电子科技大学

University of Electronic Science and Technology of China

Generalized linear model

Jiaming Liu

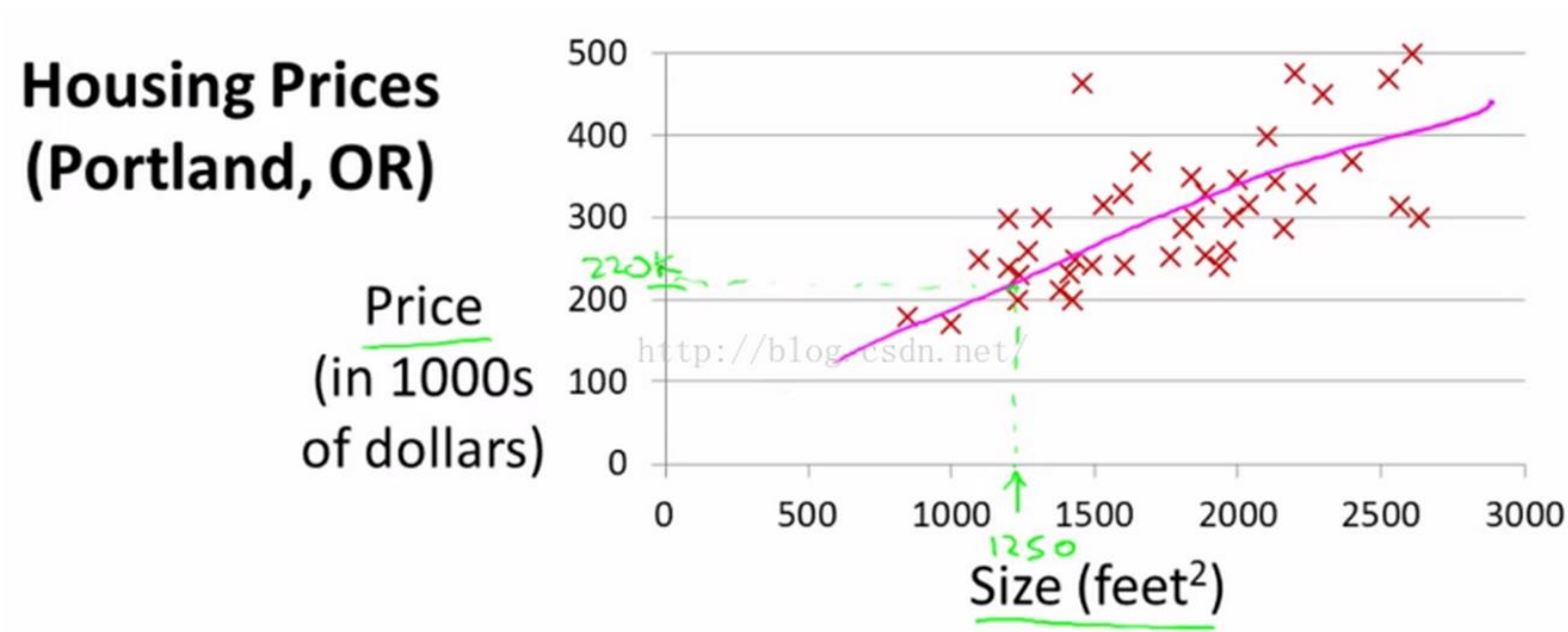
Data Mining Lab,
Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn

Outline

- Linear regression
- Classification and logistic regression
- Generalized linear model

Part Ⅰ. Linear regression

Linear regression



hypothesis function: $h_\theta(x) = \theta_0 + \theta_1 x$

Hypothesis function

What if we have more features?

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

let $x_0 = 1$

How to learn this model?

Find a set of θ so that $h(x)$ is close to given examples.

Cost function

So we define the cost function:

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

We want to choose θ so as to minimize $J(\theta)$.

