机器学习（本科生公选课）GEC6531

第3节 逻辑回归 Logistic 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

课堂名称：GEC6531 机器学习（公选课）

课堂编号：OA628

1、扫码关注公众号：微助教服务号。

2、点击系统通知：“点击此处加入【GEC6531 机器学习（公选课）】课堂”，填写学生资料加入课堂。

*如未成功收到系统通知，请点击公众号下方“学生” - “全部(A)” - “加入课堂” --- “输入课堂编号”手动加入课堂

二维码有效期至：2024-11-16

2. 微信扫码签到

线性规划中用最大似然估计（MLE）求参的矩阵理解

今天的目录

- **逻辑回归**
  - 分类问题
  - 逻辑回归的表示
  - 决策边界
- **逻辑回归求解**
  - 损失函数
  - 用最大似然估计（MLE）求参
  - 用最大后验估计（MAP）求参
  - 正则化
  - 梯度下降

今天的目录

■ 逻辑回归

- 分类问题
- 逻辑回归的表示
- 决策边界

■ 逻辑回归求解

- 损失函数
- 用最大似然估计（MLE）求参
- 用最大后验估计（MAP）求参
- 正则化
- 梯度下降

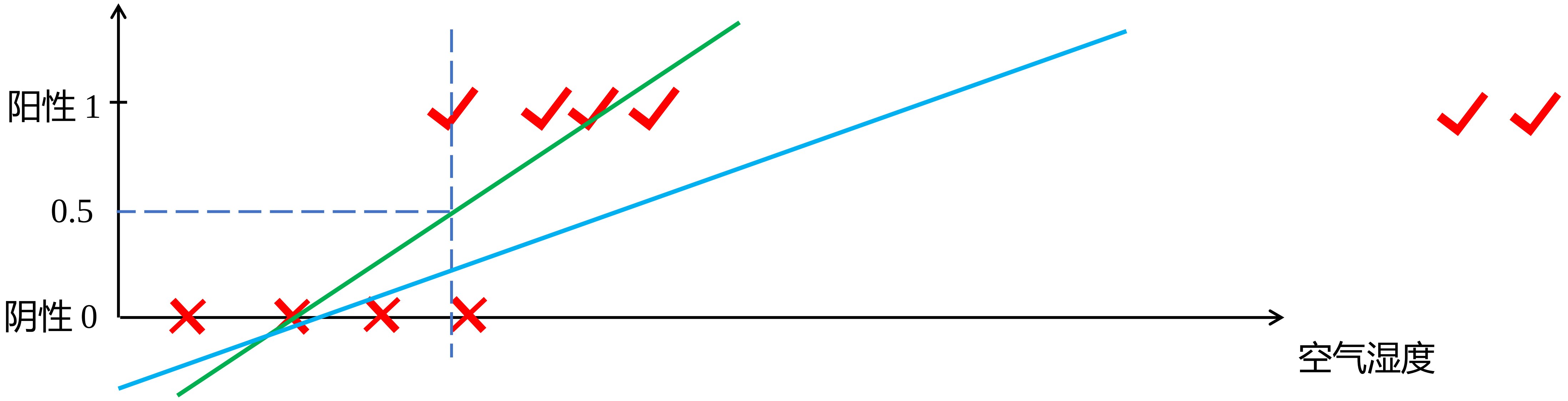
分类问题

- 是否垃圾邮件
- 是否诈骗电话
- 是否下雨

$$y \in \{0, 1\}$$

0: “阴性” (例如：不是垃圾邮件、不是诈骗电话、不下雨等)
1: “阳性” (例如：是垃圾邮件、是诈骗电话、下雨等)

