机器学习
Machine Learning

北京航空航天大学计算机学院
School of Computer Science and Engineering, Beihang University
刘庆杰　陈佳鑫

2025年春季学期
Spring 2025

部分内容来源于C. Bishop和A. NG等人的课程以及互联网资源

课前回顾

内容提要

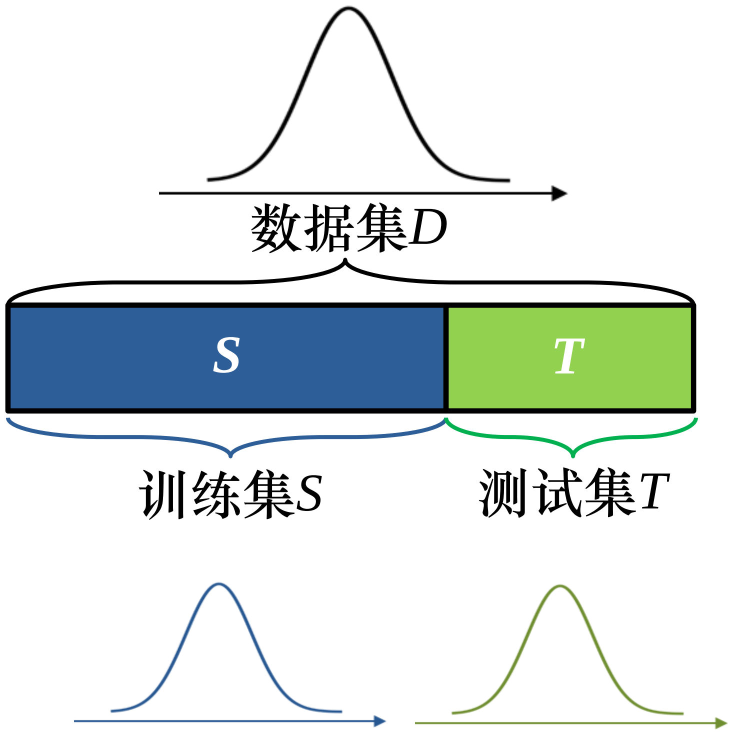
- 2.1 数据集的划分方法
- 2.2 模型的性能度量
- 2.3 概率统计基础
- 2.4 贝叶斯决策理论
- 2.5 参数化概率密度估计方法
- 2.6 非参数概率密度估计方法
- 2.7* 矩阵理论基础（拓展）

数据集划分的基本要求

● 模型评估

➢ 1. 数据集划分：分为训练集和测试集两部分

- 原则：测试集应尽可能与训练集互斥，测试样本不在训练集中出现
- 目标：将数据集D划分为训练集S和测试集T两部分，在训练集上建立模型，在测试集上评估性能
- 假设：测试样本从原样本真实分布中独立同分布采样得到
- 方法：留出法、自助法、交叉验证法



➢ 2. 性能度量：模型在测试集（新样本）上进行度量，也叫泛化性能

模型的性能度量

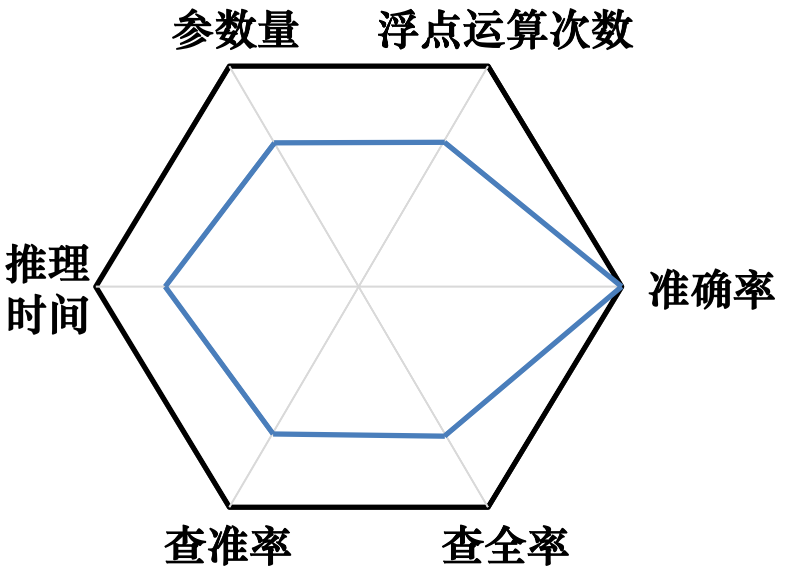
● 模型评估

- ➢ 1. 数据集划分：分为训练集和测试集两部分
- ➢ 2. **性能度量**：性能度量：模型在测试集（新样本）上进行度量，也叫泛化性能
  - ■ 常用性能度量：错误率和精度（分类任务）
  - ■ 仅能评估是否正确分类，无法提供更全面的评估

示例1：发动机合格检测

精度可以评估“检测出有多少发动机是合格的”

无法评估“检测出的合格发动机有多少是真正合格的”



一些基本概念

- 概率 (Probability)
  - 对随机事件发生可能性大小的度量
- 联合概率 (Joint Probability)
  - A和B共同发生的概率，称事件A和B的联合概率，记作P(A, B)或P(A∩B)
- 条件概率 (Conditional Probability)
  - 事件B已发生的条件下，事件A发生的概率，记作P(A|B)

贝叶斯公式

- **贝叶斯公式 (Bayes' Theorem)**
  - 贝叶斯公式给出了“结果”事件A已经发生的条件下，“原因”事件B的条件概率，对结果的任何观测都将增加我们对原因事件B的真正分布的知识

后验概率：给定观测数据后，某事件发生的概率

先验概率：在没有观测数据之前，某事件发生的初始概率

似然概率：给定事件发生的情况下，观测数据出现的概率

$$\boxed{P(B_i|A)} = \frac{\boxed{P(B_i)}\boxed{P(A|B_i)}}{\boxed{\sum_{j=1}^{n} P(B_j)P(A|B_j)}}$$

证据因子：观测数据的边际概率，即所有可能事件下观测数据的总概率

贝叶斯决策

- 贝叶斯决策是统计决策理论中的一个基本方法，用于解决分类问题
  - 已知条件：
    - 1、属于一定数量类别的数据，类别为标签为： ω_i , $i = 1, 2, \ldots, c$
    - 2、各类别 ω_i的类先验概率$P(\omega_i)$和类条件概率密度$P(x|\omega_i)$
  - 基本思想：根据贝叶斯公式计算后验概率，基于最大后验概率进行判决
  - 判决函数：最大化后验概率

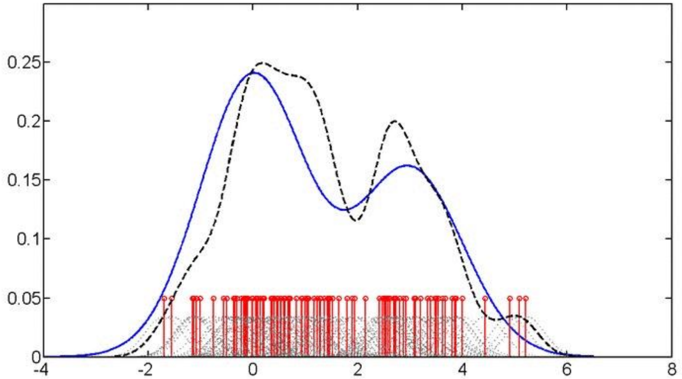
$$x \in \omega_k \text{ 当且仅当} k = \arg\max_i \{P(\omega_i|x)\}\text{，其中} P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} P(x|\omega_j)P(\omega_j)}$$

2.5 参数化概率密度估计方法

- 概率密度估计的概念
- 参数化概率密度估计方法的概念
- 极大似然估计方法
- 贝叶斯估计方法
- 估计量的性质与评价标准

概率密度估计

- **概率密度估计的任务：**
  - 根据观测样本数据估计类条件概率密度函数 $P(x|\omega_i)$和类先验概率$P(\omega_i)$
- **为什么需要估计概率密度**
  - 概率密度估计可以建模原始数据分布，辅助精细地了解数据特性。进而帮助识别数据中的异常值、或者生成新数据



- **概率密度估计的方法：**
  - 参数化方法：已知概率密度函数的形式，其中几个参数未知
  - 非参数化方法：概率密度函数的形式未知

https://en.wikipedia.org/wiki/Density_estimation

参数化概率密度估计

- 参数化概率密度估计的任务
  - 已知：概率密度函数的形式，若干参数未知
  - 目标：依据样本估计未知参数的值
- 典型方法
  - 极大似然估计：把待估计参数看做是确定的量，只是其取值未知。最佳估计就是使产生已观测到样本的概率最大的那个值
  - 贝叶斯估计：把待估计参数看做是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，从而利用样本信息修正参数的初始估计值的过程

极大似然估计

- **极大似然估计的假设条件**
  - $P(x|\omega_i)$具有某种确定的解析函数形式，只有部分参数θ未知；
  - 参数θ通常为向量，如一维正态分布$N(\mu,\sigma^2)$中的μ、σ
  - 参数θ是确定的未知量，不是随机量
  - 各类样本集x_i，$i=1,2,\cdots,c$ 满足独立同分布条件(i.i.d.)，即x_i均为从概率密度函数为$P(x|\omega_i)$的总体分布中独立抽取出来的
  - 各类样本只包含本类分布的信息；因此，$P(x|\omega_i)$可记为$P(x|\omega_i;\theta_i)$或$P(x;\theta_i)$
- **基于上述假设，各类条件概率密度可根据各类样本分别估计**

极大似然估计

- **似然函数**
  - 针对一类已知样本$X = \{x_i, i = 1,2,\ldots,N\}$，定义参数$\theta$下观测到样本集$X$的联合分布概率密度，称为相对于样本集$X$的$\theta$的似然函数

$$l(\theta) = P(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} P(x_i; \theta)$$

- **基本思想**
  - 在θ可能的取值范围内选择使似然函数达到最大的参数值作为参数θ的估计值
  - 形式化描述为：求$\hat{\theta}$，使得$l(\hat{\theta}) = \max_{\theta} l(\theta)$
  - 如果参数$\theta = \hat{\theta}$时，$l(\theta)$最大，则$\hat{\theta}$是最可能的参数估计值。它是关于样本集的函数，记作：$\hat{\theta} = d(x_1, x_2, \cdots, x_N) = d(X)$，称为极大似然估计量
  - 为便于分析求解，实际应用中往往采用对数似然函数：$H(\theta) = \ln l(\theta)$

极大似然估计

求解

- 若似然函数连续可微，最大似然函数估计量就是方程$\frac{dl(\theta)}{d\theta}=0$或$\frac{dH(\theta)}{d\theta}=0$的解
- 若未知参数不止一个，即$\theta=[\theta_1,\theta_2,\cdots,\theta_s]^T$，则需联立以下$s$个方程组求解

$$\frac{dH(\theta)}{d\theta_i}=0, i=1,2,\ldots,s$$

极大似然函数的求解性质

- 若似然函数连续可导，存在最大值且必要条件方程有唯一解，则该解就是极大似然估计量
- 如果似然函数有多个解，则使似然函数值最大者为极大似然估计量
- 若似然函数单调，可根据极大似然思想，将似然函数最大值点作为参数的极大似然估计值

极大似然估计

- **极大似然估计的求解步骤**

  - 1、确定需要估计的概率分布$P(x|\theta)$，其中$\theta=[\theta_1,\theta_2,\dots,\theta_s]^T$
  - 2、构造似然函数$l(\theta)=P(x_1,x_2,\dots,x_N;\theta)=\prod_{i=1}^{N}P(x_i;\theta)$
  - 3、求对数似然函数$H(\theta)=\ln l(\theta)$
  - 4、令$\frac{dH(\theta)}{d\theta_i}=0, i=1,2,\dots,s$，联立求解

极大似然估计

- 示例：单变量正态分布
  - 已知
    - 参数：$\boldsymbol{\theta}=[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2]$，　$\boldsymbol{\theta}_1=\boldsymbol{\mu},\boldsymbol{\theta}_2=\boldsymbol{\sigma}^2$
    - 概率密度函数：$\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma}}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$
    - 样本集：　$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N\}$
  - 目标：估计参数$[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2]$

极大似然估计示例：单变量正态分布

- 解：
  - 1、求似然函数： $l(\theta)=P(x_1,x_2,\dots,x_n;\theta)=\prod_{i=1}^{n}P(x_i;\theta)$
  - 2、求对数似然函数：$H(\theta)=\ln l(\theta)=\sum_{i=1}^{N}\ln P(x_i;\theta)$
  - 3、构造方程组：$\begin{cases}\frac{\partial H}{\partial \mu}=0\\ \frac{\partial H}{\partial \sigma^2}=0\end{cases}$，$\begin{cases}\frac{1}{\sigma^2}\left[\sum_{i=1}^{N}x_i-N\mu\right]=0\\ -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2=0\end{cases}$
  - 4、联立求解：$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$，$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i$

贝叶斯估计

贝叶斯估计的基本思想

- 把待估计参数 θ 看作是具有先验分布$P(\theta)$的随机变量，其取值与样本集 X 有关，贝叶斯估计利用样本集 X 将 θ 的先验概率分布修正为后验概率分布
- 贝叶斯决策用于分类，计算离散形式的后验概率值，而贝叶斯估计则用于回归，计算连续形式的后验概率密度函数

贝叶斯估计损失函数

- 把 θ 估计为 $\hat{\theta}$ 所造成的损失，记为$\lambda(\hat{\theta},\theta)$
- 不同于离散形式贝叶斯决策的损失表，由于参数化概率密度估计为连续值估计，因此常采用损失函数，常用平方误差损失函数$\lambda(\hat{\theta},\theta)=(\theta-\hat{\theta})^2$

贝叶斯估计

● 贝叶斯估计相关概念

- **条件期望损失：** $R(\hat{\theta}|x)=\int_{\Theta}\lambda(\hat{\theta},\theta)P(\theta|x)d\theta$，其中，$x\in E^d$，$\theta\in\Theta$，$E^d$为样本集，$\Theta$为待估参数集

- **期望风险：** $R=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)P(x;\theta)d\theta dx$

$$=\int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)P(\theta|x)P(x)d\theta dx=\int_{E^d}R(\hat{\theta}|x)P(x)dx$$

- **贝叶斯估计量：** 使条件期望损失最小的估计量$\hat{\theta}$，即

$$\hat{\theta}=\operatorname{argmin}\left(R(\hat{\theta}|x)\right)=\operatorname{argmin}\left(\int_{\Theta}\lambda(\hat{\theta},\theta)P(\theta|x)d\theta\right)$$

- **定理2.5.1：** 若采用平方误差损失函数，则θ的贝叶斯估计量是在给定样本集X时θ的条件期望，即$\hat{\theta}=E(\theta|x)=\int_{\Theta}\theta P(\theta|x)d\theta$

定理2.5.1证明过程，可参考《模式识别》-边肇祺张学工P52

贝叶斯估计

贝叶斯估计步骤

- 1、确定参数 θ 所遵从的先验分布： $P(\theta)$
- 2、求样本集的联合分布：$P(X|\theta)=\prod_{i=1}^{N}P(x_i|\theta)$
- 3、求 θ 的后验概率分布：$P(\theta|X)=\frac{P(X|\theta)P(\theta)}{\int_\Theta P(X|\theta)P(\theta)d\theta}$
- 4、求 θ 的贝叶斯估计量（定理2.5.1）： $\hat{\theta}=\int_\Theta \theta P(\theta|x)d\theta$

贝叶斯估计

- **示例：单变量正态分布**
  - **已知**
    - 参数：$\theta = [\theta_1, \theta_2]$， $\theta_1 = \mu, \theta_2 = \sigma^2$ ，其中θ_2已知
    - 概率密度函数：$P(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$
    - 样本集：$X = \{x_1, x_2, \dots, x_N\}$
  - **目标：估计均值θ_1**

贝叶斯估计示例：单变量正态分布

- 解：（仅保留关键步骤）
  - 1、确定概率密度函数形式：$P(x|\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$
  - 2、设估计量μ遵从先验分布：$\mu \sim \mathcal{N}(\mu_0,\sigma_0^2)$
  - 3、根据观测样本x求得μ的后验分布：$\mu|x \sim \mathcal{N}(\frac{\mu_0\tau_0+n\bar{x}\tau}{\tau_0+n\tau},\frac{1}{\tau_0+n\tau})$，其中$\tau=\frac{1}{\sigma^2}$，$\tau_0=\frac{1}{\sigma_0^2}$
  - 4、μ的贝叶斯估计量为

$$E(\mu|x)=\frac{\mu_0\tau_0+n\bar{x}\tau}{\tau_0+n\tau}=w\mu_0+(1-w)\bar{x}$$，其中n为样本数，$w=\frac{\tau_0}{\tau_0+n\tau}$

  - 5、当$n=0$时，估计量$\hat{\mu}=\mu_0$，当$n\to\infty$时，估计量$\hat{\mu}=\bar{x}$

具体可参考《模式识别》-边肇祺张学工P56-57

估计量的性质与评价标准

● 估计量的性质

无偏性	渐进无偏性	有效性	一致性
$\boldsymbol{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N)\right]=\boldsymbol{\theta}$	$\boldsymbol{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{x}_N)\right]\xrightarrow{N\to\infty}\boldsymbol{\theta}$	对估计$\widehat{\boldsymbol{\theta}}_1$，$\widehat{\boldsymbol{\theta}}_2$，若方差$\boldsymbol{\sigma}^2\left(\widehat{\boldsymbol{\theta}}_1\right)<\boldsymbol{\sigma}^2\left(\widehat{\boldsymbol{\theta}}_2\right)$，则$\widehat{\boldsymbol{\theta}}_1$估计更有效	当样本数无穷多时，每一次估计都在概率意义上任意接近真实值，即：$\forall\varepsilon>0$， $\lim\limits_{N\to\infty}P\left(\left\|\widehat{\boldsymbol{\theta}}_N-\boldsymbol{\theta}\right\|>\varepsilon\right)=0$

- ➢ **结合无偏性和有效性，要求估计量能够在多次估计中，以较小的方差平均地表示真实值**
- ➢ **极大似然估计是无偏的，在样本充足时有效性更好，一致性更强；**
- ➢ **贝叶斯估计可能有偏，在样本量有限且有合理的先验信息时更有效。**

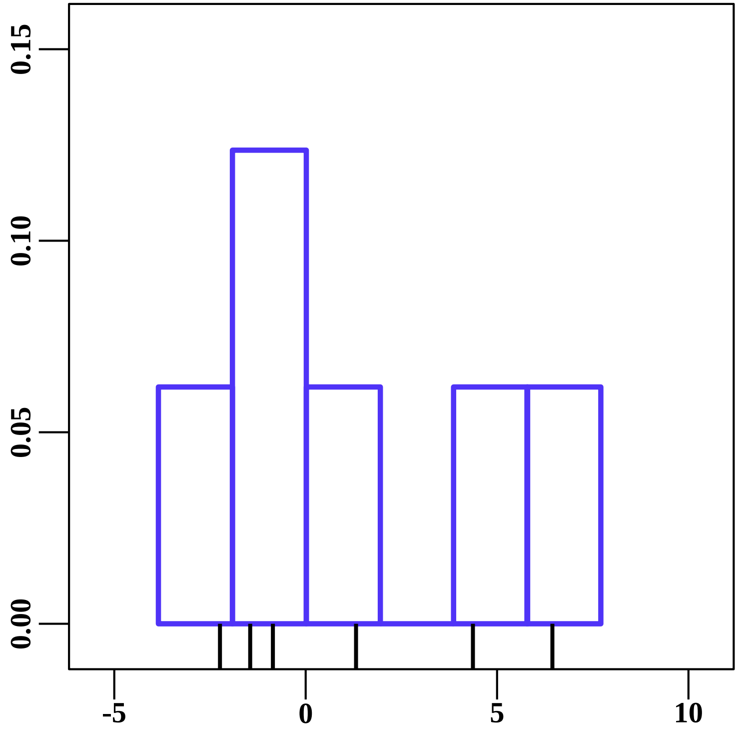
2.6 非参数概率密度估计方法

- 什么是非参数估计?
- Parzen窗算法
- k近邻算法

什么是非参数估计?