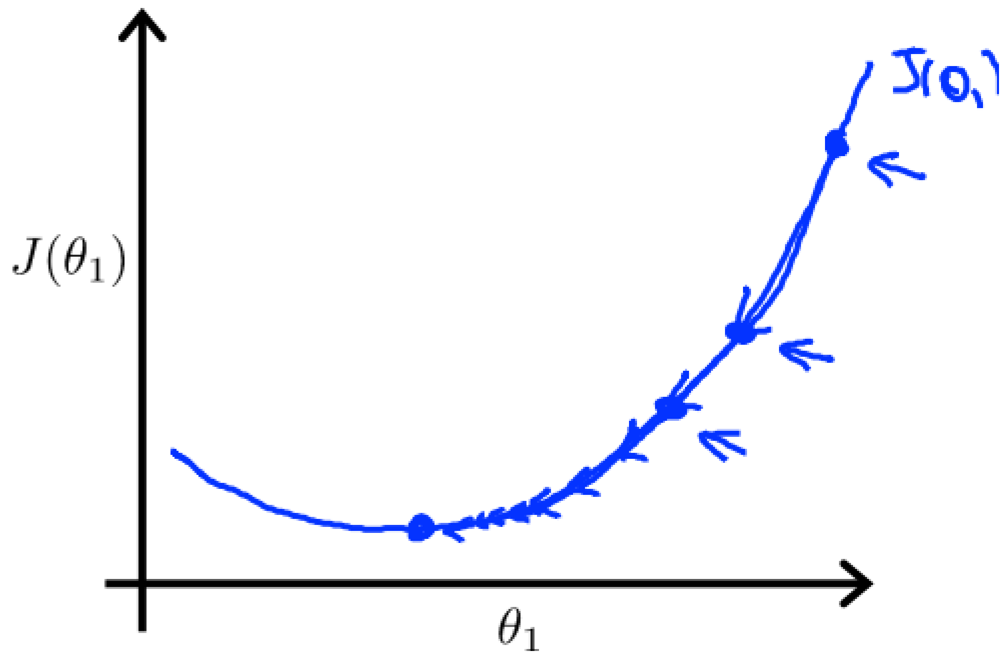
Question1. Why we choose the ordinary least squares method?
Question2. Where does the $'\frac{1}{2}'$ come from?

Gradient descent

Gradient descent can converge to a local minimum, even with the learning rate α fixed.

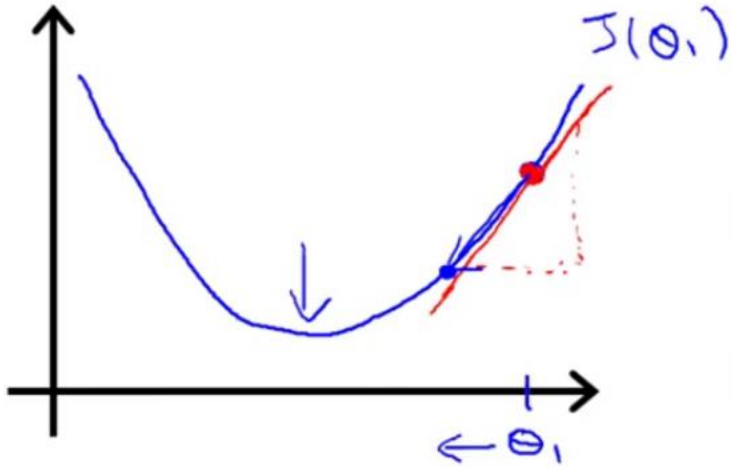
$$\theta_1 := \theta_1 - \alpha \boxed{\frac{d}{d\theta_1} J(\theta_1)}$$

As we approach a local minimum, gradient descent will automatically take smaller steps. So, no need to decrease α over time.