用线性回归解决分类问题



■ **通过阈值来分类**

- 如果 h(x) >= 0.5, 预测 y=1
- 如果 h(x) < 0.5, 预测 y=0

线性回归分类的值域问题

■ 分类

- y = 0 或者 1
- 但是 h (x) 可能 >1 或者 <0， 希望 0 <= h(x) <= 1

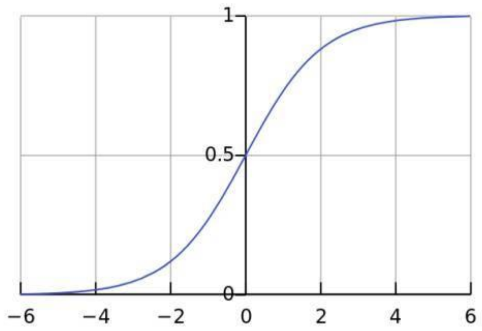
■ 逻辑回归

- Logistic regression: 0 <= h(x) <= 1

名字叫逻辑回归是历史原因，但实际上是分类方法

这里“逻辑”是音译，来自逻辑函数： logistic function
也叫sigmoid function

$$g(s) = \frac{1}{1+\exp(-s)}$$



今天的目录

■ 逻辑回归

- 分类问题
- 逻辑回归的表示
- 决策边界

■ 逻辑回归求解

- 损失函数
- 用最大似然估计（MLE）求参
- 用最大后验估计（MAP）求参
- 正则化
- 梯度下降

逻辑回归

把逻辑函数里的自变量替换成 $\mathbf{w}^{\mathrm{T}}\mathbf{x}$

阳性概率

$$g(s) = \frac{1}{1+\exp(-s)} \quad \Rightarrow \quad \mathrm{p}(\mathrm{y}=1|\mathrm{x},\mathrm{w}) = \frac{1}{1+\mathrm{e}^{-\mathrm{w}^{\mathrm{T}}\mathrm{x}}}$$

- $h(\mathbf{x}) = g(\mathbf{w}^{\mathrm{T}}\mathbf{x})$

分类判断

- 如果 P (y=1| **x, w**) >= 0.5 则阳性, 即类 1
- 否则阴性，即类 0

阴性概率

- P (y=0| **x, w**) = 1 - P (y=1| **x, w**) ; $\quad \mathrm{p}(\mathrm{y}=0|\mathrm{x},\mathrm{w}) = \dfrac{1}{1+\mathrm{e}^{\mathrm{w}^{\mathrm{T}}\mathrm{x}}}$
- $P(y|X) = \dfrac{1}{1+e^{-y(w^TX)}}$