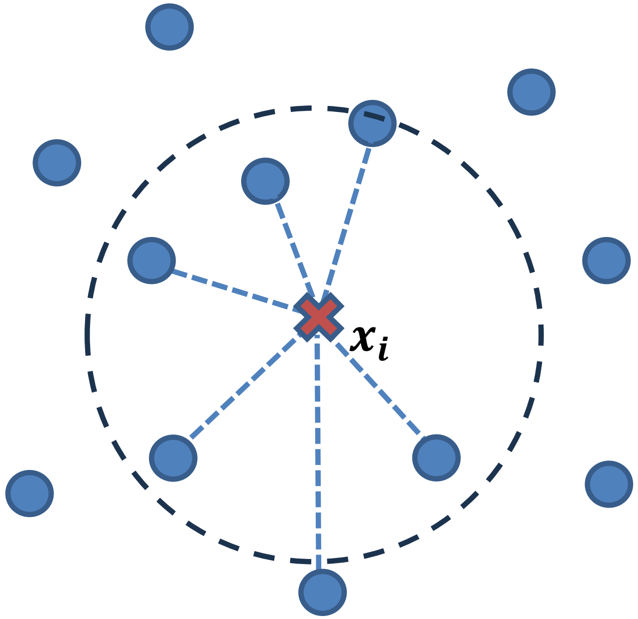
- **非参数估计**
  - 概率密度**函数的形式未知**的模型，直接依赖于数据本身来进行推断和估计

- **常见的非参数估计方法**
  - Parzen窗法
  - k近邻法



非参数估计方法

- 非参数估计方法
  - 基本思路：要估计x_i点的密度，可把相关样本在该点的“贡献”相加近似作为其概率密度，进而可以以此方法估计每个点的概率密度
  - 具体流程
    - 1.计算x_i点处与概率密度相关的贡献点
    - 2.确定贡献点对x_i点处的贡献
    - 3.重复步骤1-2计算所有点处的概率密度



非参数估计方法

非参数概率密度估计

- 假设N为样本总数，以x_i为中心的区域R（足够小，体积为V）内的k个点对估计x_i的概率密度$p(x)$有贡献，则R中落入k个样本的概率为：

$$P_R = k/N = \int_R p(x)dx = \hat{p}(x)V$$

- 估计得到的概率密度$\hat{p}(x_i)$为： $\hat{p}(x_i) = k/NV$
- 当满足以下条件时，概率估计$\hat{p}(x_i)$收敛于$p(x_i)$：
  - 贡献点的区域大小越小越好 $\lim\limits_{N\to\infty} V_N = 0$
  - 贡献点越多越好 $\lim\limits_{N\to\infty} k_N = \infty$
  - 贡献点与总样本数比例越小越好 $\lim\limits_{N\to\infty} \frac{k_N}{N} = 0$

非参数估计方法

- **Parzen法**
  - 使区域体积序列V_N以N的某个函数的关系不断缩小
  - 同时限制k_N和k_N/N

有限的N, V选择很敏感

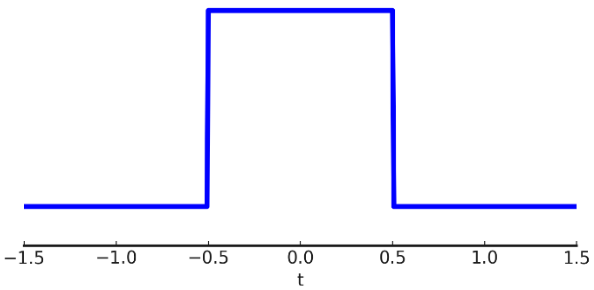
↓

动态变化V的取值

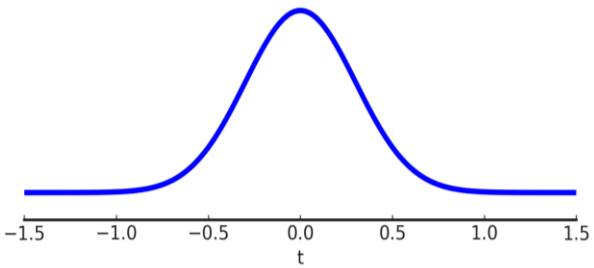
- **k近邻法**
  - 使落入区域样本数k_N为N的某个函数
  - 选择不同的V_N使区域包含x的k_N个近邻

非参数估计方法——Parzen窗法

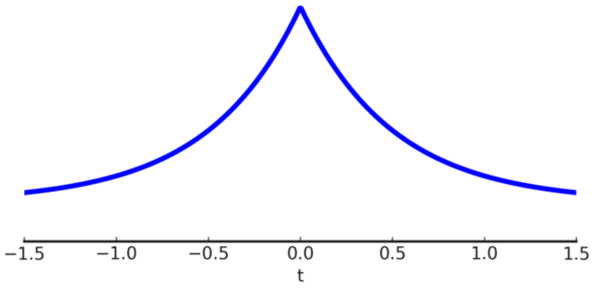
- **Parzen窗法使用窗函数对贡献点进行选择**
- **窗函数**
  - ➢ 形式：$k(x, x_i)$，反映x_i对$p(x)$的贡献同时进行区域选择
  - ➢ 条件：$k(x, x_i) \geq 0, \int k(x, x_i)dx = 1$
  - ➢ 窗函数选择
    - ■ 方窗函数、正态窗函数、指数窗函数等
  - ➢ 窗宽选择
    - ■ 原则：样本数多则选小窗宽；样本数少则选大窗宽



方窗函数



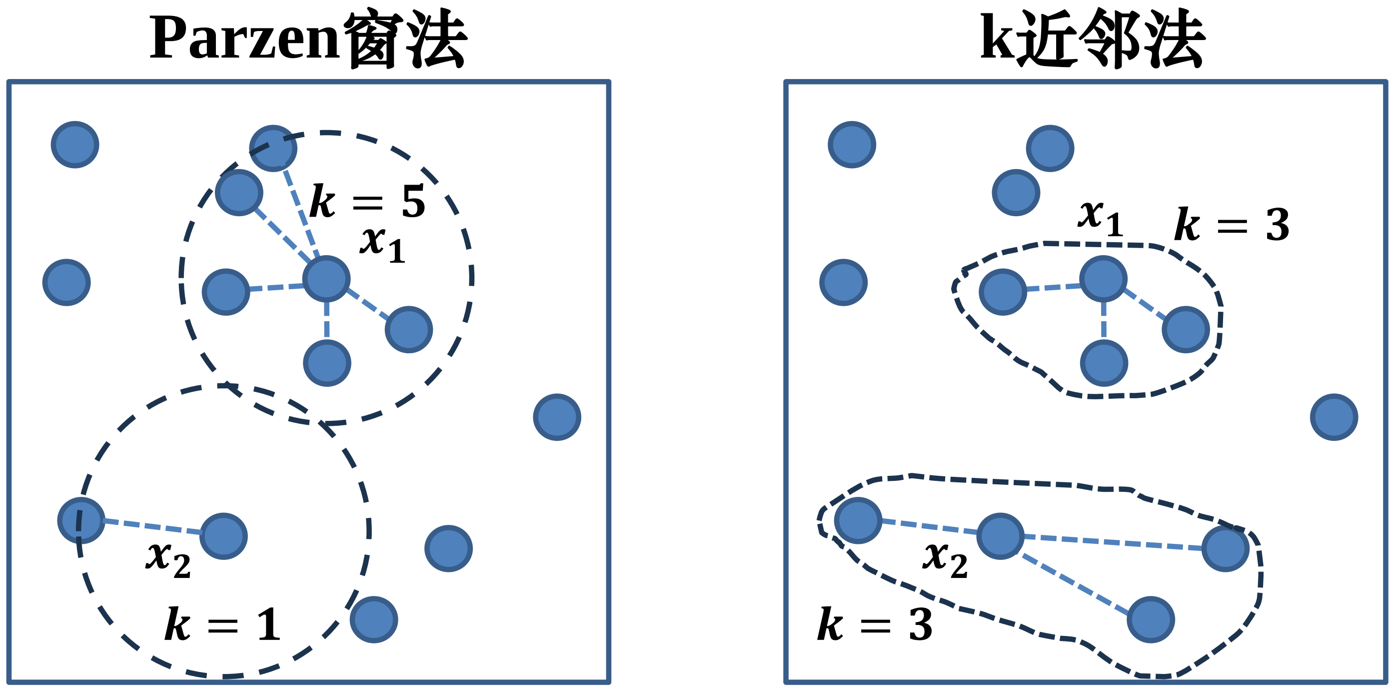
正态窗函数



指数窗函数

非参数估计方法——k近邻法

- k近邻法使用k近邻算法对贡献点进行选择
  - 选择样本x_i一定范围内确定个数的k个样本后根据k/NV计算概率密度
  - k近邻法更适用于样本分布不均匀的数据



k近邻算法

- **k近邻算法流程：**
  - 选择一个正整数k，表示需要考虑的邻近样本的数量
  - 对于待预测的样本，计算其与训练集中所有样本之间的距离
    - 常用的距离度量包括欧氏距离、曼哈顿距离、余弦相似度等
  - 选择k个最近邻居：根据计算得到的距离，选择距离待预测样本最近的k个邻居
- **k值选择：**
  - k值决定了决策的局部性，k值越大，模型越平滑，越小则越敏感

小结

- Parzen窗法与k近邻法均使用广泛
- 在样本分布不均匀时k近邻法比Parzen窗法表现更好
- 在高维空间中k近邻法比Parzen窗法更易应用，且可通过技术手段缓解维度问题
- k近邻法和Parzen窗法在边界附近都可能会遇到估计偏差

第3章：线性模型

Chapter 3: Linear Models

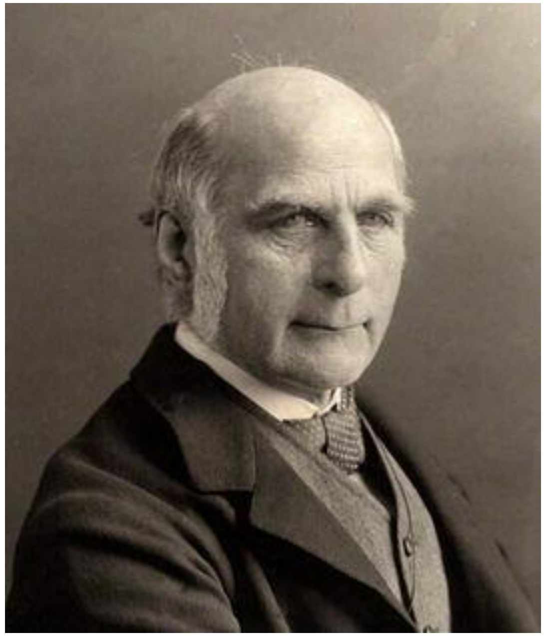
3.1 什么是线性回归

- “回归”的起源
- 线性回归的概念

“回归”起源

- 研究父母身高与子女身高之间的关系

【英国生物学家Francis Galton】
1822-1911



若父母身高高于平均身高，则其子女身高倾向于倒退生长，即会比其父母身高矮一些而更接近于大众平均身高。

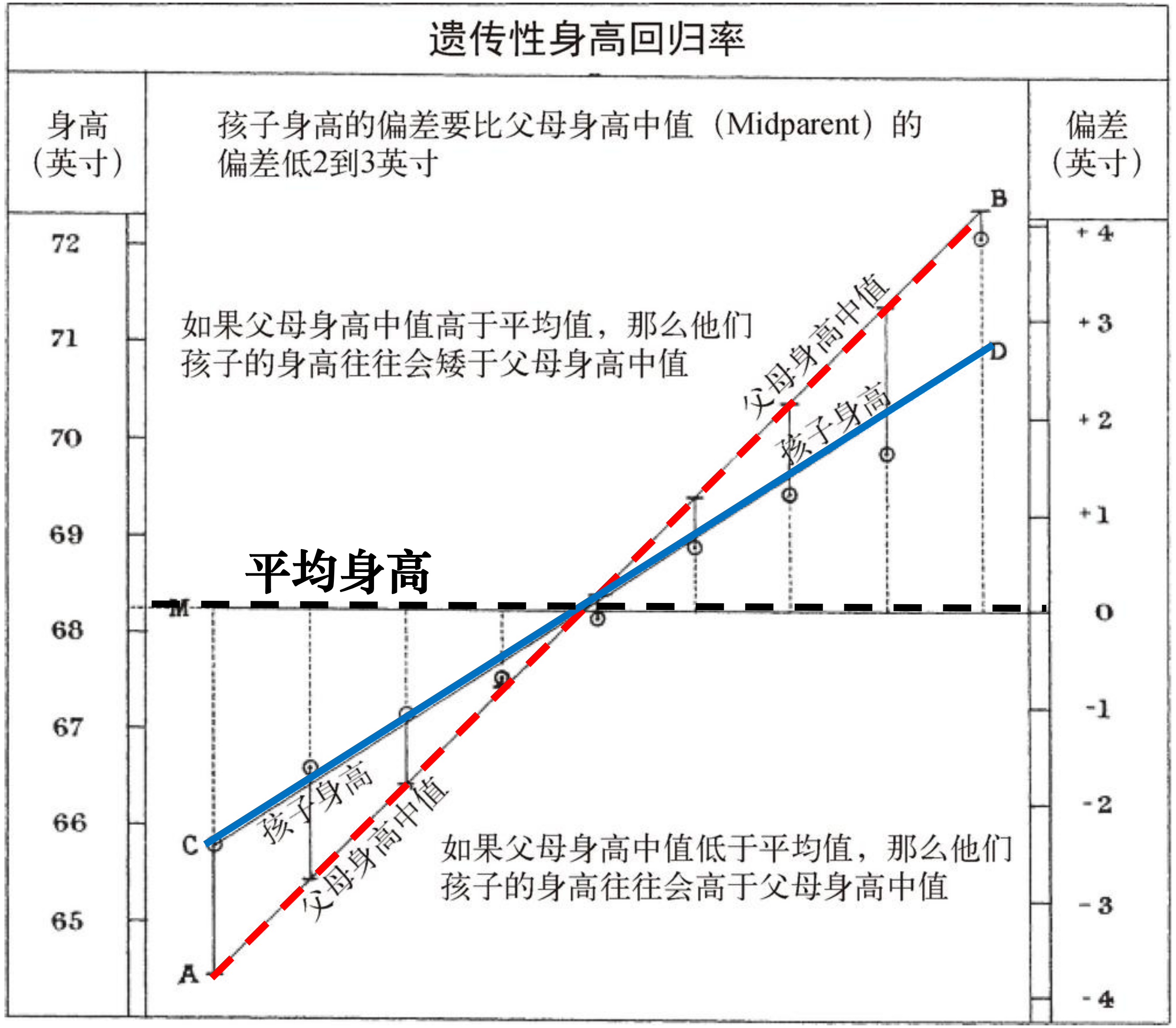
若父母身高小于平均身高，则其子女身高倾向于向上生长，即会更接近于大众平均身高。

此现象被Galton称之为回归现象，即Regression

“回归”起源

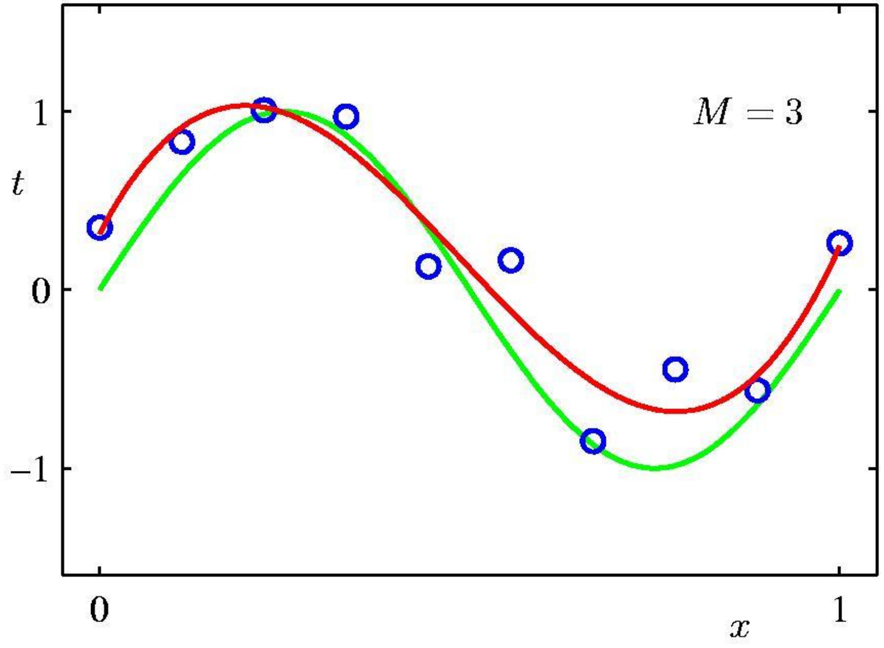
- 父母身高与子女身高之间关系

后代的身高倾向于“回归”到**平均值**



回归的概念

- **回归**
  - 回归属于**监督学习**，其目标是将输入向量分配至由一个或多个**连续变量**组成的输出*
  - 数学表示：给定输入数据x和一个连续型输出y，目标是找到一个函数$f: x \to y$，使$f(x)$与y间的差距尽可能小



*Christopher M. Bishop, Pattern Recognition and Machine Learning

回归举例-房价估计

- 房屋售价与其多种因素有关，现统计了一些房屋的售价和对应的因素数据（包括面积、卧室数量、层数、房龄等）

面积(m^2)	卧室数量(间)	层数	房龄(年)	价格(万元)
210.4	5	1	45	460
141.6	3	2	40	232
153.4	3	2	30	315
85.2	2	1	36	178
…	…	…	…	…

- 如何利用已知的数据构建回归模型，实现对房屋价格的估计?