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

Gradient descent



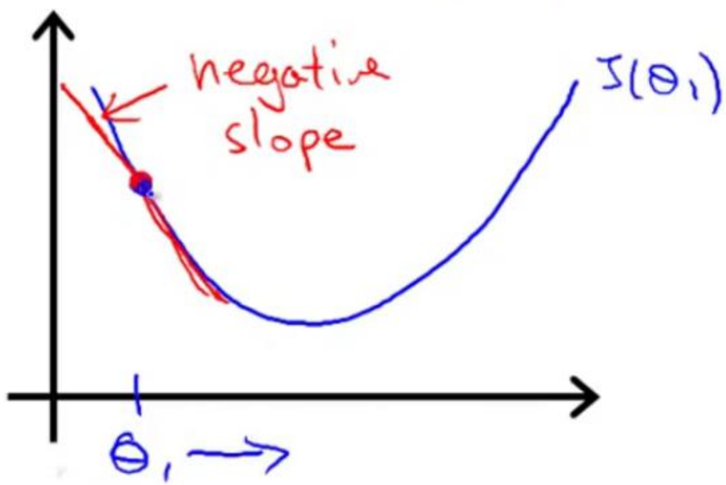
$(\theta_1 \in \mathbb{R})$

$\frac{\partial}{\partial \theta_1}$

$$\theta_1 := \theta_1 - \alpha \boxed{\frac{d}{d\theta_1} J(\theta_1)}$$

≥ 0

$$\theta_1 := \theta_1 - \underline{\alpha} \cdot (\text{positive number})$$



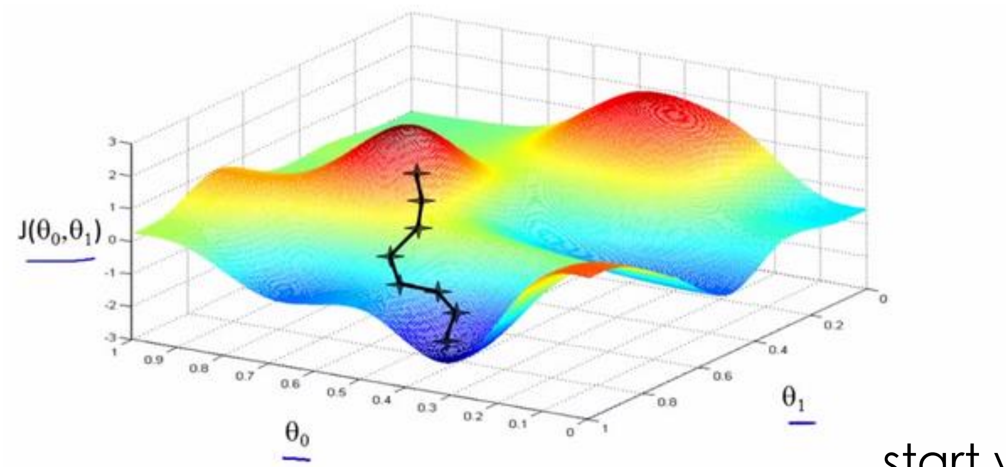
$$\frac{d}{d\theta_1} J(\theta_1)$$

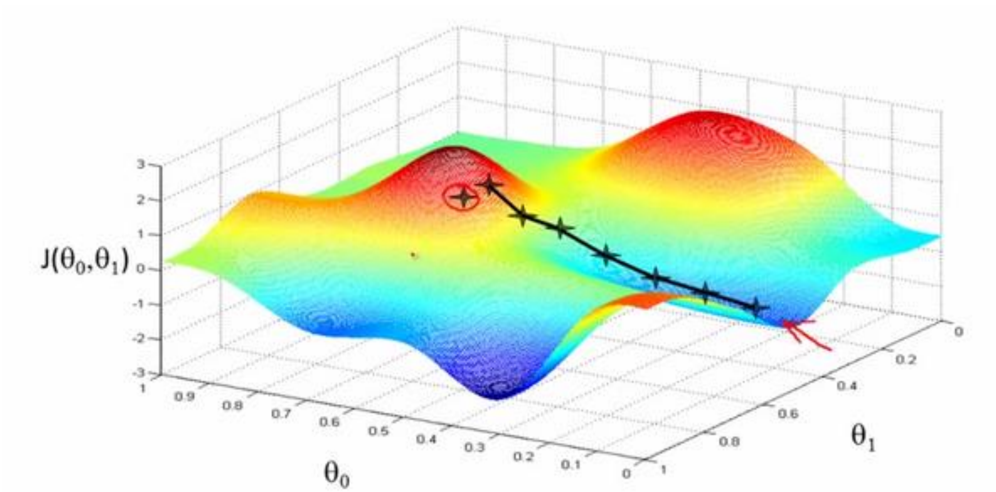
≤ 0

$$\theta_1 := \underline{\theta_1} - \alpha\,(\text{negative number})$$

Andrew Ng

Gradient descent

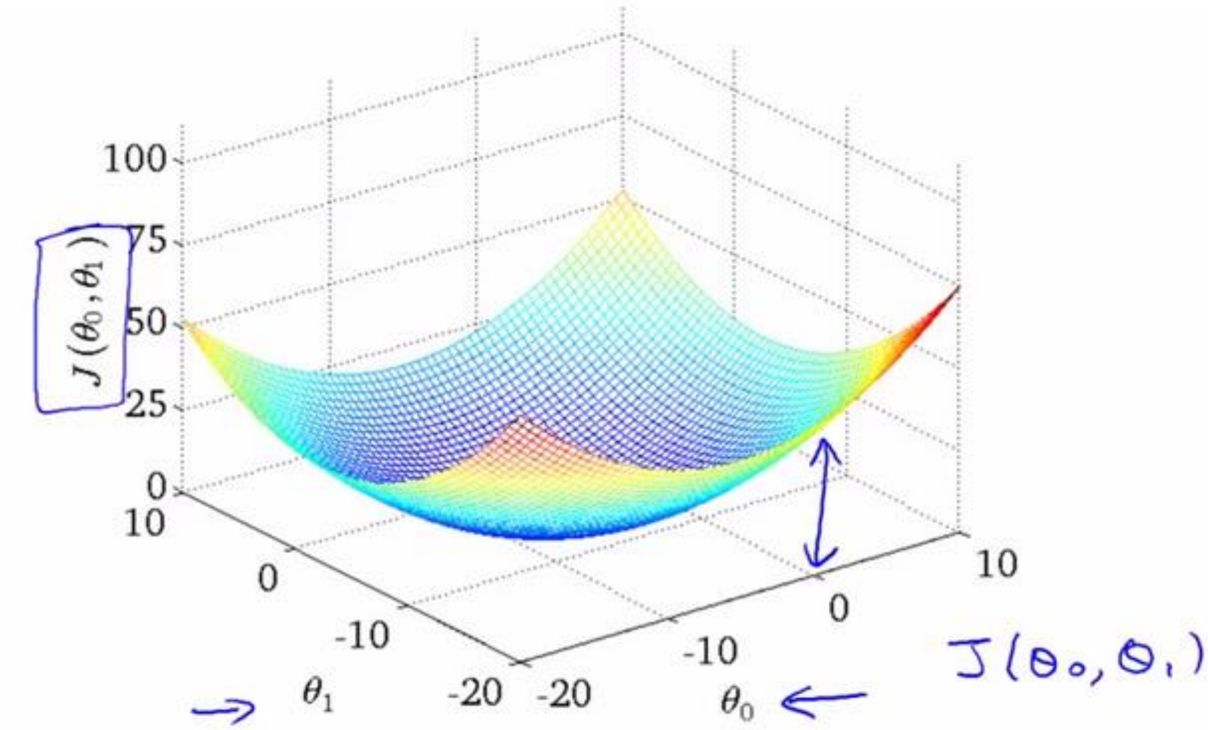