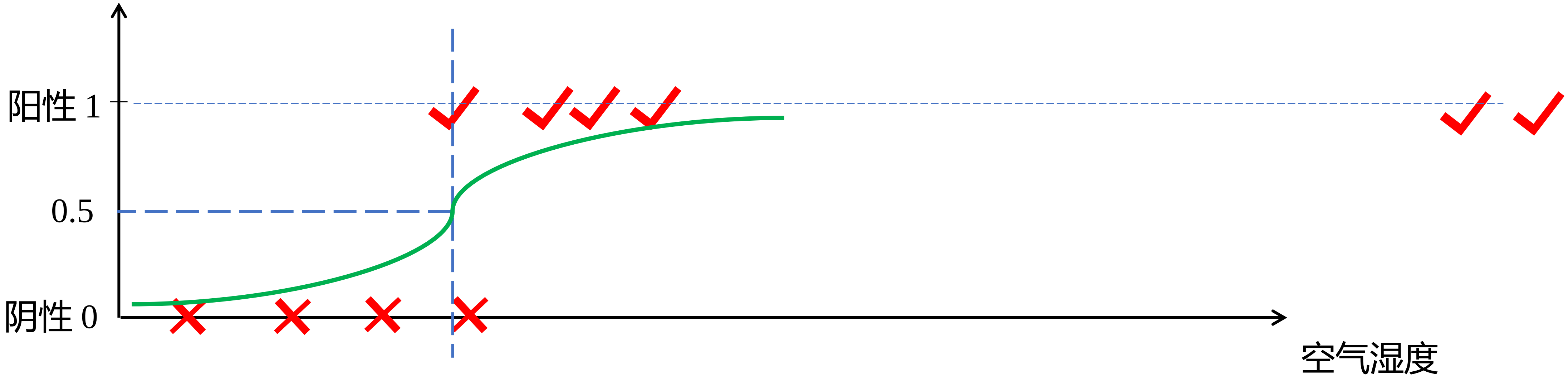
逻辑回归的表示

■ **把逻辑函数里的自变量替换成 $\mathbf{w}^T\mathbf{x}$**

阳性概率

$$g(s) = \frac{1}{1+\exp(-s)} \Rightarrow p(y=1|x,w) = \frac{1}{1+e^{-w^T x}}$$

- $h(\mathbf{x}) = g(\mathbf{w}^T\mathbf{x})$

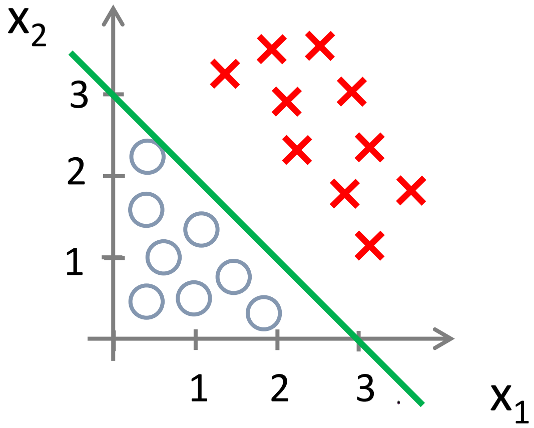


今天的目录

- **逻辑回归**
  - 分类问题
  - 逻辑回归的表示
  - 决策边界
- **逻辑回归求解**
  - 损失函数
  - 用最大似然估计（MLE）求参
  - 用最大后验估计（MAP）求参
  - 正则化
  - 梯度下降

决策边界 (Decision Boundary)

$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$



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

Predict "$y = 1$" if $\underbrace{-3 + x_1 + x_2 \geq 0}_{\theta^T x}$

$\rightarrow h_\theta(x) = 0.5$

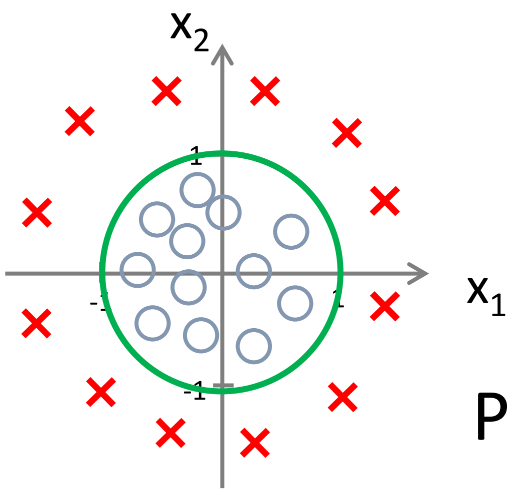
$\boxed{x_1 + x_2 = 3}$

$\rightarrow x_1 + x_2 \geq 3$

$x_1 + x_2 < 3$

$\rightarrow y = 0$

非线性决策边界 (Decision Boundary)