面积(m^2)	卧室数量(间)	层数	房龄(年)	价格(万元)
150	3	2	30	?

测试样本

回归举例-房价估计

训练样本

面积(m^2) x_1	卧室数量(间) x_2	层数 x_3	房龄(年) x_4	价格(万元) 连续变量y
210.4	5	1	45	460
141.6	3	2	40	232
153.4	3	2	30	315
85.2	2	1	36	178
…	…	…	…	…

N

监督学习

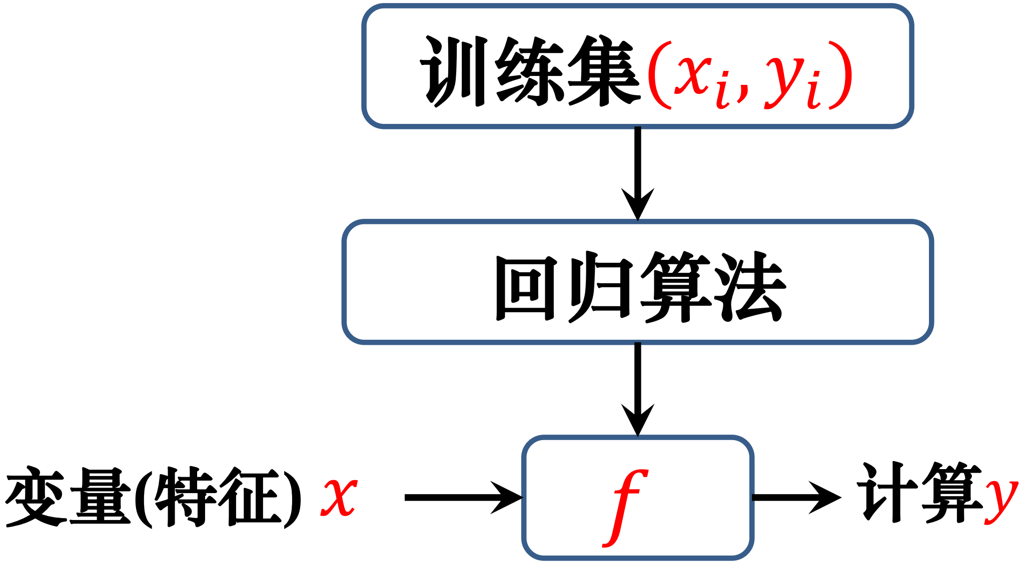
N：训练样本个数

x：输入变量/“特征”

y：输出变量/目标变量

线性回归的概念

● 问题建模



如何表示 f ？

● 线性回归

➢ 假设函数 f 为输入x的线性函数

$$f(x) = w_0 + w_1 x_1 + \cdots + w_m x_m$$
$$= w_0 + \sum_{i=1}^{m} w_i x_i$$
$$= \sum_{i=0}^{m} w_i x_i$$

增加一维 $x_0 = 1$
表示截距项，
转为齐次形式

➢ 写成向量形式

$$f(x) = \boxed{\mathbf{w}^{\mathrm{T}}} x$$

线性回归的概念

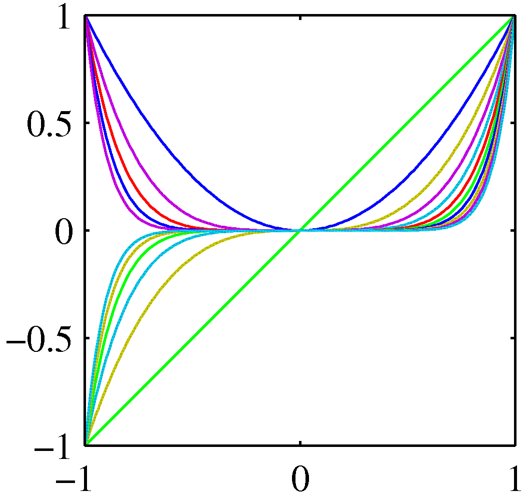
基函数

$$y(x,\mathbf{w}) = \sum_{j=0}^{M-1} w_j \boxed{\phi_j(x)} = \mathbf{w}^{\mathrm{T}}\phi(x)$$

最简单的情况下： $\phi_j(x) = x_j$

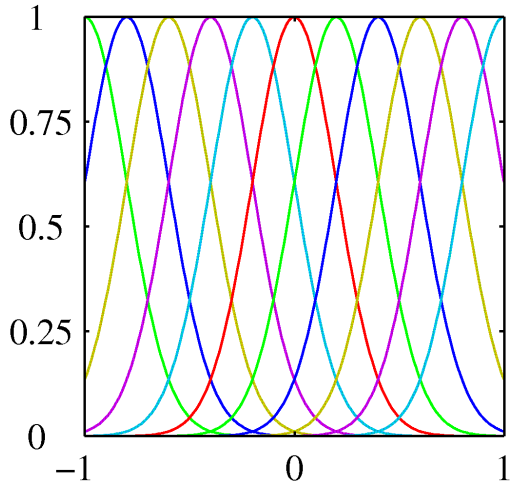
- **多项式基函数**

$$\phi_j(x) = x^j$$



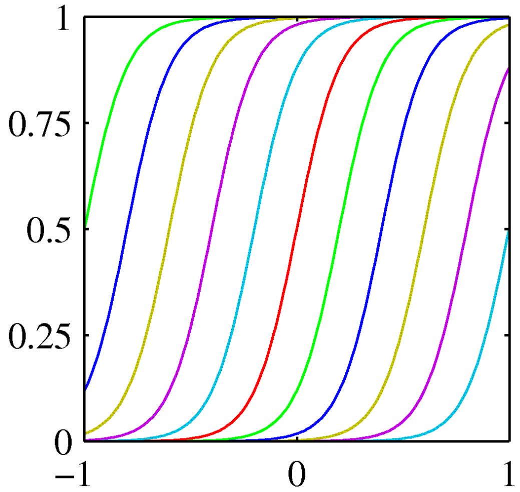
- **高斯基函数**

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$



- **Sigmoid基函数**

$$\phi_j(x) = \sigma\left\{-\frac{x-\mu_j}{s}\right\} \quad \sigma(a) = \frac{1}{1+\exp(-a)}$$



3.2 线性回归的求解

- 标准方程组
- 梯度下降法
- 最大似然估计
- 贝叶斯估计

问题回顾

面积(m²) x_1	卧室数量(间) x_2	层数 x_3	房龄(年) x_4	价格(万元) 连续变量y
210.4	5	1	45	460
141.6	3	2	40	232
153.4	3	2	30	315
85.2	2	1	36	178
…	…	…	…	…

训练样本

N

N：训练样本个数
x：输入变量/“特征”
y：输出变量/目标变量

- 线性回归问题建模
  - 假设函数 f 为输入x的线性函数

$$f(x) = w_0 + w_1 x_1 + \cdots + w_m x_m$$
$$= w_0 + \sum_{i=1}^{m} w_i x_i$$
$$= \sum_{i=0}^{m} w_i x_i$$

  - 写成向量形式

$$f(x) = \mathbf{w}^{\mathrm{T}} x$$

问题分析

- 问题本质：确定模型中的参数w
- 基本思想：基于训练集最小化预测值$\hat{y} = f(x)$与真实输出值y的差异
- 定义目标函数 (又叫代价函数)
  - 例如：平均绝对误差、均方误差和均方根误差
- 以基于均方误差的目标函数为例 $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w}) = \arg\min_{\mathbf{w}} \left[\frac{1}{2} \sum_{i=1}^{N} (f(x_i) - y_i)^2 \right]$
- 求解方法：标准方程组法、梯度下降法

标准方程组法(Normal Equations)

- **目标函数** $J(\mathbf{w}) = \sum_{i=1}^{N}(f(x_i) - y_i)^2 = \sum_{i=1}^{N}(\mathbf{w}^{\mathrm{T}}x_i - y_i)^2$
- **写成矩阵形式** $J(\mathbf{w}) = \sum_{i=1}^{N}(\mathbf{w}^{\mathrm{T}}x_i - y_i)^2 = (\mathbf{Xw} - \mathbf{y})^{\mathrm{T}}(\mathbf{Xw} - \mathbf{y})$
- **其中**

$$\mathbf{X} = \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \cdots \\ x_N^{\mathrm{T}} \end{pmatrix} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1m} \\ x_{20} & x_{21} & & x_{2m} \\ \vdots & & \ddots & \vdots \\ x_{N0} & x_{N1} & \cdots & x_{Nm} \end{bmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}^{\mathrm{T}}$$

标准方程组法(Normal Equations)

- **对目标函数直接求导**，并令其导数等于0，求得极值
- 求导

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} (\mathbf{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}} (\mathbf{X}\mathbf{w} - \mathbf{y})$$
$$= 2\mathbf{X}^{\mathrm{T}} (\mathbf{X}\mathbf{w} - \mathbf{y}) = 0$$

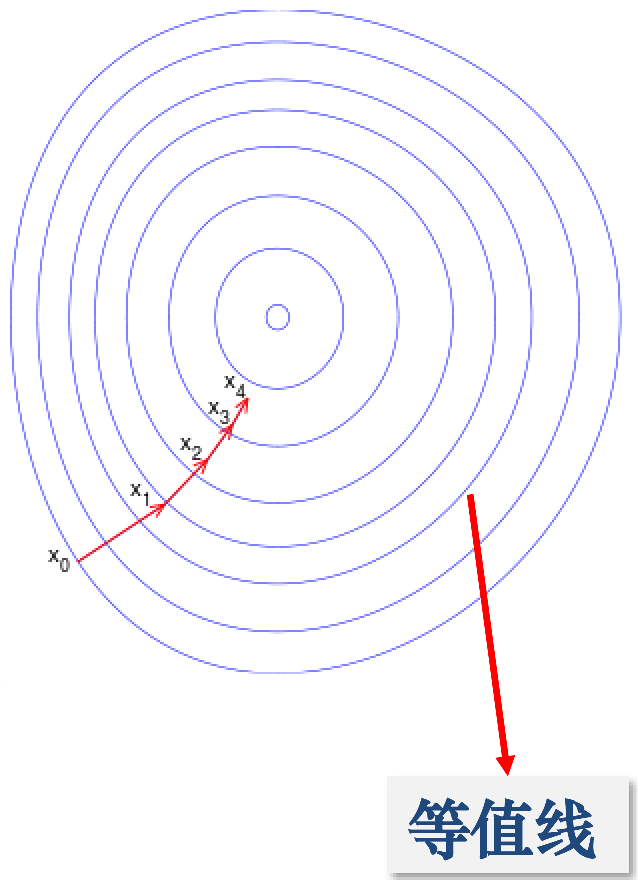
$$\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- 得到模型的参数

$$\mathbf{w}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

梯度下降法(Gradient Descent)

- 梯度下降是一个最优化算法，是求解无约束优化问题最基础的方法之一
- 基本思想
  - 更新参数让目标函数沿**负梯度方向**下降
- 一般流程
  - 首先对参数**w**赋值
  - 更新**w**的值，使得$J(w)$按**梯度下降**的方向减少直到收敛
  - 越接近目标值，搜索步长越小，前进越慢



梯度下降法(Gradient Descent)

- 损失函数：$J(\mathbf{w})=\sum_{i=1}^{N}(f(x_i)-y_i)^2$
- 目标：$\min_{\mathbf{w}} J(\mathbf{w})$
- 梯度下降法的步骤
  - 给定初始值w^0，这个值可以是随机生成的，也可以是一个全零的向量
  - 更新w使得$J(\mathbf{w})$越来越小，α为学习率或更新步长

$$w_j^t=w_j^{t-1}-\alpha\frac{\partial}{\partial w_j}J(\mathbf{w})=w_j^{t-1}-\alpha\sum_{i=1}^{N}(f(x_i)-y_i)\cdot x_{i,j}$$

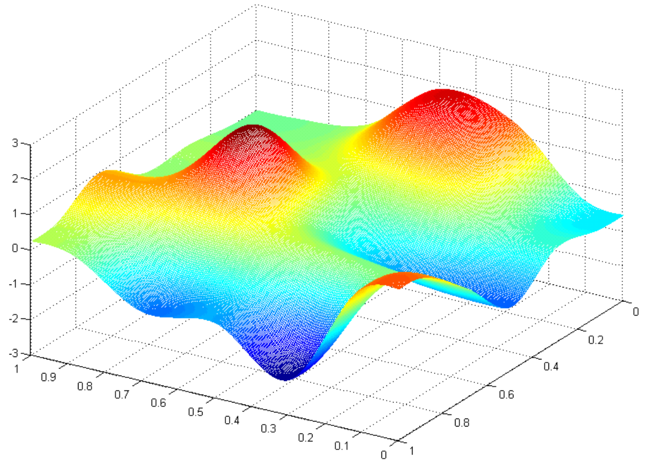
  - 同时更新函数值的各维 $f(x_i)=[w^t]^{\mathrm{T}}x_i$

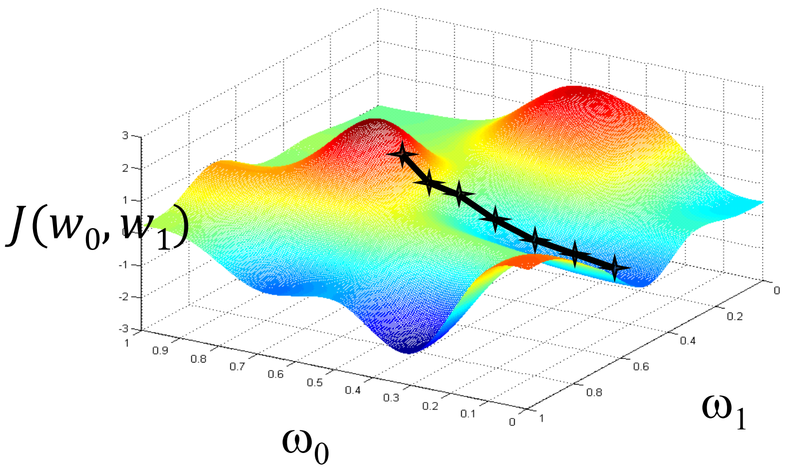
批处理梯度下降(Batch Gradient Descent)

- 在梯度下降方法中，每次更新都利用所有数据

$$w_j^t = w_j^{t-1} - \alpha \boxed{\sum_{i=1}^{N} (f(x_i) - y_i) \cdot x_j} \rightarrow \text{批处理}$$

- 对于凸函数，可以达到全局最优
- 在大样本条件下，批处理梯度下降的迭代速度很慢





随机梯度下降(Stochastic Gradient Descent)

- 基本思想：每次只用一个样本(x_r, y_r)
- 随机梯度下降的特点
  - 更新速度较快 $w_j^t = w_j^{t-1} - \boxed{\alpha(f(x_r) - y_r) \cdot x_r}$
  - 更适用于大样本数据情况
  - 对于更复杂的优化目标，可以跳出局部最优解
- 变化形式
  - 在线学习(Online Learning)，每次“看”一个样本，对所有样本循环使用多次(一次循环称为一个Epoch)
  - 每次可以看一些样本(Mini-Batch)