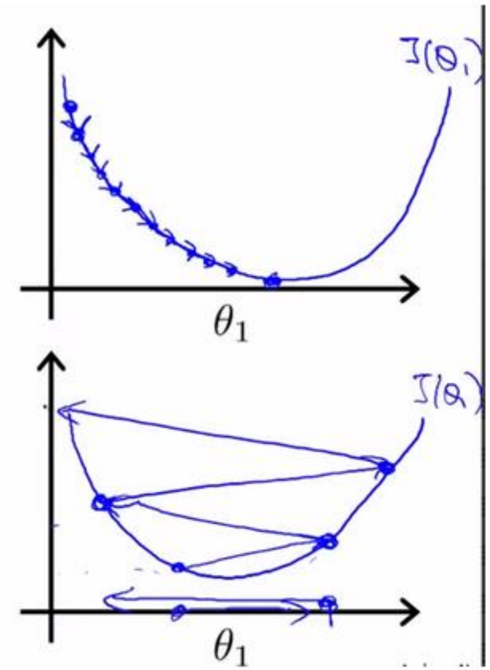


start with diff starts , end in diff ends
local optimum algorithm

Gradient descent

数据挖掘实验室
Data Mining Lab





$J(\theta)$ is a quadratic function—global optimum

Different alpha

Batch gradient descent

数据挖掘实验室
Data Mining Lab

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

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

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

Repeat until convergence {

$$\theta_j := \theta_j + \alpha\sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)} \qquad \text{(for every } j\text{)}.$$