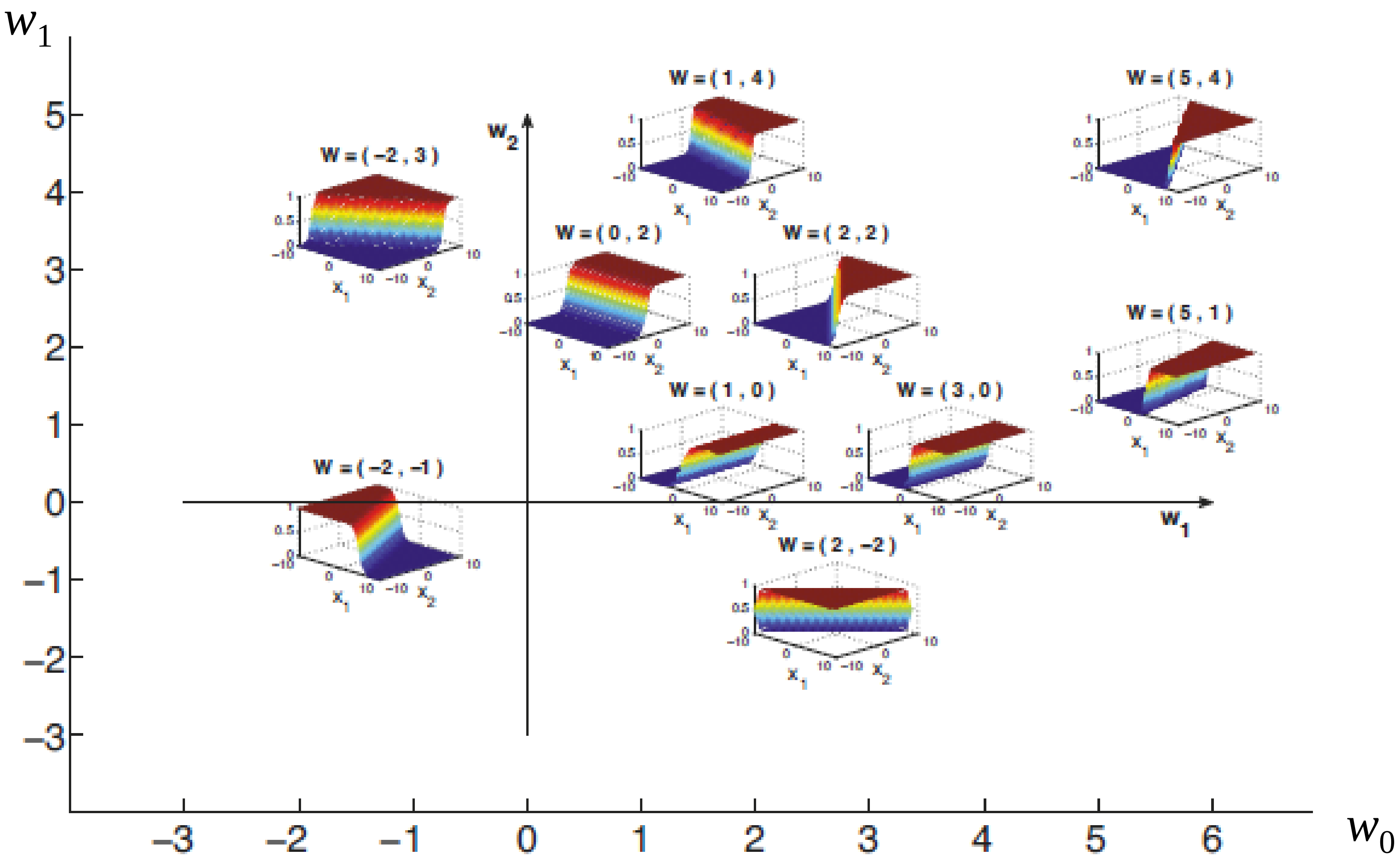


$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

决策边界直观理解 (Decision Boundary)



今天的目录

- **逻辑回归**
  - 分类问题
  - 逻辑回归的表示
  - 决策边界
- **逻辑回归求解**
  - 损失函数
  - 用最大似然估计（MLE）求参
  - 用最大后验估计（MAP）求参
  - 正则化
  - 梯度下降

损失函数 (Loss function / Cost function)