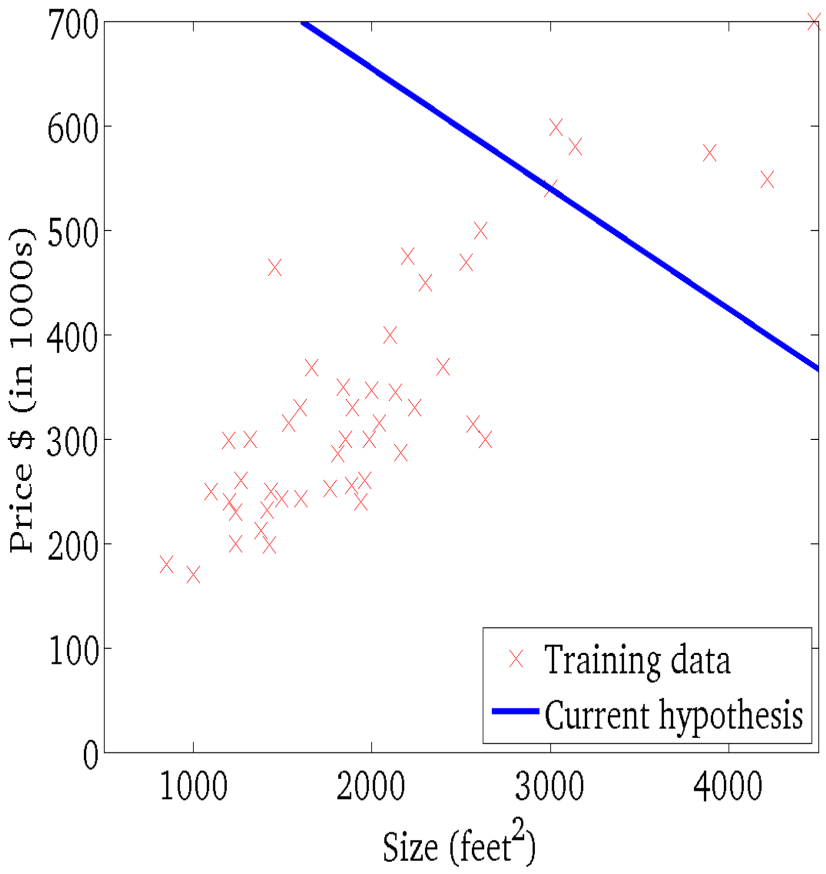
随机梯度下降

- 问题描述：如何利用已知的数据构建回归模型，实现对房屋价格的估计(简化房屋价格的影响因素，只保留房屋面积)
- 问题建模：构建$f_{\mathbf{w}}(x) = w_1 x + w_0$来根据房屋面积$\boldsymbol{x}$估计房屋价格$\boldsymbol{y}$
- 代价函数：$J(\mathbf{w}) = \sum_{i=1}^{N}(f(x_i) - y_i)^2$
- 问题求解：使用随机梯度下降法求解

随机梯度下降

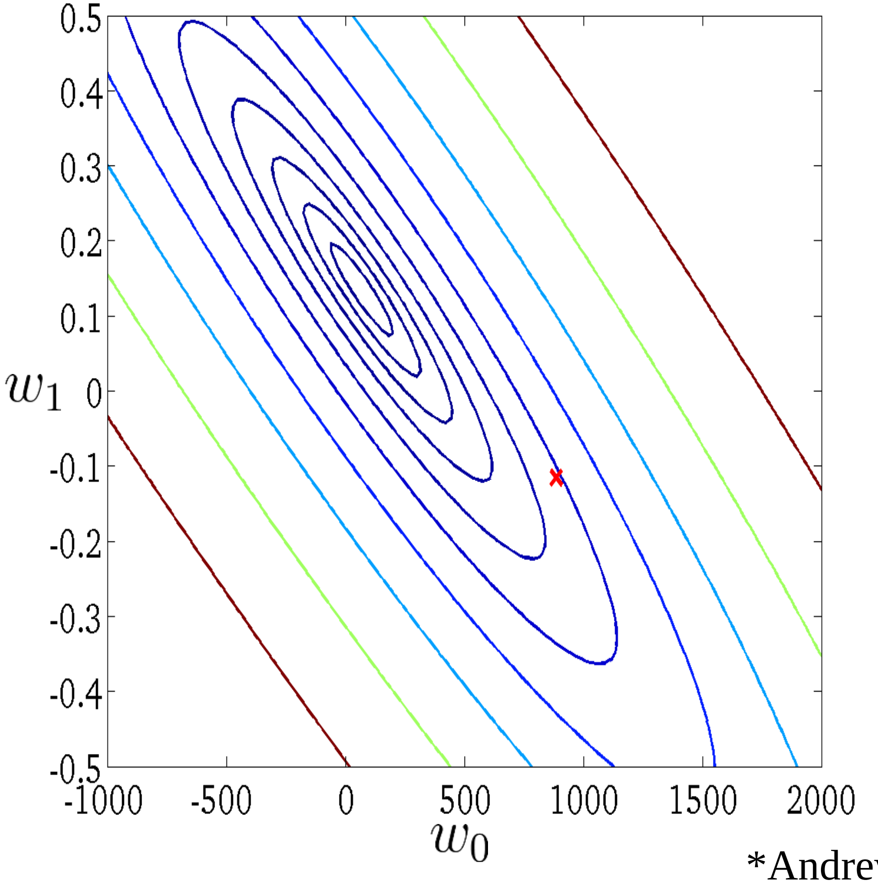
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

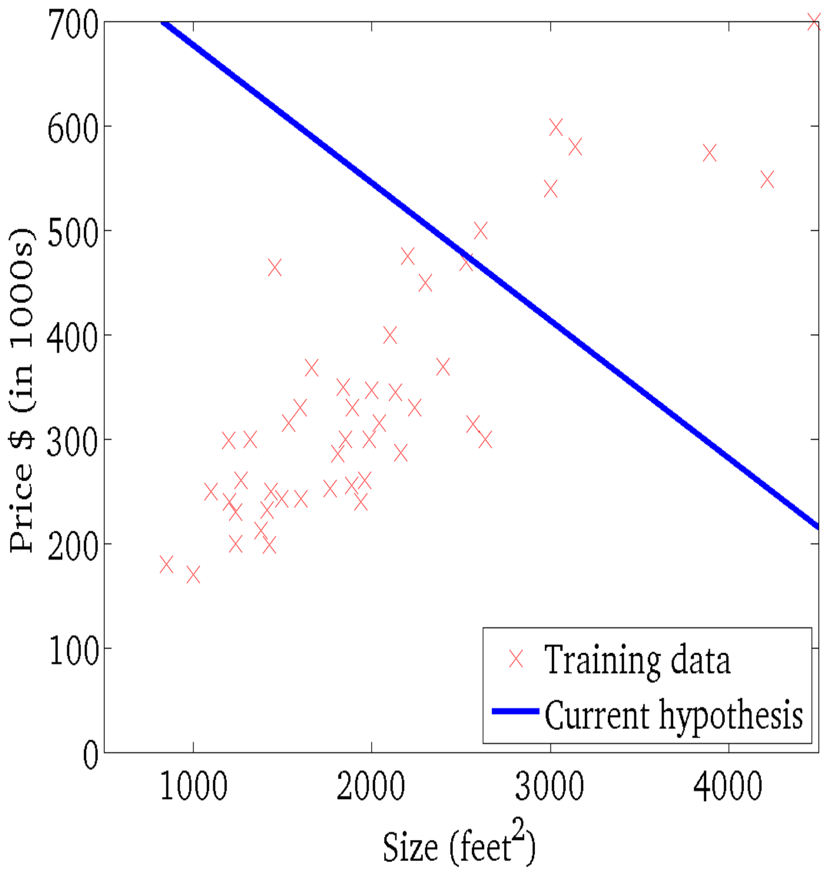


*Andrew Ng, Machine Learning

随机梯度下降

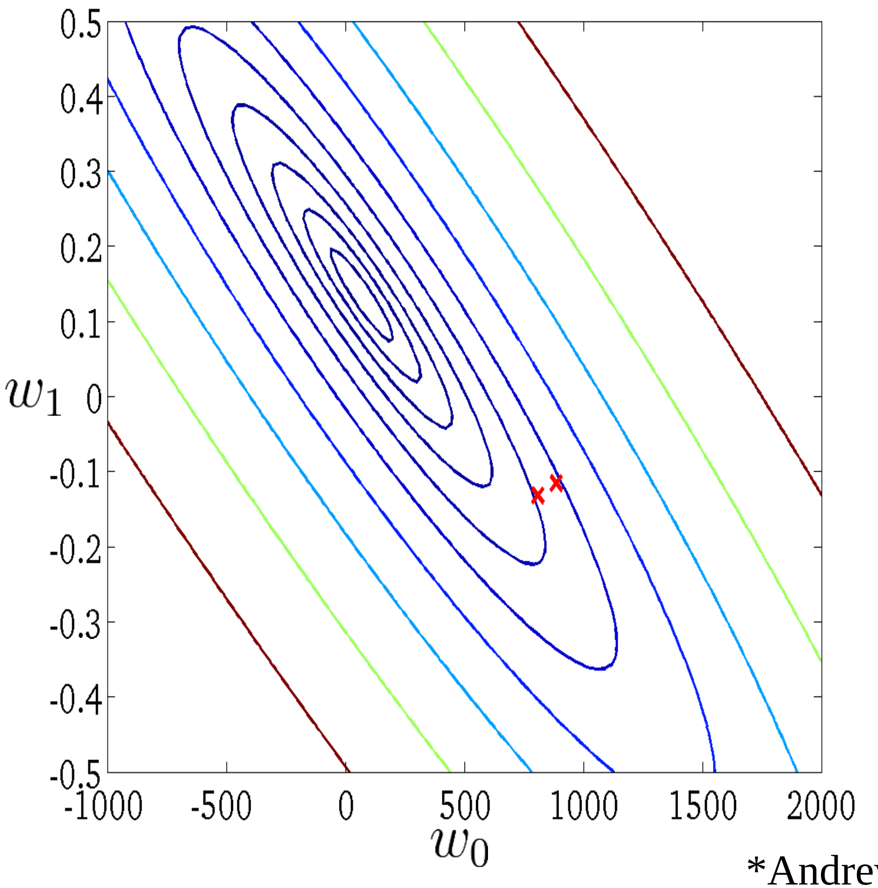
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

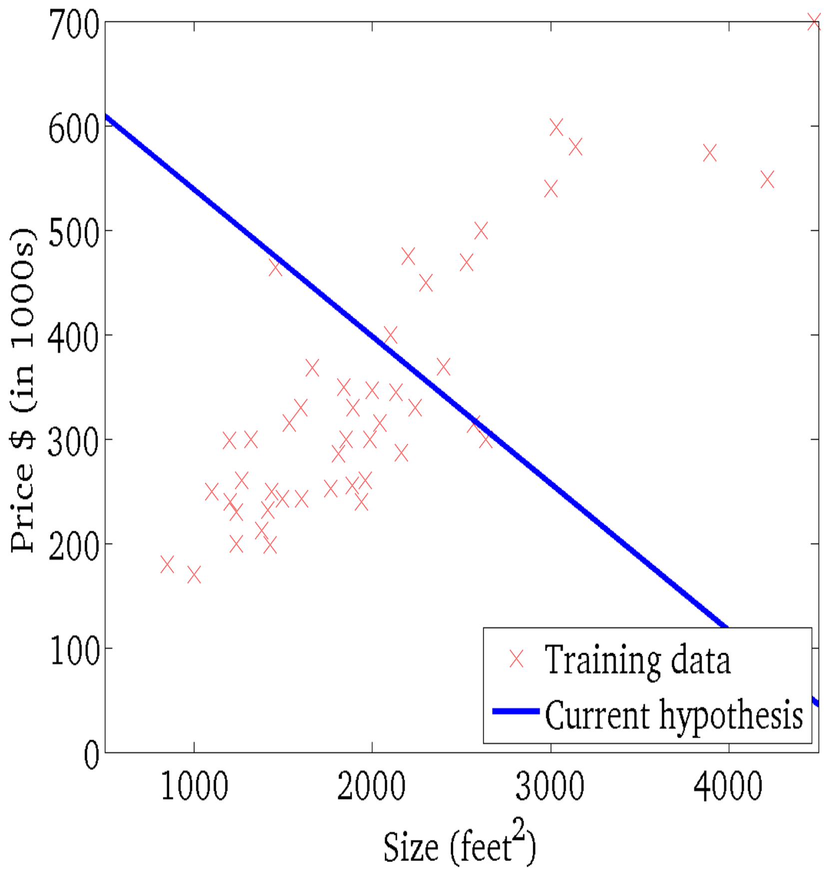


*Andrew Ng, Machine Learning

随机梯度下降

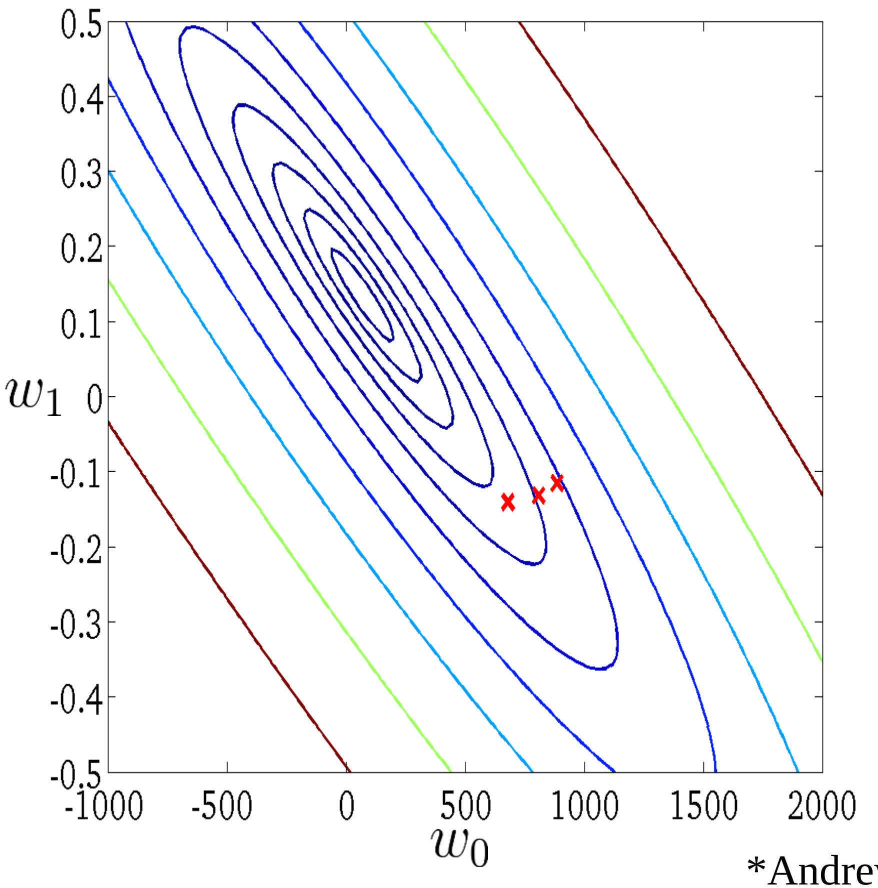
$$f_{\mathbf{w}}(x)$$

(对于固定的w_1, w_0, 这是关于x的函数)



$$J(\mathbf{w})$$

(关于参数w_1, w_0的函数)

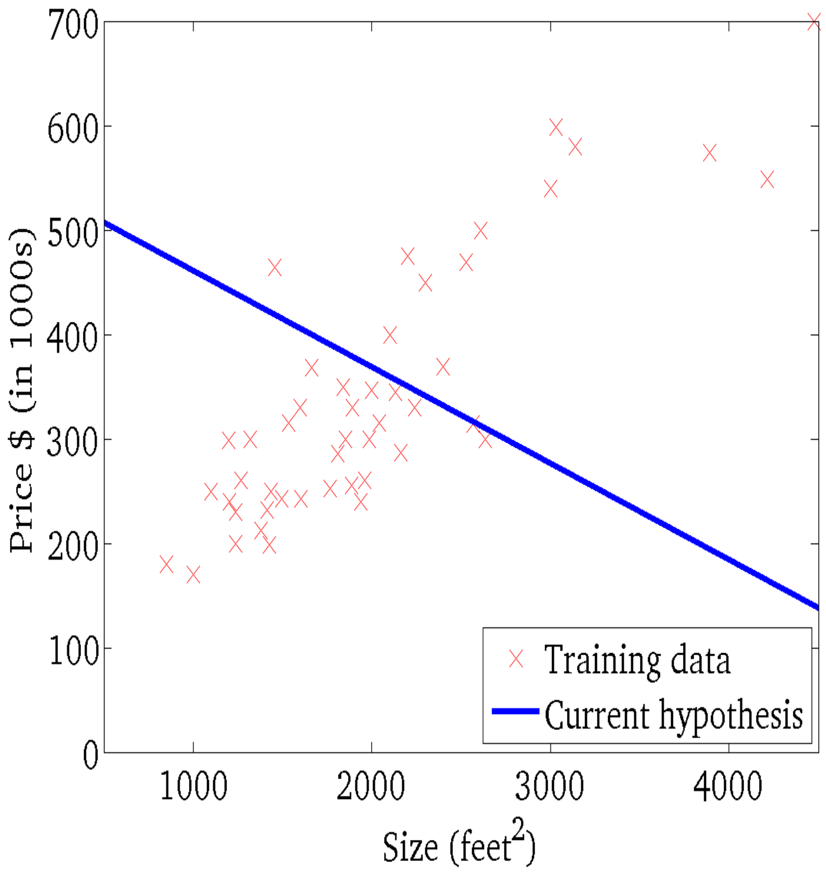


*Andrew Ng, Machine Learning

随机梯度下降

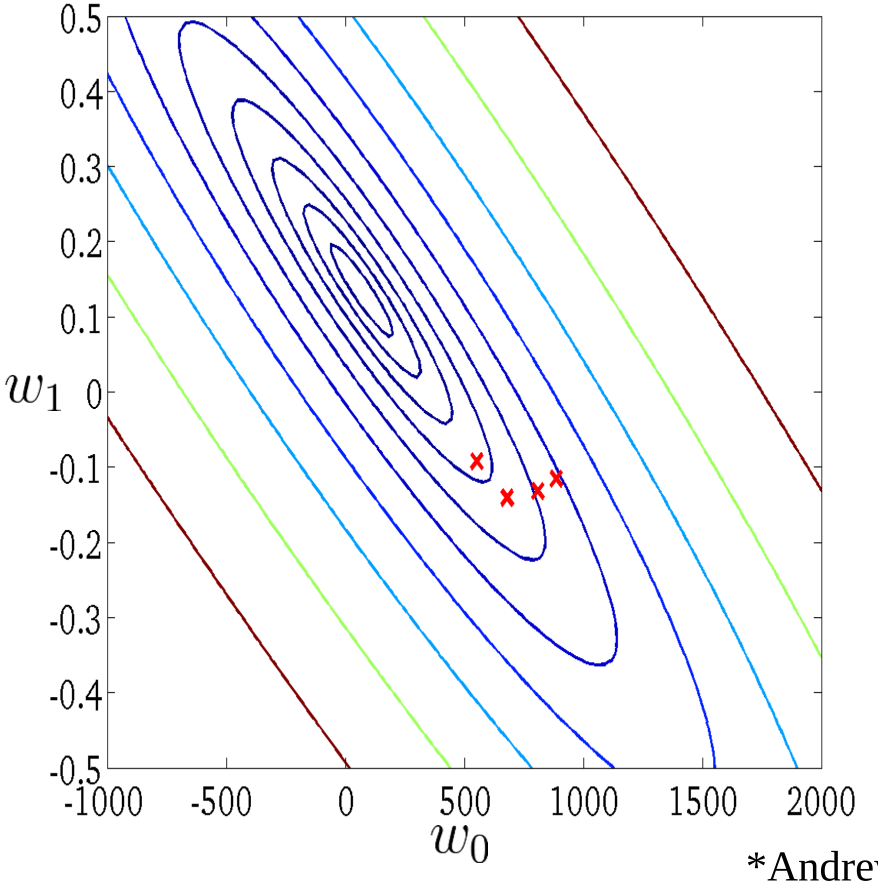
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