}

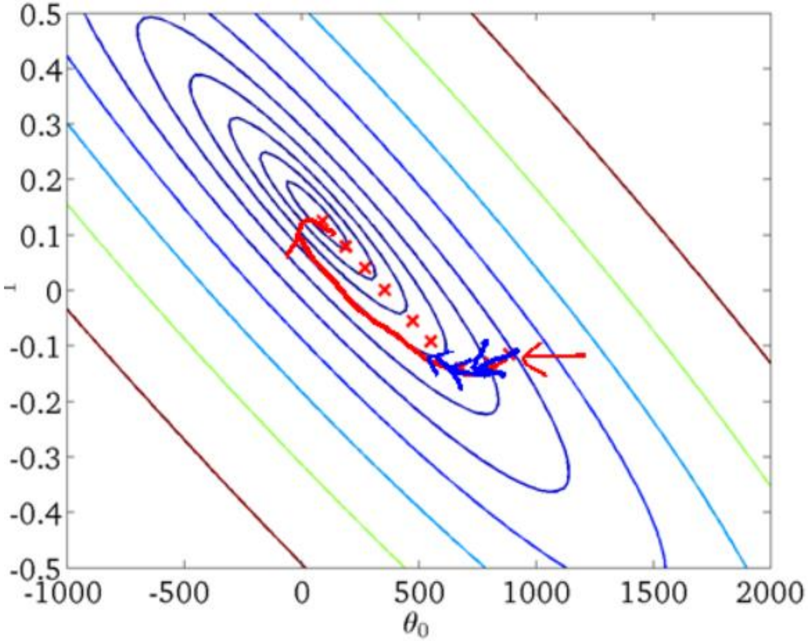
Batch gradient descent

Stochastic gradient descent

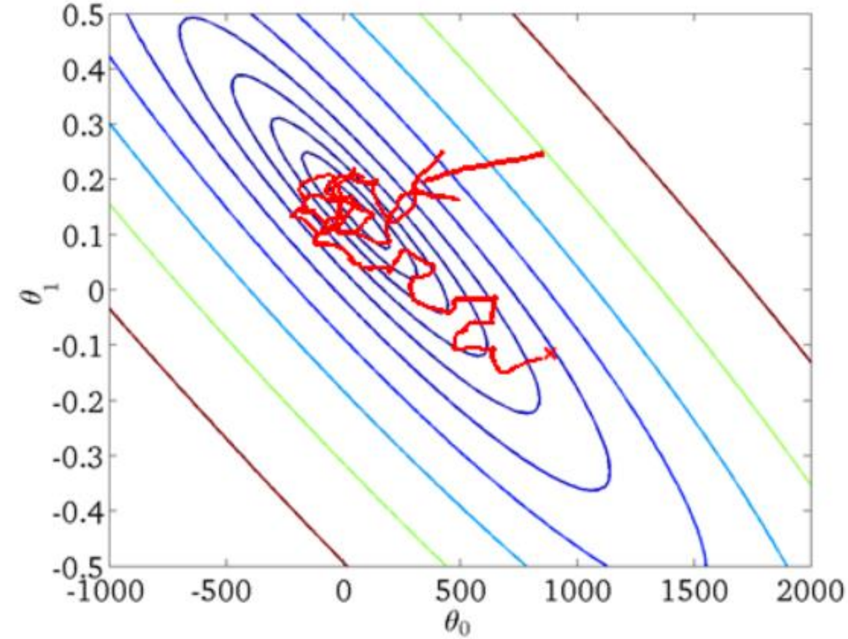
$$
\begin{aligned}
&\text{Loop } \{ \\
&\qquad \text{for i=1 to m, } \{ \\
&\qquad\qquad \theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)} \qquad (\text{for every } j). \\
&\qquad \} \\
&\}
\end{aligned}
$$

1. Batch gradient descent has to scan through the entire training set before taking a single step—a costly operation if m is large;
2. Stochastic gradient descent gets θ "close" to the minimum much faster than batch gradient descent;
3. The parameters θ will keep oscillating around the minimum of J(θ);
4. Particularly when the training set is large, stochastic gradient descent is often preferred over batch gradient descent.

Stochastic gradient descent



BGD



SGD

Probabilistic interpretation

- Why might the least-squares cost function J, be a reasonable choice?
- Let us assume that the target variables and the inputs are related via the equation:

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

assume that $\epsilon^{(i)} \sim N(0, \sigma^2).$

- According to *central limit theorem*, for the most commonly studied scenarios, when independent random variables are added, their sum tends toward a normal distribution.

Probabilistic interpretation

The density of $\epsilon^{(i)}$ is given by

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

This implies that

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

Probabilistic interpretation

数据挖掘实验室
Data Mining Lab
LESS IS MORE

likelihood function:

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

log likelihood:

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

Probabilistic interpretation

Hence, maximizing $\ell(\theta)$ gives the same answer as minimizing

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

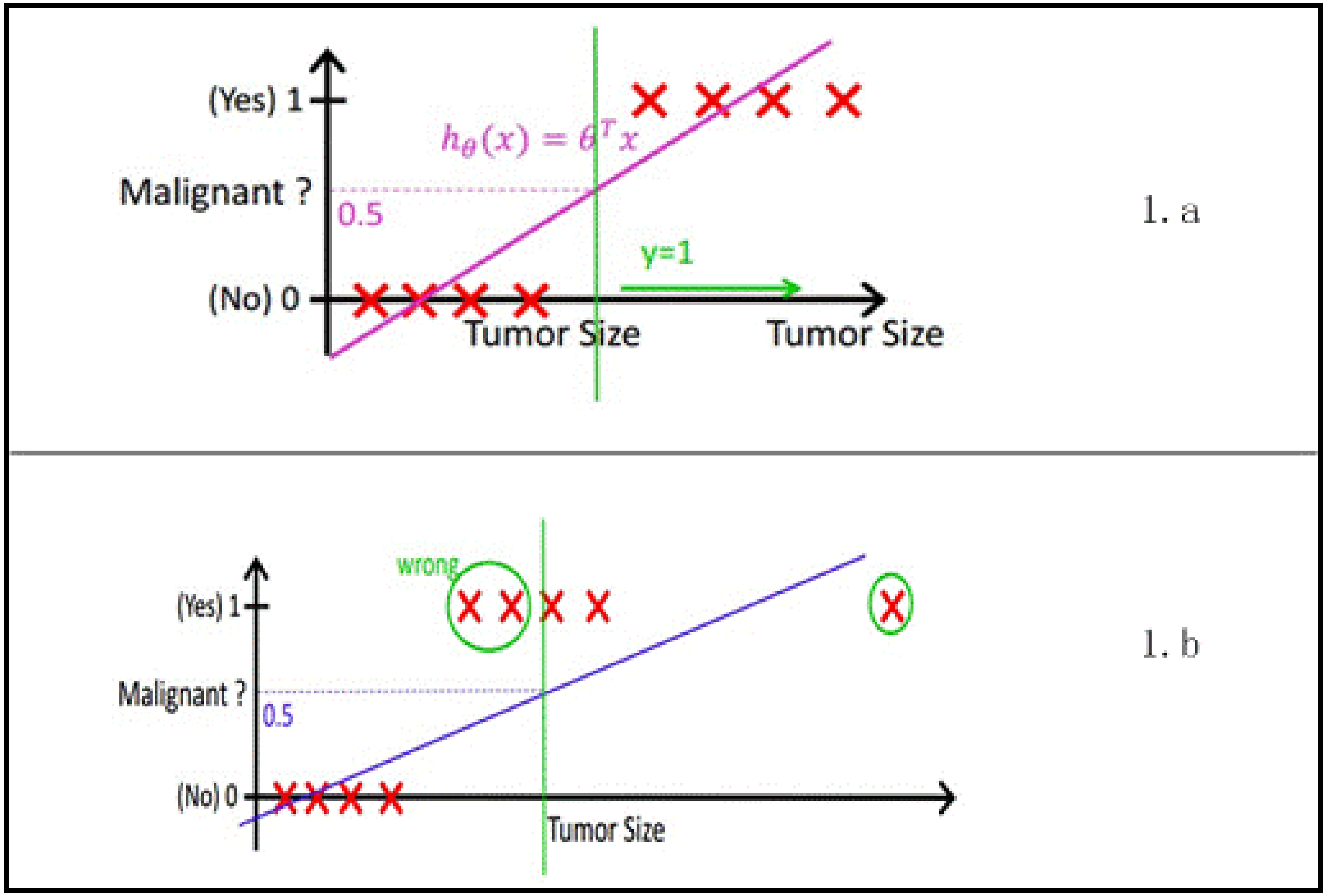
which we recognize to be $J(\theta)$, our original least-squares cost function.

To summarize: When errors follow a Gaussian distribution, least-squares regression can be justified as a very natural method that's just doing maximum likelihood estimation.

Note that, our final choice of θ did not depend on what was σ^2.

Part Ⅱ. Classification and logistic regression

Classification and logistic regression



Classification and logistic regression

Classification: $\boxed{y = 0 \text{ or } 1}$

$h_\theta(x)$ can be > 1 or < 0

Logistic Regression: $0 \leq h_\theta(x) \leq 1$

Classification

We need a $h_\theta(x)$ for logistic regression.

Logistic Regression Model

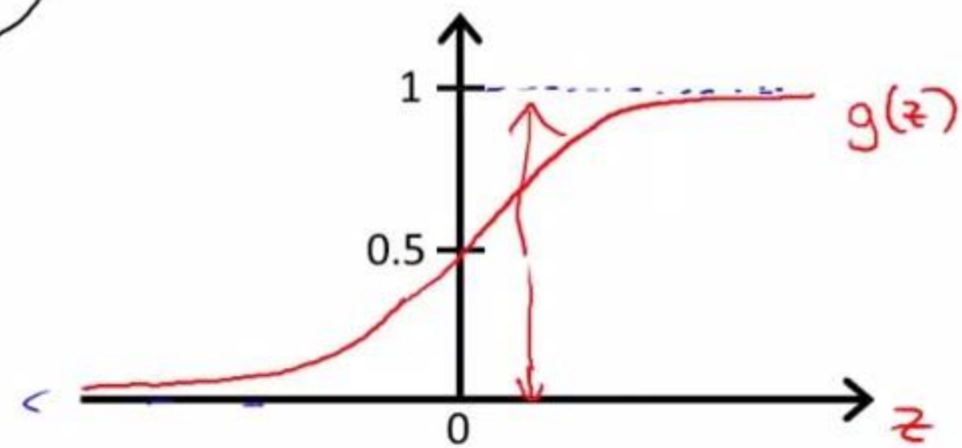
Want $\boxed{0 \leq h_\theta(x) \leq 1}$