预测概率：

$$P(y=1|x) = h_\theta(x) = \frac{1}{1+e^{(-\theta^T x)}} = \sigma(\theta^T x) \tag{1}$$

$$P(y=0|x) = 1 - P(y=1|x) = 1 - h_\theta(x) \tag{2}$$

结合 (1) 和 (2):

$$P(y|x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y} \tag{3}$$

y 表示标签，取值为 0 或 1

对给定的 m 个样本，使用最大似然估计 (MLE):

$$L(\theta) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)};\theta) = \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}} \tag{6}$$

$$\longrightarrow \ell(\theta) = \log L(\theta) = \sum_{i=1}^{m} (y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(y^{(i)}))) \tag{7}$$

$$J(\theta) = -\frac{1}{m}\ell(\theta)$$

$J(\theta)$ 是设计的损失函数

逻辑回归的MLE

- 在 MLE 中，我们选择**最大化条件似然**的参数。条件似然 $P(Y|X, w)$ 是以训练数据中特征向量 x_i 为条件的观测值 $Y \in R^n$ 中的概率。其中，$X = [x_1, \ldots, x_i, \ldots, x_n] \in R^{d \times n}$。
- 我们选择使该函数最大化的参数，并假设对于给定的输入特征 x_i 和 w 时 y_i 之间是独立的。

$$P(\mathbf{y} \mid X, \mathbf{w}) = \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \mathbf{w}).$$

我们对上式取 log：

$$\log\left(\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})\right) = -\sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

$$\hat{\mathbf{w}}_{MLE} = \underset{\mathbf{w}}{\operatorname{argmax}} -\sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

逻辑回归的MLE

$$\hat{\mathbf{w}}_{MLE} = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

- 我们需要估计参数 w。为了求出最小值对应的参数，可以试着求出 $\nabla_w \sum_i^n log(1 + e^{y_i w^T x_i}) = 0$。
- 该方程没有闭式解。我们在负对数似然上使用梯度下降来寻找近似解:

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

逻辑回归的MAP

- 在 MAP 估计中，我们将 w 作为一个随机变量，并可以指定它的先验分布。假设先验为： $w \sim N(0\ \sigma^2 I)$。这是逻辑回归的高斯近似。
- 我们在 MAP 中的目标是找到对给定数据的**最大化后验**的模型参数。

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)} \propto P(D|w)P(w)$$

$$\begin{aligned} P(w|D) &= P(w|X,y) \propto P(X,y|w)P(w) \\ &= P(y|X,w)P(X,w)P(w) \propto P(y|X,w)P(w) \end{aligned}$$

$$\begin{aligned} \hat{w}_{MAP} &= \underset{w}{\operatorname{argmax}}\ log(P(y|X,w)P(w)) \\ &= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} log(1 + e^{y_i w^T x_i}) + \lambda w^T w \end{aligned}$$

$\lambda = \frac{1}{2\sigma^2}$。同样，这个函数没有闭式解，但我们可以在负对数后验上使用梯度下降来找到最优的参数。

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}) + \lambda \mathbf{w}^\top \mathbf{w}$$

今天的目录

■ 逻辑回归

- 分类问题
- 逻辑回归的表示
- 决策边界

■ 逻辑回归求解

- 损失函数
- 用最大似然估计（MLE）求参
- 用最大后验估计（MAP）求参
- 正则化
- 梯度下降

正则化 (Regularization)