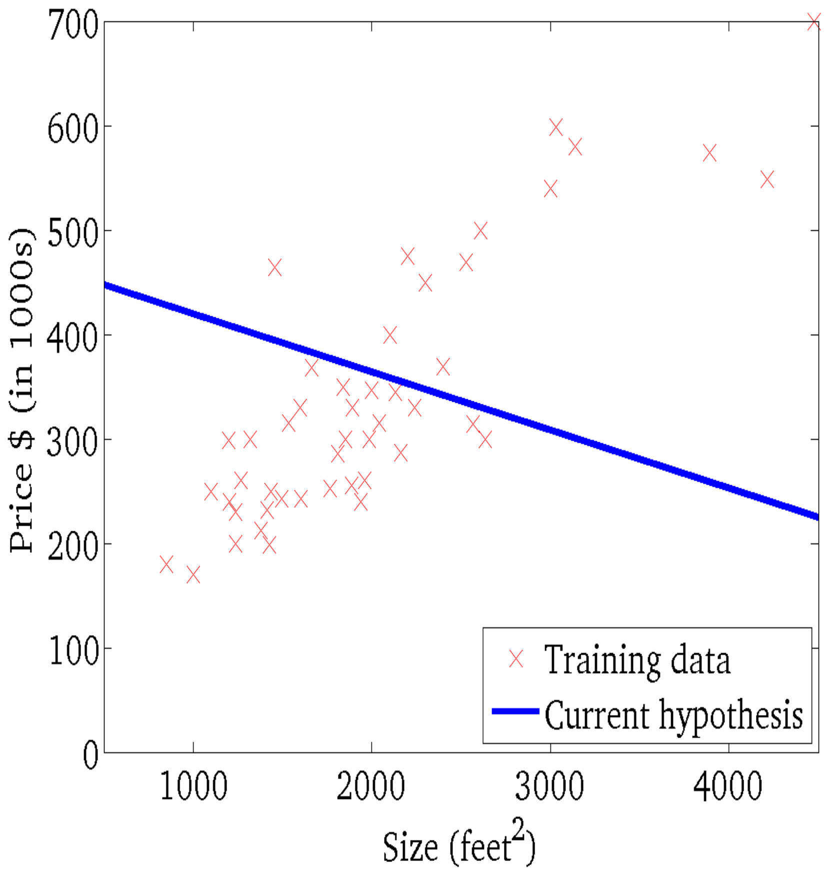


*Andrew Ng, Machine Learning

随机梯度下降

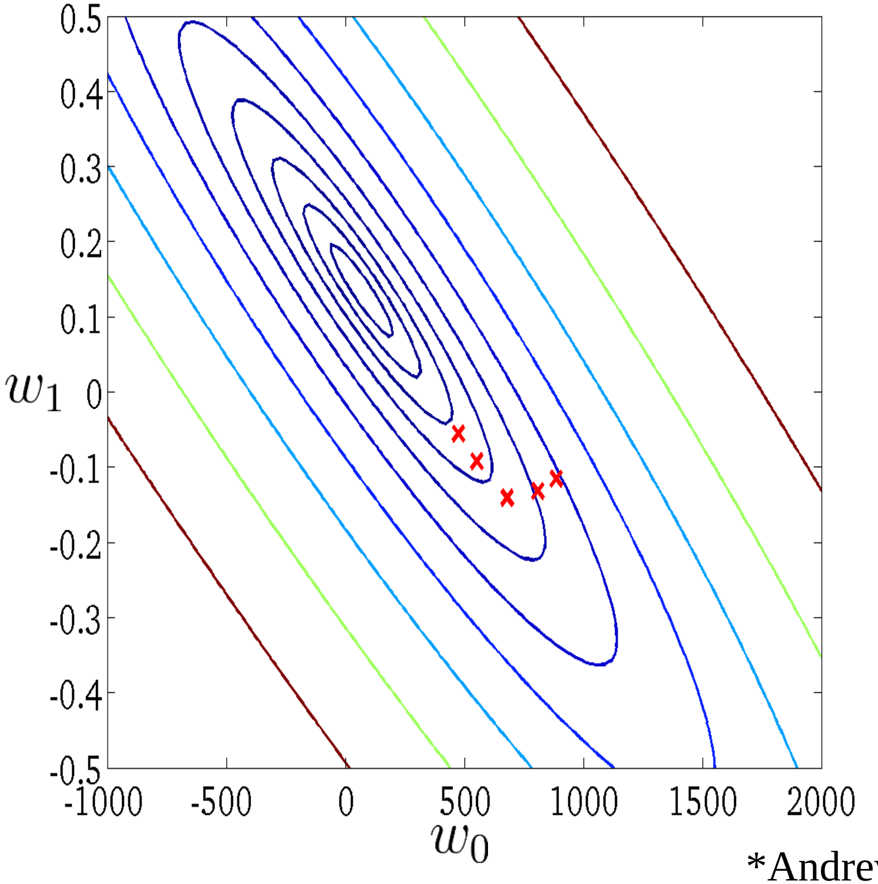
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

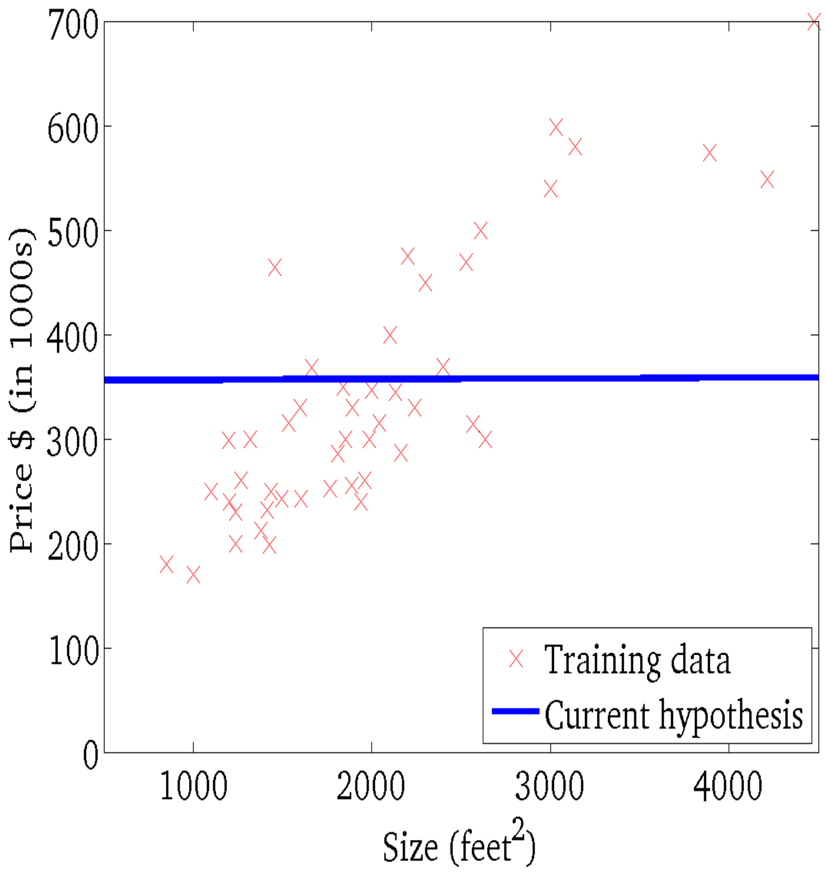


*Andrew Ng, Machine Learning

随机梯度下降

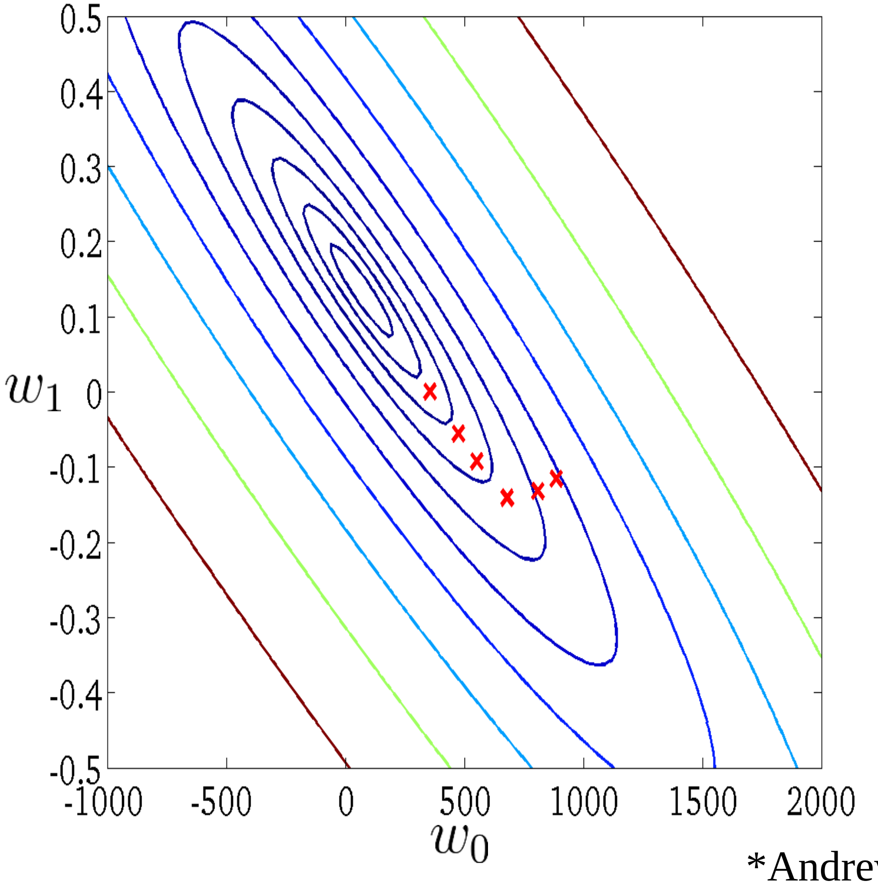
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

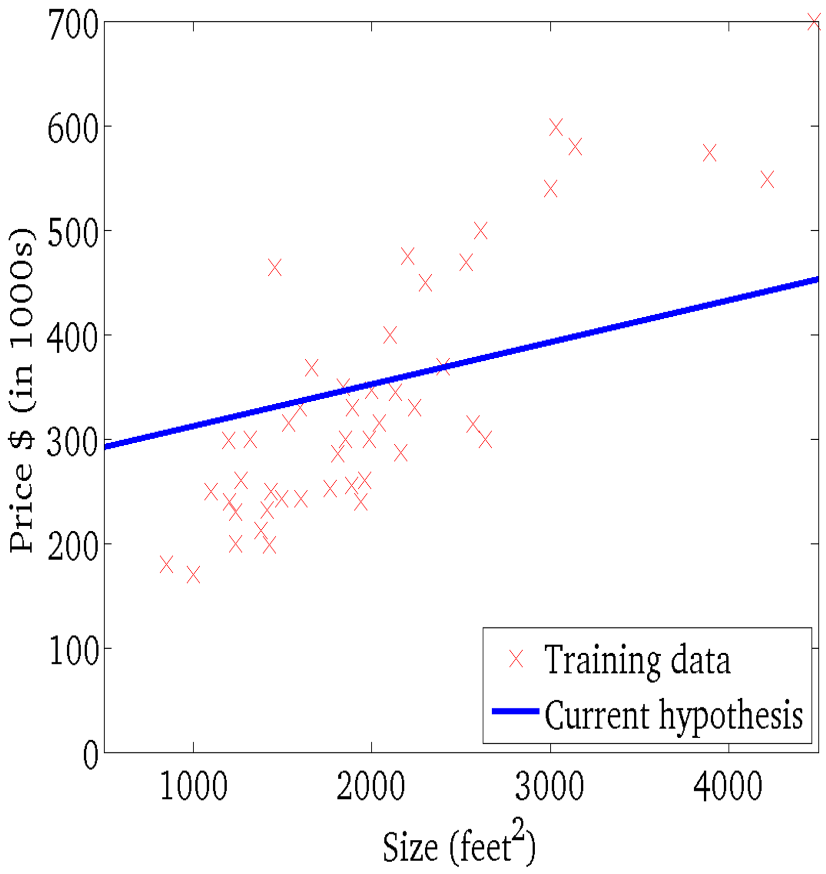


*Andrew Ng, Machine Learning

随机梯度下降

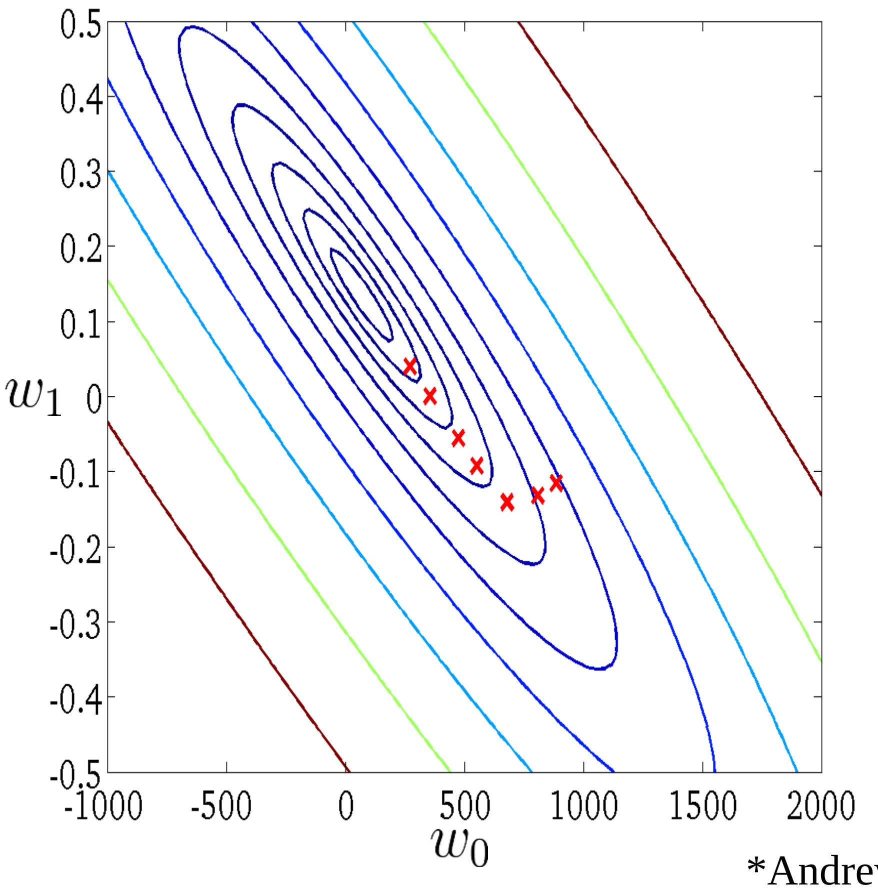
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

