- **用最大似然估计（MLE）求参**
- **用最大后验估计（MAP）求参**
- **扩展**
  - 简单和多重线性回归
  - 广义线性模型
  - 层次线性模型
- **应用**

应用

金融

资本资产定价模型使用线性回归和 beta 的概念来分析和量化投资的系统风险。这直接来自线性回归模型的 beta 系数，该模型将投资回报与所有风险资产的回报联系起来。

经济学

线性回归是经济学中主要的实证工具。可被用来预测消费支出、固定投资支出、库存投资、购买一个国家的出口产品、进口支出、持有流动资产的需求、劳动力需求和劳动力供给。

环境科学

线性回归在环境科学中有着广泛的应用。在加拿大，环境影响监测方案使用对鱼类和底栖生物调查的统计分析来衡量纸浆厂或金属矿山废水对水生生态系统的影响。

机器学习

线性回归在机器学习等人工智能领域发挥着重要作用。线性回归算法因其相对简单和广为人知的特性而成为监督机器学习的基本算法之一。

总结

- **MLE方法：**

  $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$
  均方损失
  没有正则项
  闭式解形式: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$

  - **最小二乘法**

- **MAP方法、岭回归**

  $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \lambda||\mathbf{w}||_2^2$
  均方损失
  ℓ_2-regularization
  闭式解形式: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$