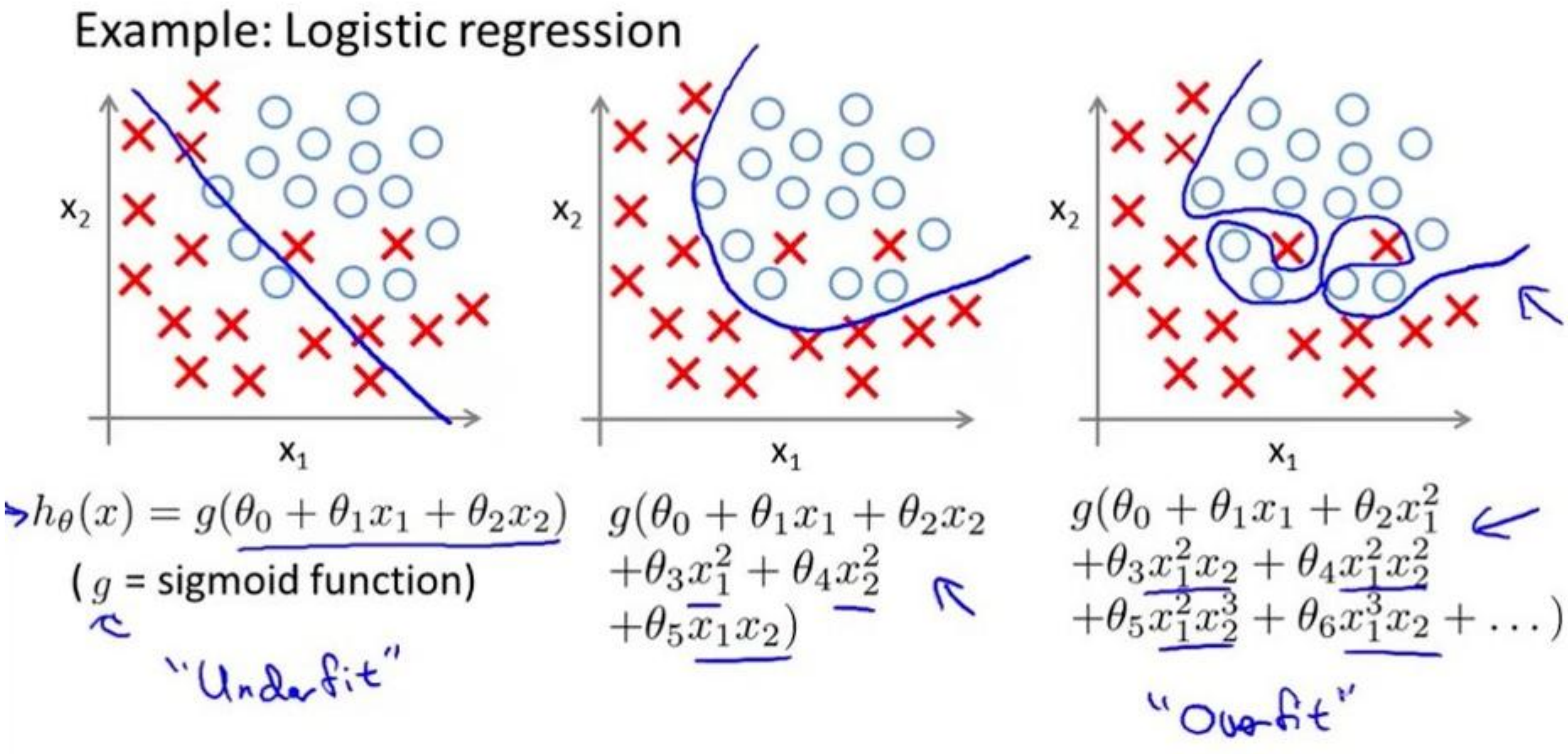
基本思想

- L_2 正则:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}))^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] \quad (8)$$