$$h_\theta(x) = g(\theta^T x)$$

$$\rightarrow g(z) = \frac{1}{1+e^{-z}}$$

$\theta^T x$

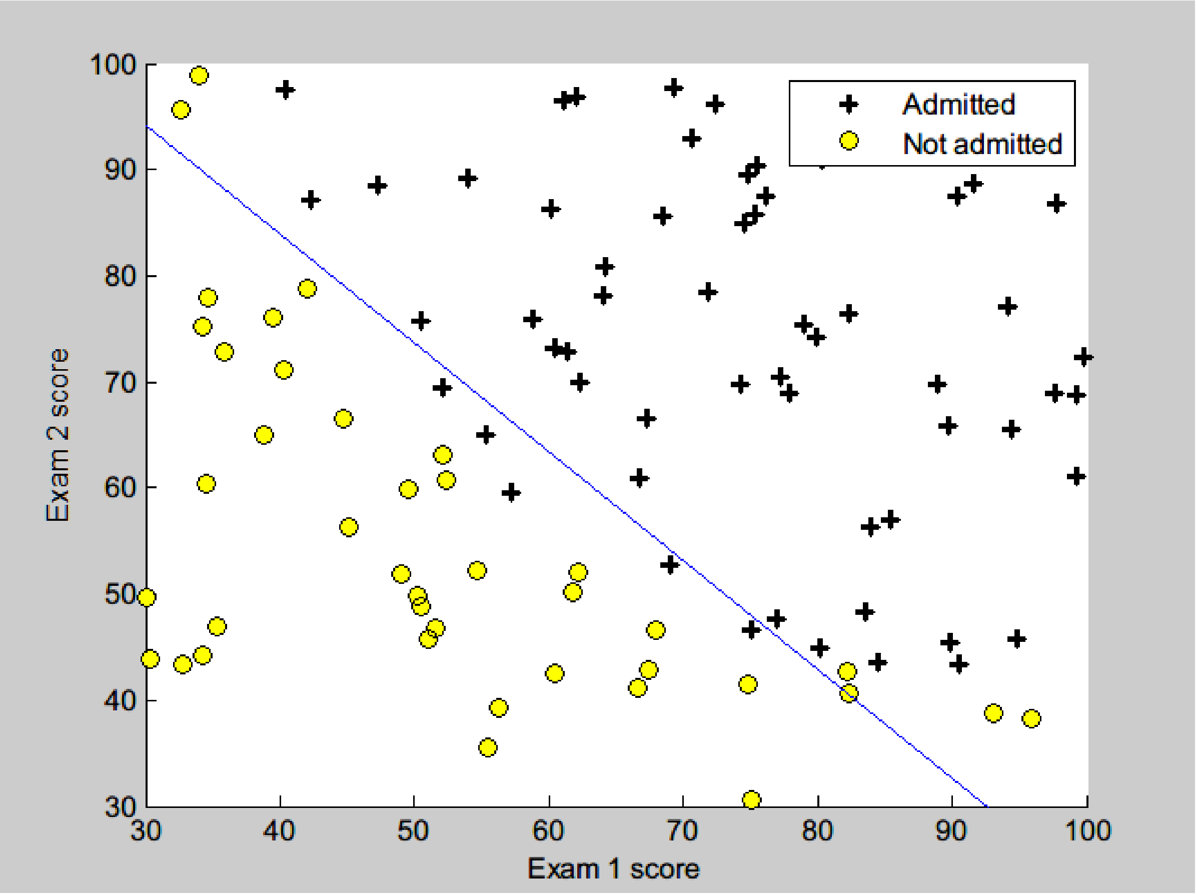
$$\rightarrow h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$



→ Sigmoid function

→ Logistic function

Classification and logistic regression



<Machine learning> ex2 by Andrew Ng

Classification and logistic regression

Let us assume that

$$\begin{aligned} P(y=1 \mid x;\theta) &= h_\theta(x) \\ P(y=0 \mid x;\theta) &= 1-h_\theta(x) \end{aligned}$$

Note that this can be written more compactly as

$$p(y \mid x;\theta) = (h_\theta(x))^y \left(1-h_\theta(x)\right)^{1-y}$$

Write down the likelihood of the parameters as

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

It will be easier to maximize the log likelihood:

$$\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left(y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x) \frac{\partial}{\partial \theta_j} \theta^T x \\
&= \left(y(1 - g(\theta^T x)) - (1 - y) g(\theta^T x) \right) x_j \\
&= (y - h_\theta(x)) \, x_j
\end{aligned}$$