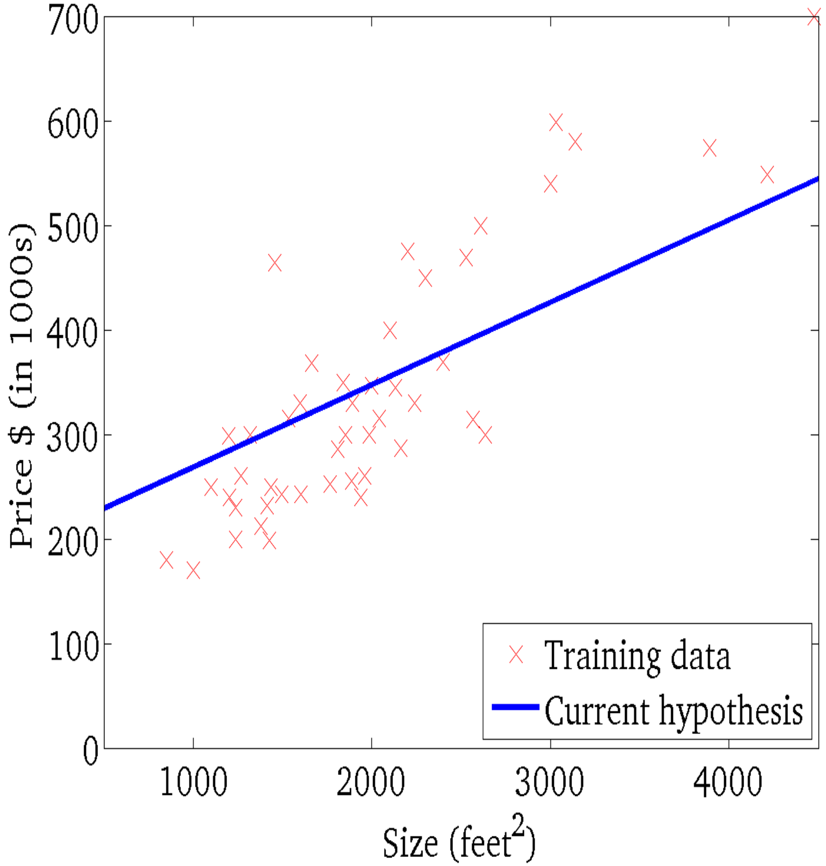


*Andrew Ng, Machine Learning

随机梯度下降

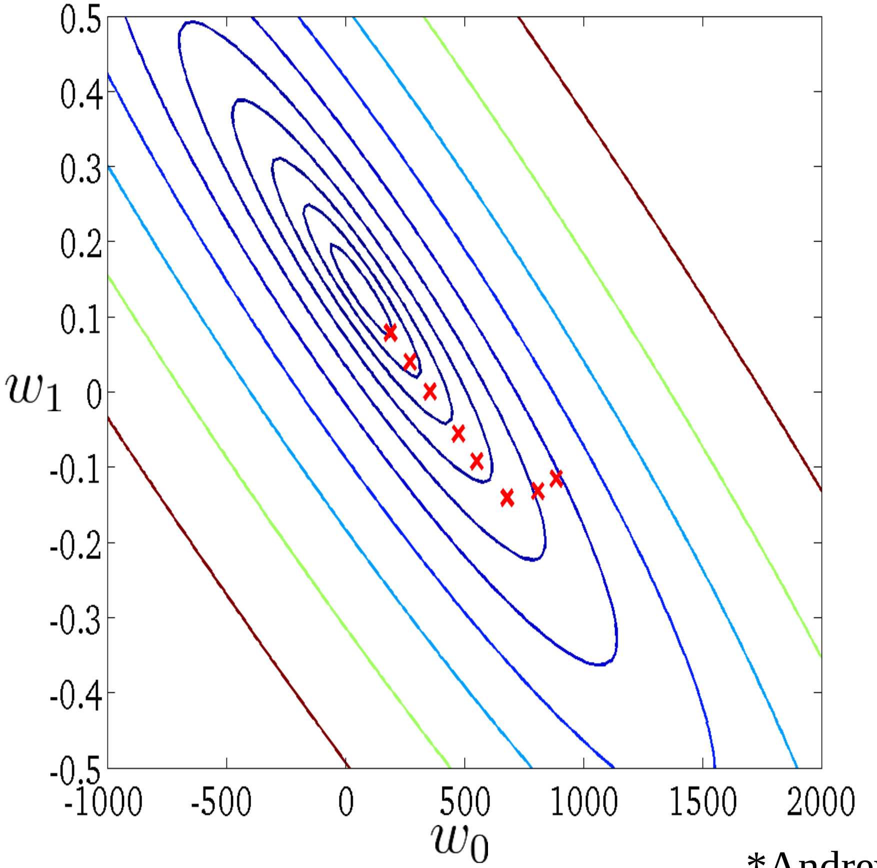
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)

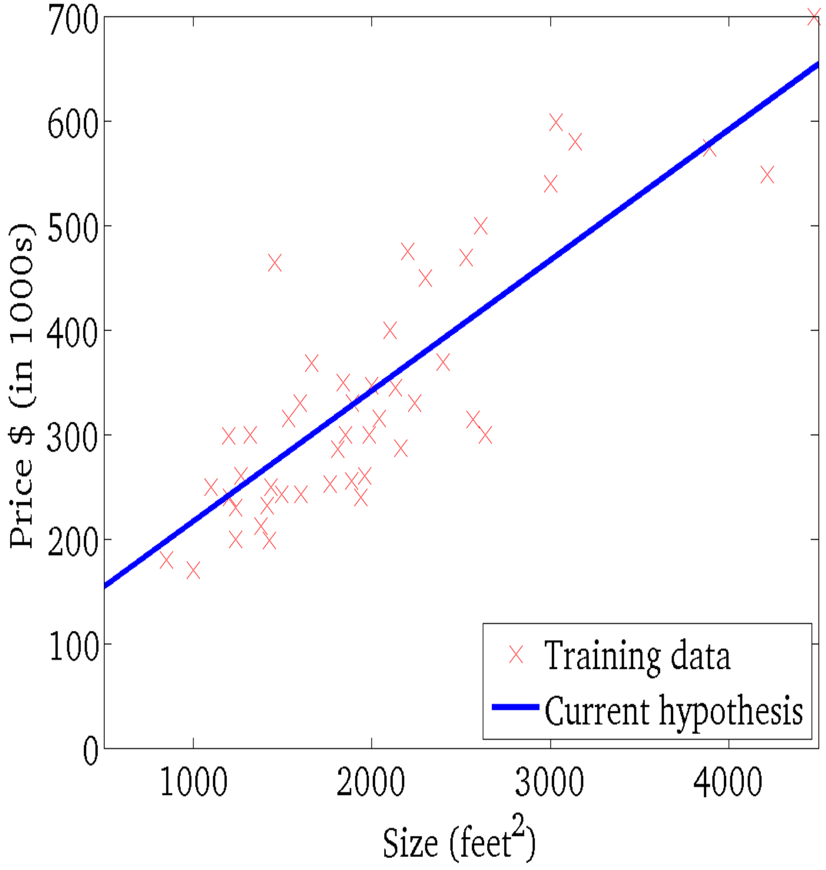


*Andrew Ng, Machine Learning

随机梯度下降

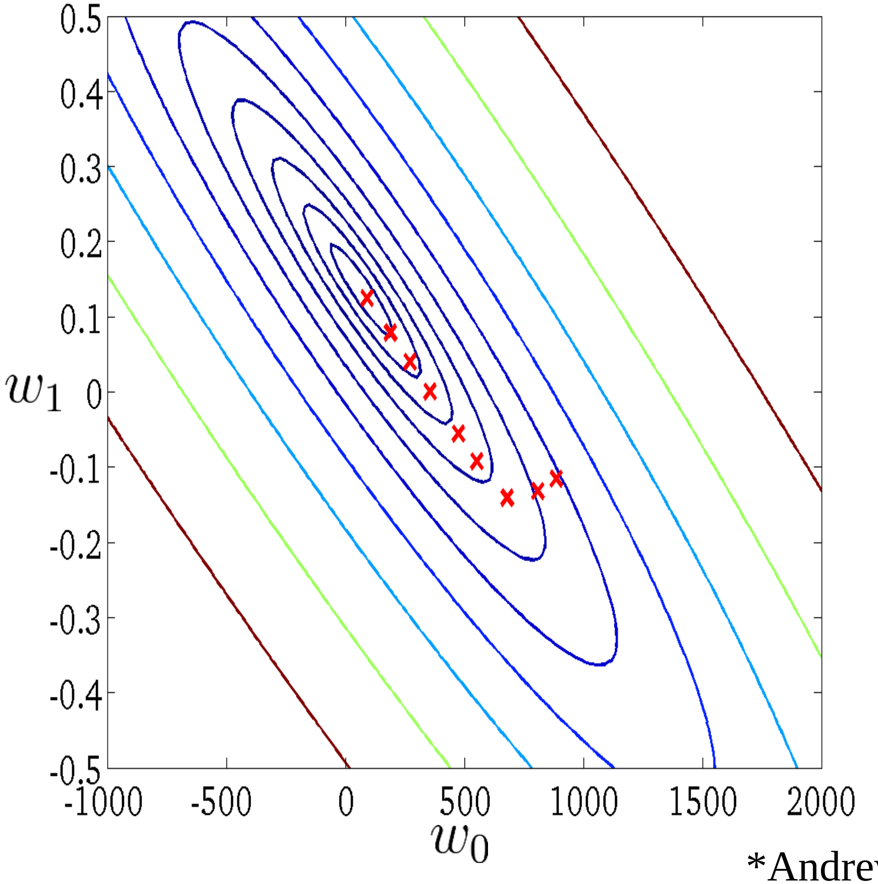
$f_{\mathbf{w}}(x)$

(对于固定的w_1, w_0, 这是关于x的函数)



$J(\mathbf{w})$

(关于参数w_1, w_0的函数)



*Andrew Ng, Machine Learning

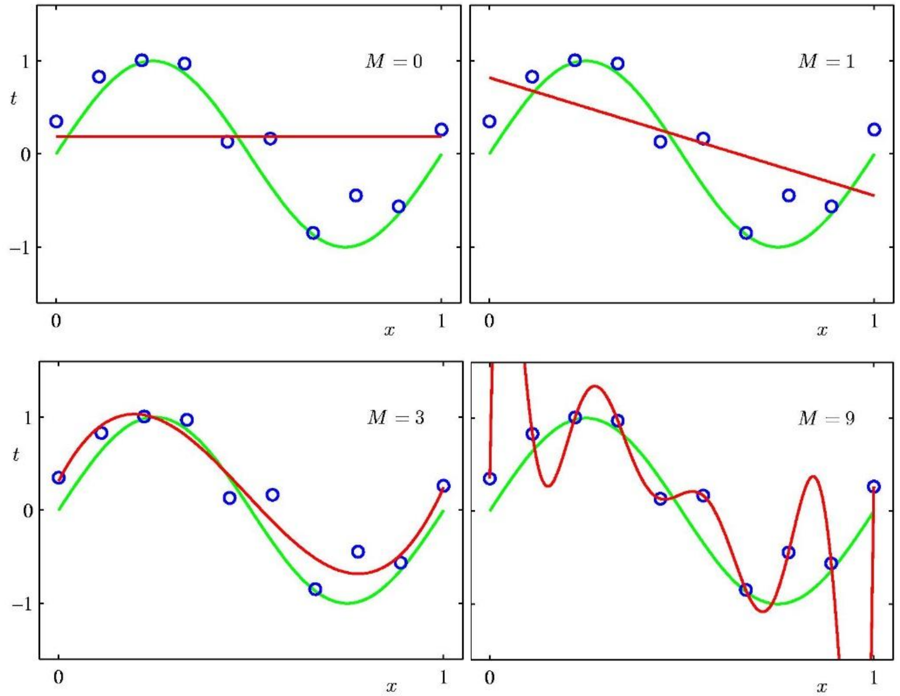
标准方程组 *v.s.* 梯度下降

	标准方程组	梯度下降
超参数	不需要超参α	需要选择α
迭代次数	不需要迭代多次	需要迭代多次
归一化	无需数据归一化	需要数据归一化
适用情况	样本量大时不适用 需要计算$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$	样本量非常大时也适用

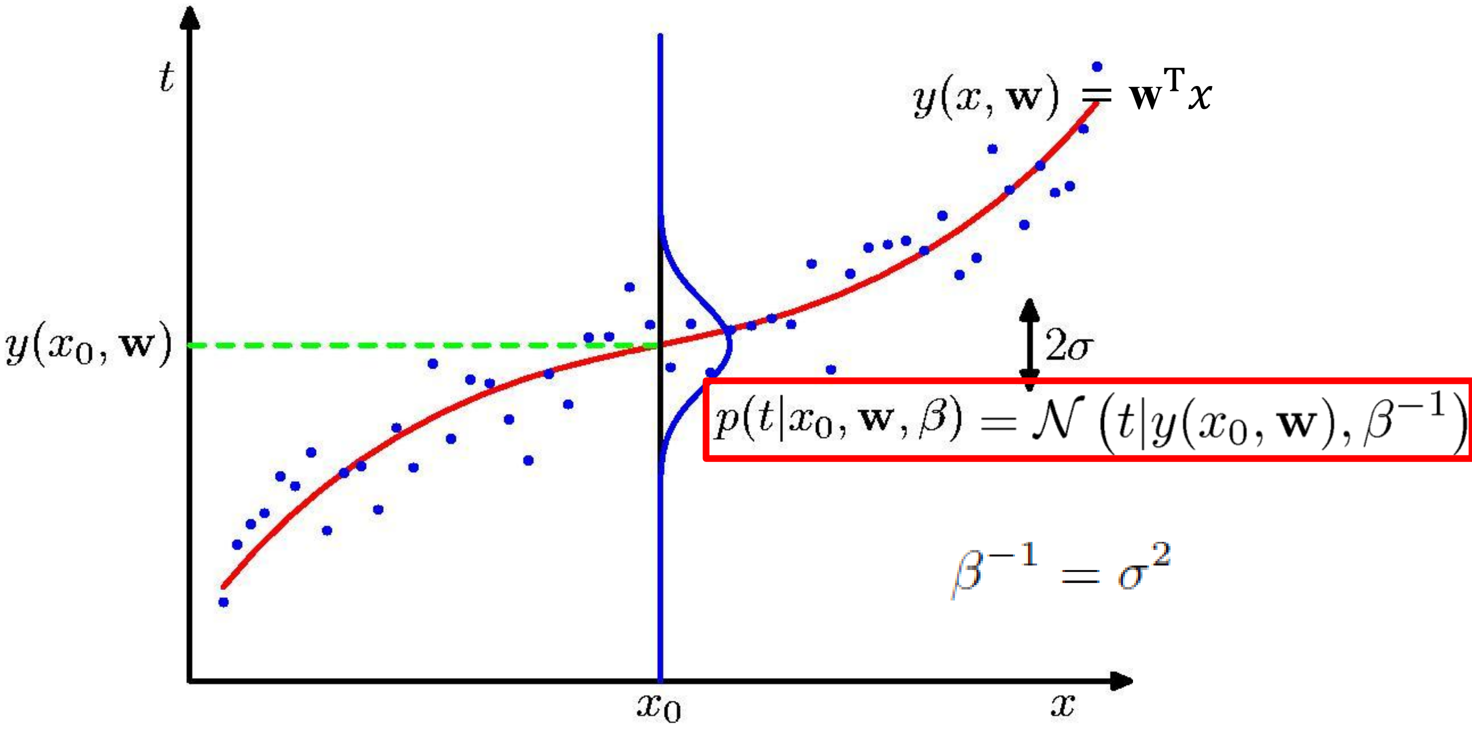
样本量较小时选用标准方程组求解
样本量较大时选用梯度下降法求解

从概率视角看回归

- 第一章的曲线拟合中，目标变量值出现不确定性，可用概率分布表示
- 假定目标值t服从以$y(x,\mathbf{w})$为均值，σ^2为方差的高斯分布



第一章中示例：曲线拟合



最大似然估计与贝叶斯估计

- 最大似然估计
  - 把待估计的参数w看做是确定的量，只是其取值未知。最佳估计就是使得观测到的样本的概率最大的那个值。

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n\,|\,y(x_n,\mathbf{w}),\beta^{-1})$$

- 贝叶斯估计
  - 把待估计的参数w看做是符合某种先验概率分布的随机变量。对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，从而利用样本信息修正了对参数的初始估计值。

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta)=\int p(t|x,\mathbf{w})\,p(\mathbf{w}|\mathbf{x},\mathbf{t})\mathrm{d}\mathbf{w}=\mathcal{N}\left(t\middle|m_N^{\mathrm{T}}\mathbf{x},\sigma_N^2(\mathbf{x})\right)$$

最大似然估计(Maximum Likelihood Estimation)

已知训练数据x和t，现有新输入变量x，需要预测目标变量t

似然函数表示为$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n\,|\,y(x_n,\mathbf{w}),\beta^{-1})$

均方误差

对数似然函数为$\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=-\boxed{\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2}+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$

通过最大化对数似然函数解得$\boxed{\mathbf{w}_{ML}}$

等价于以均方误差为损失函数的线性回归的解

此时高斯分布的参数 $\frac{1}{\beta_{ML}}=\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}_{ML})-t_n\}^2$

得到概率分布 $p(t|x,\mathbf{w}_{ML},\beta_{ML})=\mathcal{N}(t\,|\,y(x,\mathbf{w}_{ML}),\beta_{ML}^{-1})$

MAP：离贝叶斯更近一步

假定w的先验分布为

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

其中α是分布精度， M是w的阶数

根据贝叶斯定理 $p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto \boxed{p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)}\boxed{p(\mathbf{w}|\alpha)}$

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n,\mathbf{w}),\beta^{-1})$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I})$$

取对数可知，最大化$\ln p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta)$即最大化

$$-\ln\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

MAP：离贝叶斯更近一步

最大化$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta)$即最小化

$$\boxed{\beta\tilde{E}(\mathbf{w}) = \underbrace{\boxed{\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2}}_{\text{平方误差}} + \underbrace{\boxed{\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}}}_{\text{正则化项}}} \rightarrow \text{正则平方误差函数}$$

令$\lambda=\frac{\alpha}{\beta}$为正则化系数，令$\nabla_w\tilde{E}(\mathbf{w})=\mathbf{0}$得解

$$\boxed{\mathbf{w}_{MAP} = (\mathbf{x}^{\mathrm{T}}\mathbf{x}+\lambda\mathbf{I})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{t}}$$

等价于以正则平方误差为损失函数的线性回归的解

贝叶斯估计

预测分布简单表示为

$$p(t|x,\mathbf{x},\mathbf{t}) = \int \boxed{p(t|x,\mathbf{w})}\,\boxed{p(\mathbf{w}|\mathbf{x},\mathbf{t})}\,\mathrm{d}\mathbf{w}$$

其中

$$p(t|x,\mathbf{w},\beta) = \mathcal{N}(t\,|\,y(x,\mathbf{w}),\beta^{-1})$$

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

预测分布涉及两个高斯分布的卷积，可采用以下形式

$$p(t|x,\mathbf{x},\mathbf{t}) = \mathcal{N}(t\,|\,m(x),s^2(x)\,)$$

贝叶斯估计

$$p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\mathbf{w})\,p(\mathbf{w}|\mathbf{x},\mathbf{t})\mathrm{d}\mathbf{w} = \mathcal{N}(t\,|\,m(x),s^2(x)\,)$$