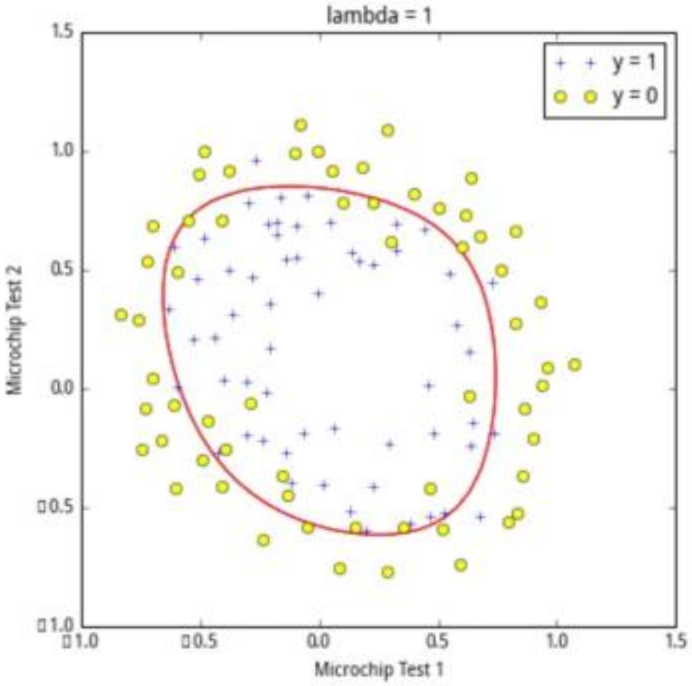
- 正则化: 防止权重变得过大
- 正则化可以通过对权值的约束在一定程度上避免过拟合

正则化 (Regularization)

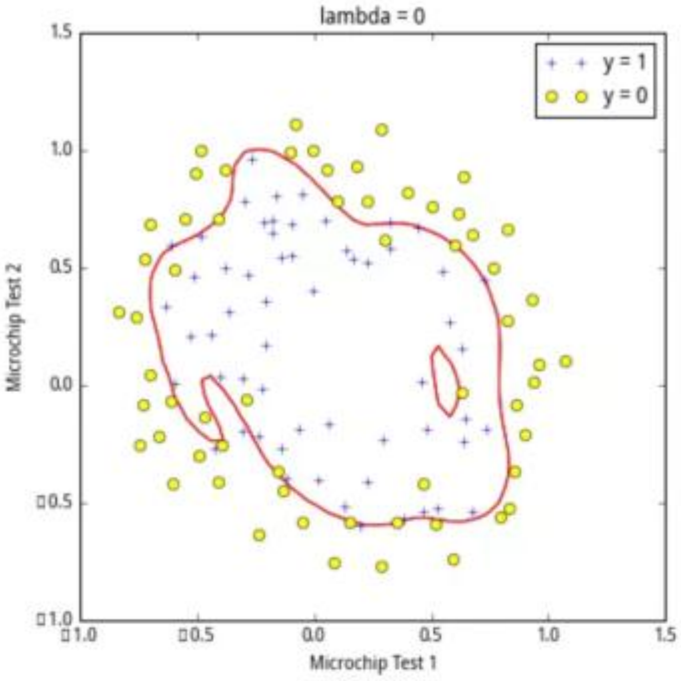


正则化 (Regularization)

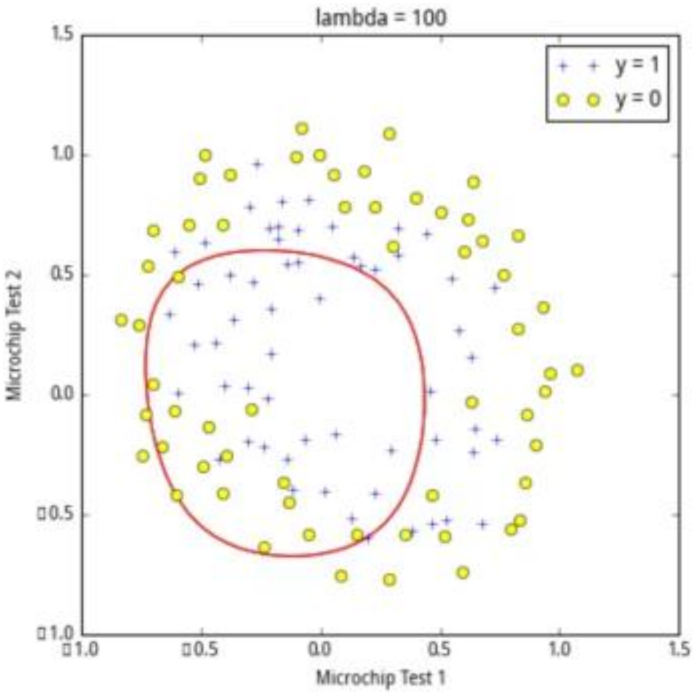
通过逻辑回归对任务进行分类



λ = 1 Good !