Stochastic Gradient ascent rule:

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

Part III. Generalized linear model

Introduction

In statistics, the **generalized linear model** (**GLM**) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

-By Wikipedia

Introduction

The GLM consists of three elements:

1. A probability distribution from the exponential family.
2. A linear predictor $\eta = \theta^T x$.
3. A link function g such that $E(Y) = \mu = g^{-1}(\eta)$

The exponential family

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

Here, η is called the **natural parameter** (also called the **canonical parameter**) of the distribution; $T(y)$ is the **sufficient statistic** (for the distributions we consider, it will often be the case that $T(y) = y$); and $a(\eta)$ is the **log partition function**. The quantity $e^{-a(\eta)}$ essentially plays the role of a normalization constant, that makes sure the distribution $p(y;\eta)$ sums/integrates over y to 1.

The exponential family

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$$\begin{aligned} P(y;\varphi) &= \varphi^y(1-\varphi)^{1-y} = \exp(\log\varphi^y(1-\varphi)^{1-y}) \\ &= \exp(y\log\varphi + (1-y)\log(1-\varphi)) \\ &= \exp\left(y\log\frac{\varphi}{1-\varphi} + \log(1-\varphi)\right) \end{aligned}$$

For Bernoulli distribution

$$b(y) = 1$$

$$T(y) = y$$

$$\eta = \log\frac{\varphi}{1-\varphi} \Rightarrow \varphi = \frac{1}{1+e^{-\eta}}$$

Sigmoid function

$$a(\eta) = -\log(1-\varphi) = 1 + e^{-\eta}$$

The logistic model is the pre probability estimation for Bernoulli distribution.

The exponential family

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$$\boxed{\text{set } \sigma^2 = 1}$$

$$\mathrm{N}(\mu,1) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2 - \frac{1}{2}\mu^2 + \mu y\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^2)\exp(\mu y - \frac{1}{2}\mu^2)$$

For Gaussian distribution

$$b(y) = \frac{1}{\sqrt{2\pi}}\, exp(-\frac{1}{2}\, y^2)$$

$$T(y) = y$$

$$\eta = \mu$$

$$a(\eta) = \frac{1}{2}\,\mu^2$$

Three assumptions

1. $y \mid x; \theta \sim Exponential\ Family(\eta)$. I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.

2. Given x, our goal is to predict the expected value of $T(y)$ given x. In most of our examples, we will have $T(y) = y$, so this means we would like the prediction $h(x)$ output by our learned hypothesis h to satisfy $h(x) = E[y|x]$.

3. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$.

(1) $y|x;\theta\ ExponentialFamily$（$\eta$）；给定样本$x$与参数$\theta$，样本分类$y$ 服从指数分布族中的某个分布；
(2) 给定一个 x，我们需要的目标函数为$h_\theta(x) = E[T(y)|x]$;
(3)$\eta = \theta^T x$。

Logistic model

From Bernoulli distribution to logistic regression model :

$$
\begin{aligned}
h_\theta(x) &= E[T(y)|x] = E[y|x] = p(y=1|x;\theta) \\
&= \varphi \\
&= \frac{1}{1+e^{-\eta}} \\
&= \frac{1}{1+e^{-\theta^T x}}
\end{aligned}
$$

The first equality follows from Assumption 2, above; the second equality follows from the fact that $y|x;\ \theta \sim Bernoulli(\varphi)$; the third equality follows from Bernoulli distribution is an exponential family distribution; the fourth equality follows from Assumption 3, above.

Linear model

From Gaussian distribution to linear model:

$$\begin{aligned} h_\theta(x) &= E(T(y)|x) = E[y|x] \\ &= \mu \\ &= \eta \\ &= \theta^T x \end{aligned}$$

The first equality follows from Assumption 2, above; the second equality follows from the fact that $y|x;\ \theta\ \sim\ N(\mu, \sigma 2)$, and so its expected value is given by μ; the third equality follows from Assumption 1; and the last equality follows from Assumption 3.

Summary

- η以不同的映射函数与其它概率分布函数中的参数发生联系，从而得到不同的模型。

- GLM将指数分布族中的所有成员都作为linear model的扩展，通过各种非线性的映射函数$g^{-1}(\eta) = E[T(y);\eta]$将线性函数映射到其他空间，从而大大扩大了线性模型可解决的问题。

- 广义线性模型通过假设一个概率分布，得到不同的模型，而梯度下降是为了求取模型中的线性部分($\theta^T x$)的参数θ的。

Thanks