代入

$$y(x,\mathbf{w}) = \sum_{j=0}^{M-1} w_j x_j = \mathbf{w}^{\mathrm{T}} x$$

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

使用边际高斯和条件高斯的转换可知

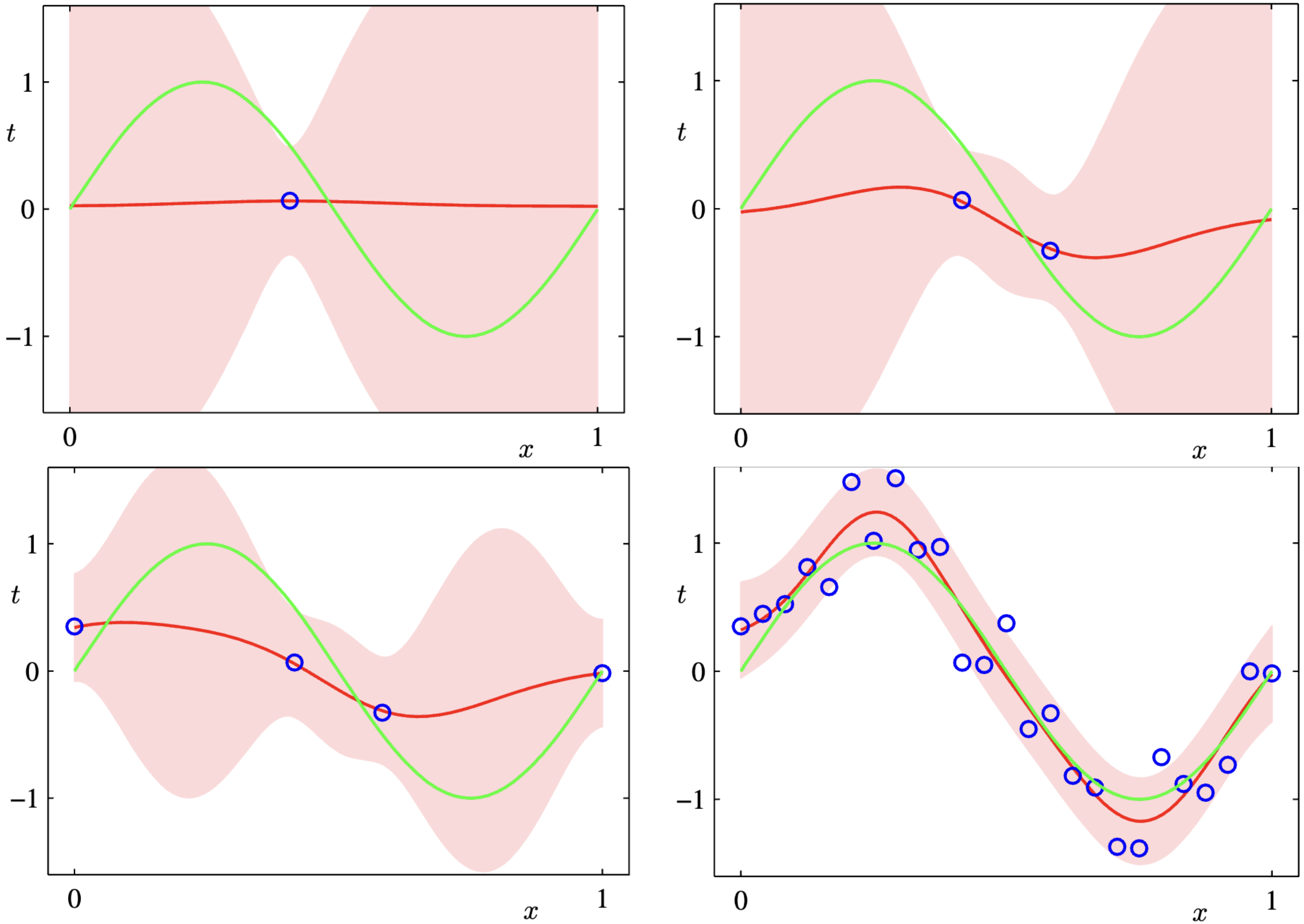
$$m(x) = \beta x^{\mathrm{T}} S \sum_{n=1}^{N} x_n t_n \qquad s^2(x) = \beta^{-1} + x^{\mathrm{T}} S x$$

其中

$$\mathrm{S}^{-1} = \alpha\mathbf{I} + \beta \sum_{n=1}^{N} x_n x_n^{\mathrm{T}}$$

贝叶斯估计

$$p(t|x_0, \mathbf{w}, \beta) = \mathcal{N}\left(t \mid \mathbf{w}^{\mathrm{T}}\emptyset(x_0), \beta^{-1}\right)$$



3.3 什么是线性分类器?

- 线性分类器的概念
- 线性判别函数的一般形式
- 线性判别函数的几何理解
- 广义线性判别函数
- 线性判别函数的齐次化

从贝叶斯分类说起

- **贝叶斯公式：**

先验概率

似然概率(类条件概率)

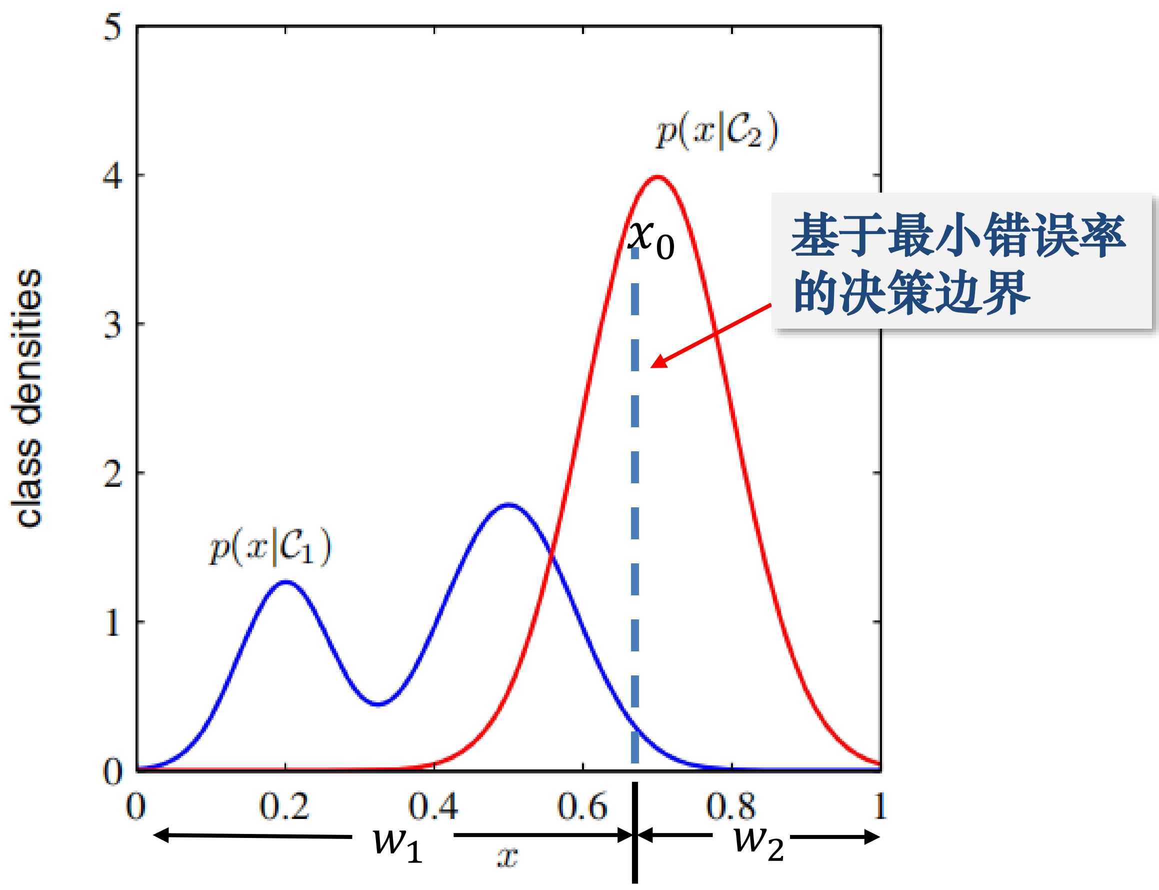
后验概率

$$\boxed{P(w_i|\mathbf{x})} = \frac{\boxed{P(w_i)}\boxed{P(\mathbf{x}|w_i)}}{\boxed{\sum_{j=1}^{n} P(w_j)P(\mathbf{x}|w_j)}}$$

证据因子

贝叶斯决策根据贝叶斯公式计算后验概率，基于最大后验概率进行判决

- **决策规则**



几个问题

- 已知类条件概率密度$p(\mathbf{x}|w_i)$和先验概率$P(w_i)$，计算后验概率$p(w_i|\mathbf{x})$进行决策

若类条件概率密度参数未知?

- 已知类条件概率密度$p(\mathbf{x}|w_i)$的参数表达式，利用样本估计$p(\mathbf{x}|w_i)$的未知参数，再利用贝叶斯定理将其转化成后验概率$p(w_i|\mathbf{x})$进行决策

若类条件概率密度形式难以确定?

- 非参数方法估计

需要大量样本

线性分类器

- 利用样本集直接设计分类器

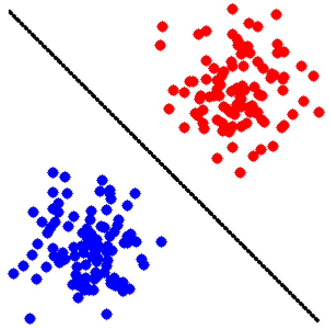
线性回归： $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$　　利用样本估计$\mathbf{w}$, 对于给定$\mathbf{x}$, 计算y

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right)$$

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases}$$

$\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \geq 0$　　C_1

$\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \leq 0$　　C_2



➢ 将分类器设计问题转化为求准则函数极值的问题；准则函数：分类器设计的某些要求的函数形式

线性判别函数的一般形式

- 二分类情况下线性判别函数一般形式为$g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$

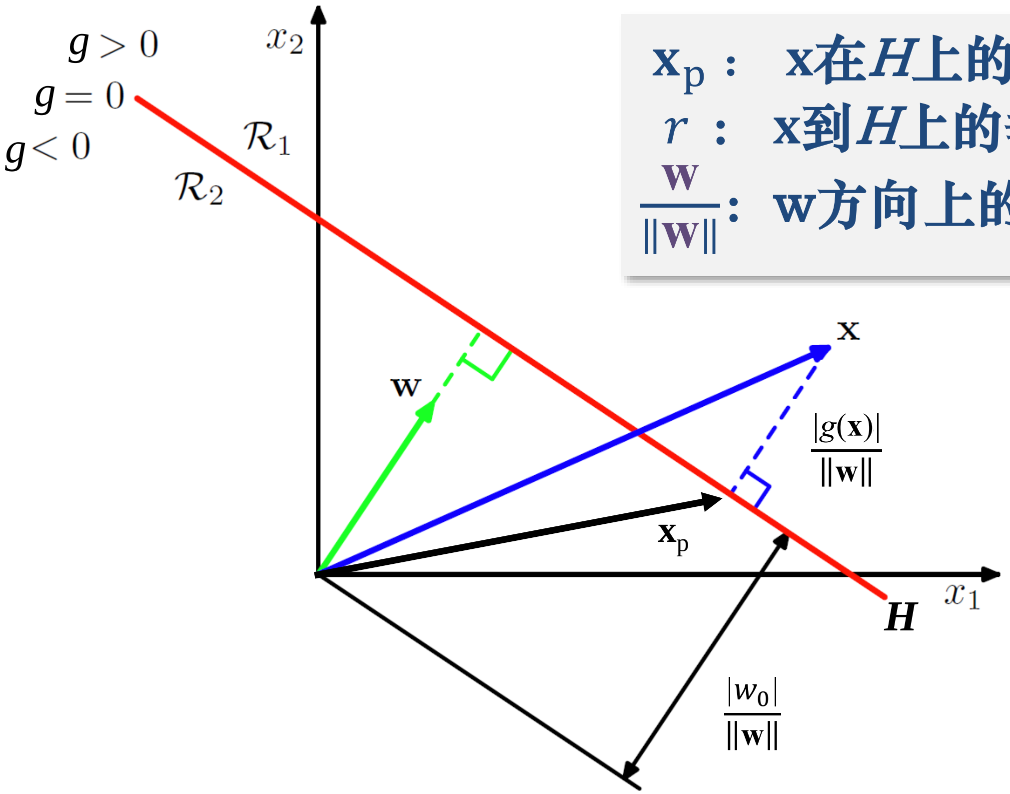
其中$\mathbf{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$是**特征向量/样本向量**；$\mathbf{w}=\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$是**权向量**；$w_0$是**阈值权** (常数)

- 分类时，如果$\begin{cases} g(\mathbf{x}) > 0, \mathbf{x} \in w_1 \\ g(\mathbf{x}) < 0, \mathbf{x} \in w_2 \\ g(\mathbf{x}) = 0, \text{可将}\mathbf{x}\text{分到任意一类或拒绝} \end{cases}$

- $g(\mathbf{x}) = 0$定义了一个决策面，当$g(\mathbf{x})$为线性函数时，决策面就是超平面

线性判别函数的几何理解

- 如果$\mathbf{x}_1$和$\mathbf{x}_2$都在决策面H上，则有$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 + w_0 = \mathbf{w}^{\mathrm{T}}\mathbf{x}_2 + w_0$，即$\mathbf{w}^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_2) = 0$，说明$\mathbf{w}$和超平面$H$上任一向量正交，即$\mathbf{w}$是$H$的法向量
- 判别函数$g(\mathbf{x})$可看成是特征空间中某点$\mathbf{x}$到超平面H距离的一种代数度量



$\mathbf{x}_{\mathrm{p}}$：$\mathbf{x}$在H上的投影向量

r：$\mathbf{x}$到H上的垂直距离

$\frac{\mathbf{w}}{\|\mathbf{w}\|}$：$\mathbf{w}$方向上的单位向量

$$\mathbf{x} = \mathbf{x}_{\mathrm{p}} + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^{\mathrm{T}}\left(\mathbf{x}_{\mathrm{p}} + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{p}} + r\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{\|\mathbf{w}\|} + w_0 \\ &= r\|\mathbf{w}\| \end{aligned}$$

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

线性判别函数的几何理解

- 若x为原点，则$g(\mathbf{x}) = w_0$；从原点到超平面H的距离$r_0 = \dfrac{w_0}{\|\mathbf{w}\|}$

  如果$w_0 > 0$, 则原点在H的正侧

  如果$w_0 < 0$, 则原点在H的负侧

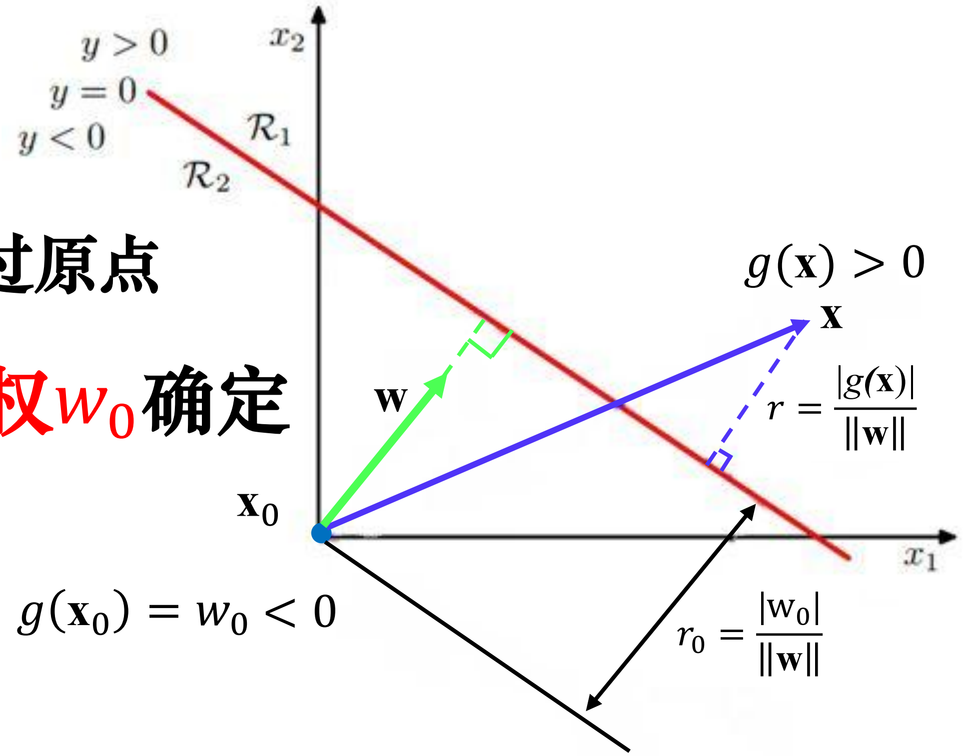
  如果$w_0 = 0$, 则$g(\mathbf{x})$具有齐次形式，超平面H通过原点