λ = 0 Overfiting



λ = 100 Underfiting

不同 λ 值的二分类效果

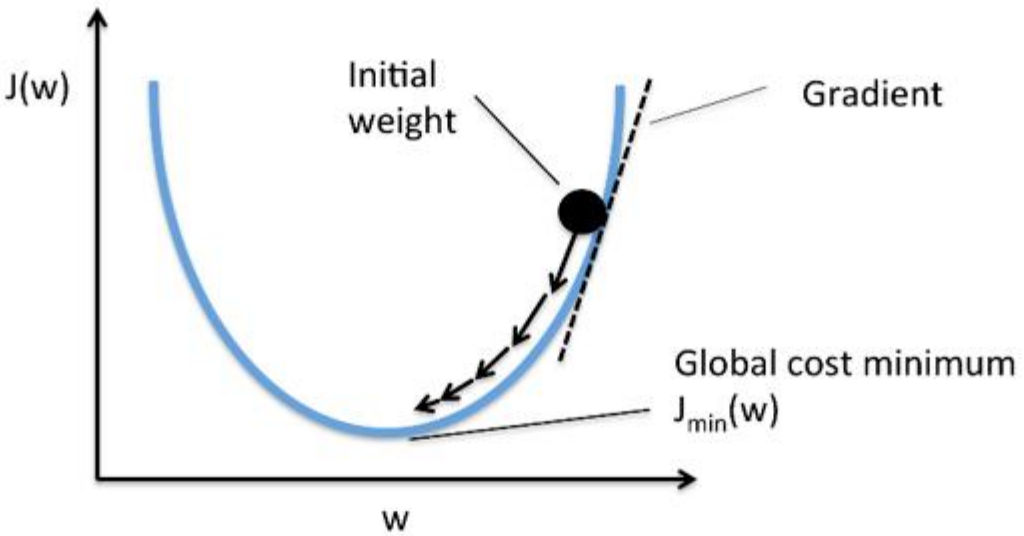
今天的目录

- **逻辑回归**
  - 分类问题
  - 逻辑回归的表示
  - 决策边界
- **逻辑回归求解**
  - 损失函数
  - 用最大似然估计（MLE）求参
  - 用最大后验估计（MAP）求参
  - 正则化
  - 梯度下降

梯度下降预览 (Gradient Descent)

损失函数 $J(\theta_0, \theta_1)$

目标 $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$



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

逻辑回归梯度下降 (Gradient Descent)

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

$$\downarrow$$

偏导数:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_i x_j^{(i)}(h_\theta(x^{(i)}) - y^{(i)})$$

$$\downarrow$$

权重更新:

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

总结

■ 逻辑回归，一种解决分类问题的常用方法，将线性回归嵌入逻辑函数

$$g(s) = \frac{1}{1+\exp(-s)} \quad \Rightarrow \quad p(y=1|x,w) = \frac{1}{1+e^{-w^T x}}$$

- $h(\mathbf{x}) = g(\mathbf{w}^{\mathrm{T}}\mathbf{x})$

■ 逻辑回归求解

- y 的概率的统一表示方法 $P(y|x;\theta) = (h_\theta(x))^y(1-h_\theta(x))^{1-y}$ y 表示标签，取值为 0 或 1
- 损失函数

$$L(\theta) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)};\theta) = \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}} \tag{6}$$

$$\longrightarrow \ell(\theta) = \log L(\theta) = \sum_{i=1}^{m}(y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(y^{(i)}))) \tag{7}$$

$$J(\theta) = -\frac{1}{m}\ell(\theta)$$

$J(\theta)$ 是设计的损失函数

- 用梯度下降求解