- 超平面方向由权向量w决定；位置由阈值权w_0确定

  判别函数$g(\mathbf{x})$正比于x点到超平面的代数距离

  当x在H的正侧时，$g(\mathbf{x}) > 0$

  当x在H的负侧时，$g(\mathbf{x}) < 0$



广义线性判别函数

- **两类别问题, X是一维样本空间**

不适用于非凸和多连通区域划分

若 $x < a$或 $x > b$，$x \in w_1$；若 $a < x < b$，$x \in w_2$

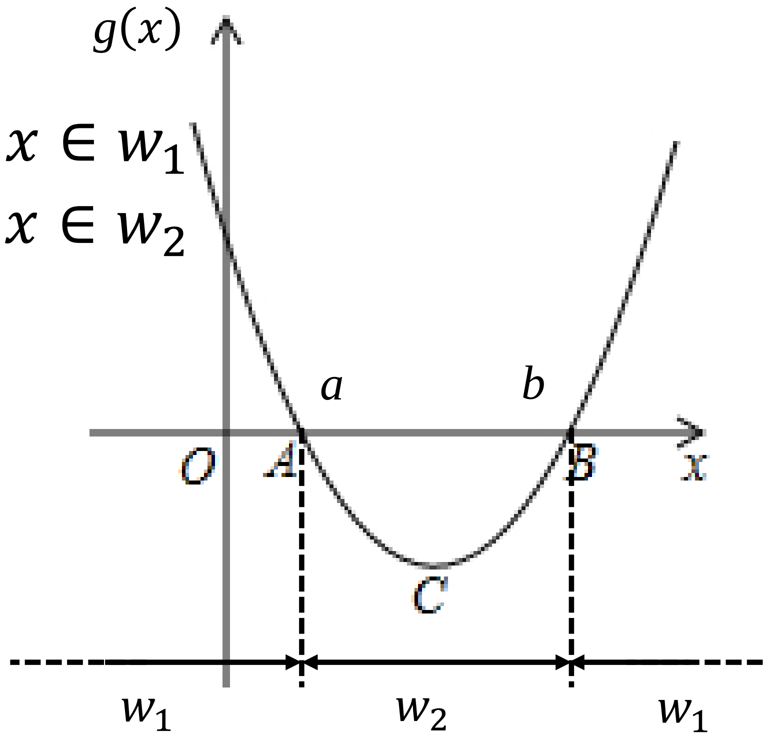
如果建立 $g(x) = (x-a)(x-b)$ 决策规则 $\begin{cases} g(x) > 0, x \in w_1 \\ g(x) < 0, x \in w_2 \end{cases}$

一般形式 $g(x) = c_0 + c_1 x + c_2 x^2$，可转化为

$$g(x) = \mathbf{a}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{3} a_i y_i \quad \text{其中}\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}=\begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \ \mathbf{a}=\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}=\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

称$g(x) = \mathbf{a}^{\mathrm{T}}\mathbf{y}$为广义线性判别函数，$\mathbf{a}$为广义权向量

当$g(x)$取二次形式时，也称为二次判别分析(QDA)*



利用线性函数的简单性解决复杂问题

维数大大增加→“维数灾难”

*更多内容详见Christopher M. Bishop, Pattern Recognition and Machine Learning, P199-200

线性判别函数的齐次化

- 线性判别函数 $g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$

改写成 $g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i}^{d} a_i y_i = \mathbf{a}^{\mathrm{T}}\mathbf{y}$ 称为线性判别函数的齐次简化

其中$\mathbf{y}=\begin{bmatrix}1\\x_1\\x_2\\\vdots\\x_d\end{bmatrix}=\begin{bmatrix}1\\\mathbf{x}\end{bmatrix}$称为增广样本向量；$\mathbf{a}=\begin{bmatrix}w_0\\w_1\\w_2\\\vdots\\w_d\end{bmatrix}=\begin{bmatrix}w_0\\\mathbf{w}\end{bmatrix}$称为增广权向量

$\hat{d} = d + 1$；y比x增加一维，保持样本空间欧式距离不变，变换后的样本仍全部位于d维子空间(原X空间)中

方程 $\mathbf{a}^{\mathrm{T}}\mathbf{y} = 0$ 在Y空间确定了一个通过原点的超平面$\widehat{H}$，它对d维子空间的划分与原决策面 $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = 0$ 对原X空间的划分完全相同。Y空间中任意一点y到$\widehat{H}$的距离：

$$r = \frac{g(\mathbf{x})}{\|\mathbf{a}\|} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{a}\|}$$

线性分类器设计的一般过程

- 利用训练样本建立线性判别函数

$$g(\mathbf{x})=\boxed{\mathbf{w}^{\mathrm{T}}}\mathbf{x}+\boxed{w_0}$$

$$g(\mathbf{x})=w_0+\sum_{i=1}^{d}w_i x_i=\sum_{i}^{d}a_i y_i=\boxed{\mathbf{a}^{\mathrm{T}}}\mathbf{y}$$

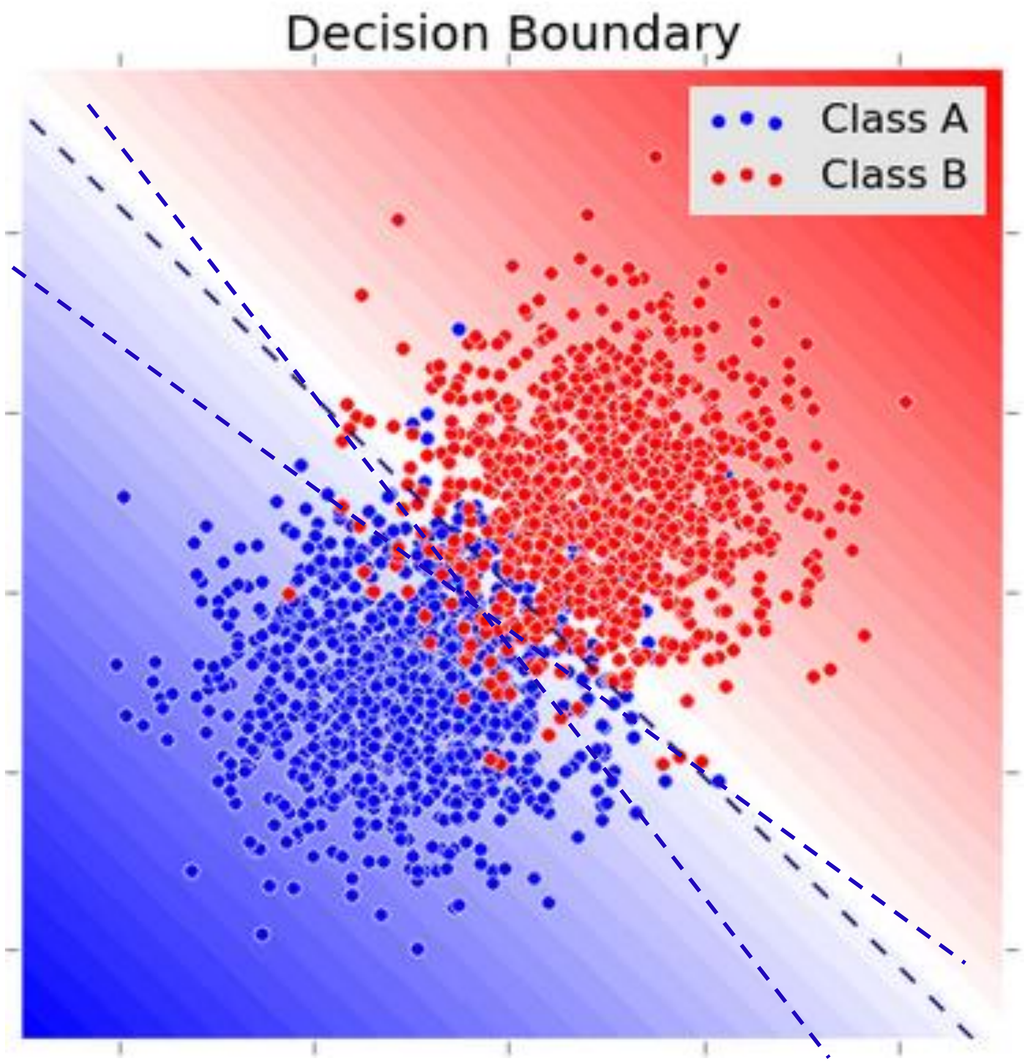
最好的结果一般出现在准则函数的极值点上，所以**将分类器设计问题转化为求准则函数极值$\mathbf{w}^*, w_0^*$ 或$\mathbf{a}^*$的问题。**

- **步骤1：**具有类别标志的样本集$\mathcal{X}=\{\mathrm{x}_1,\mathrm{x}_2,\ldots\mathrm{x}_\mathrm{N}\}$或其增广样本集$\mathcal{Y}$
- **步骤2：**确定准则函数$\mathcal{J}$，满足①$\mathcal{J}$是样本集和$\mathbf{w}, w_0$或$\mathbf{a}$的函数；②$\mathcal{J}$的值反映分类器的性能，其极值对应“最好”的决策
- **步骤3：**优化求解准则函数极值$\mathbf{w}^*, w_0^*$ 或$\mathbf{a}^*$

最终得到线性判别函数：$g(\mathbf{x})=\mathbf{w}^{*\mathrm{T}}\mathbf{x}+w_0$ 或$g(\mathbf{x})=\mathbf{a}^{*\mathrm{T}}\mathbf{y}$，对于未知类别样本$\mathbf{x}_\mathrm{k}$，计算$g(\mathbf{x}_\mathrm{k})$并通过决策规则判断其类别

准则函数的设计

- Fisher准则
- 感知机准则
- 最小二乘准则



3.4 Fisher准则

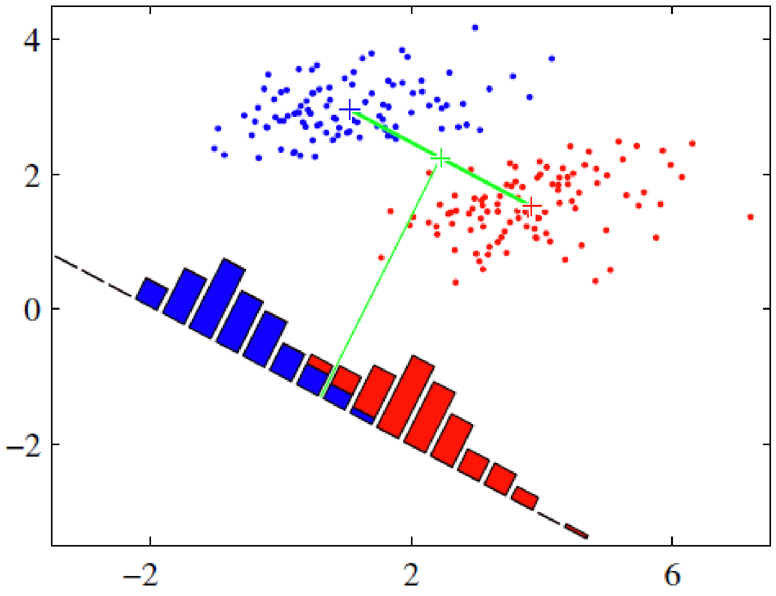
- Fisher准则的概念
- Fisher准则的求解

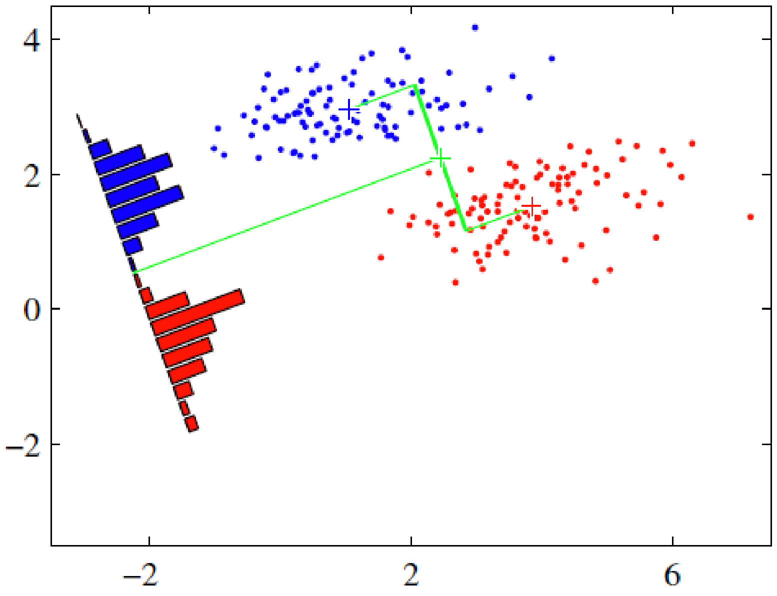
Fisher准则的概念

- 考虑把 d 维空间的样本投影到一条直线上形成一维空间。在一般情况下总可以找到某个方向，使样本在这个方向的直线上的投影分开得最好

【R. A. Fisher, 1936】

Fisher准则就是要解决
如何根据实际情况找到这条最好的、最易于分类的投影方向的问题





Fisher准则的推导

- **寻找最好投影方向$\mathbf{w}^*$**
  - 以二分类问题为例，d维样本$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$, 其中$N_1$个属于$w_1$类记为子集$X_1$, N_2个属于w_2类记为子集X_2
  - 在d维X空间

各类样本的均值向量$\mathbf{m}_\mathrm{i}$

$$\mathbf{m}_\mathrm{i} = \frac{1}{N_i}\sum_{\mathbf{x}\in\mathcal{X}_i}\mathbf{x}, i = 1,2$$

样本类内离散度矩阵$\mathbf{S}_\mathrm{i}$和总类内离散度矩阵$\mathbf{S}_\mathrm{W}$

$$\mathbf{S}_\mathrm{i} = \sum_{\mathbf{x}\in\mathcal{X}_i}(\mathbf{x}-\mathbf{m}_\mathrm{i})(\mathbf{x}-\mathbf{m}_\mathrm{i})^\mathrm{T}, i = 1,2$$

$$\mathbf{S}_\mathrm{W} = \mathbf{S}_1 + \mathbf{S}_2$$

$$\mathbf{S}_\mathrm{W} = P(w_1)\mathbf{S}_1 + P(w_1)\mathbf{S}_2$$

样本类间离散度矩阵$\mathbf{S}_\mathrm{b}$

$$\mathbf{S}_\mathrm{b} = (\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^\mathrm{T}$$

$$\mathbf{S}_\mathrm{b} = P(w_1)P(w_2)(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^\mathrm{T}$$

Fisher准则

- 寻找最好投影方向$\mathbf{w}^*$
  - 以二分类问题为例，d维样本$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$, 其中$N_1$个属于$w_1$ 类记为子集X_1, N_2个属于w_2类记为子集X_2
  - 在一维Y空间$\mathbf{y}_\mathrm{n} = \mathbf{w}^\mathrm{T}\mathbf{x}_\mathrm{n}$

    各类样本均值$\widetilde{\mathbf{m}}_\mathrm{i}$

$$\widetilde{\mathbf{m}}_\mathrm{i} = \frac{1}{N_i}\sum_{\mathbf{y}\in\boldsymbol{\eta}_\mathrm{i}} \mathbf{y}, i = 1,2$$

    样本类内离散度$\tilde{\mathbf{S}}_\mathrm{i}^2$ 和总类内离散度$\tilde{\mathbf{S}}_\mathrm{W}$

$$\tilde{\mathbf{S}}_\mathrm{i}^2 = \sum_{\mathbf{y}\in\boldsymbol{\eta}_\mathrm{i}} (\mathbf{y} - \widetilde{\mathbf{m}}_\mathrm{i})^2, i = 1,2$$

$$\tilde{\mathbf{S}}_\mathrm{W} = \tilde{\mathbf{S}}_1^2 + \tilde{\mathbf{S}}_2^2$$

Fisher准则

- 希望投影后在一维*Y*空间中各类样本尽可能分开，即两类均值之差越大越好
- 同时希望各类样本内部尽量密集，即类内离散度越小越好

➢ **Fisher准则定义为类间方差与类内方差之比：**

$$J_F(\mathbf{w}) = \frac{\boxed{(\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2)^2}}{\boxed{\tilde{\mathbf{S}}_1^2 + \tilde{\mathbf{S}}_2^2}}$$

两类均值之差越大越好

类内离散度越小越好

Fisher准则

- **求$J_F(\mathbf{w})$取得最大值的w** $J_F(\mathbf{w})=\frac{(\tilde{\mathrm{m}}_1-\tilde{\mathrm{m}}_2)^2}{\tilde{\mathrm{S}}_1^2+\tilde{\mathrm{S}}_2^2}$

$$\tilde{\mathrm{m}}_\mathrm{i}=\frac{1}{N_i}\sum_{\mathrm{y}\in\eta_i}\mathrm{y}=\frac{1}{N_i}\sum_{\mathrm{x}\in\mathcal{X}_\mathrm{i}}\mathbf{w}^\mathrm{T}\mathbf{x}=\mathbf{w}^\mathrm{T}\left(\frac{1}{N_i}\sum_{\mathrm{x}\in\mathcal{X}_\mathrm{i}}\mathbf{x}\right)=\mathbf{w}^\mathrm{T}\mathbf{m}_\mathrm{i}$$

$$(\tilde{\mathrm{m}}_1-\tilde{\mathrm{m}}_2)^2=(\mathbf{w}^\mathrm{T}\mathbf{m}_1-\mathbf{w}^\mathrm{T}\mathbf{m}_2)^2=\mathbf{w}^\mathrm{T}(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^\mathrm{T}\mathbf{w}=\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{b}\mathbf{w}$$

$$\tilde{\mathrm{S}}_\mathrm{i}^2=\sum_{\mathrm{y}\in\eta_\mathrm{i}}(\mathrm{y}-\tilde{\mathrm{m}}_\mathrm{i})^2=\sum_{\mathrm{x}\in\mathcal{X}_\mathrm{i}}\mathbf{w}^\mathrm{T}\mathbf{x}-\mathbf{w}^\mathrm{T}\mathbf{m}_\mathrm{i})^2=\mathbf{w}^\mathrm{T}\sum_{\mathrm{x}\in\mathcal{X}_\mathrm{i}}(\mathbf{x}-\mathbf{m}_\mathrm{i})(\mathbf{x}-\mathbf{m}_\mathrm{i})^\mathrm{T}\mathbf{w}=\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{i}\mathbf{w}$$

$$\tilde{\mathrm{S}}_1^2+\tilde{\mathrm{S}}_2^2=\mathbf{w}^\mathrm{T}(\mathbf{S}_1+\mathbf{S}_2)\mathbf{w}=\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{W}\mathbf{w}\quad\Rightarrow\quad J_F(\mathbf{w})=\frac{\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{b}\mathbf{w}}{\mathbf{w}^\mathrm{T}\mathbf{S}_\mathrm{W}\mathbf{w}}$$

- **Lagrange乘子法求解：**

$$\mathbf{w}^*=\mathbf{S}_\mathrm{W}^{-1}(\mathbf{m}_1-\mathbf{m}_2)$$

- **分类时确定分界阈值y_0，与$y=\mathbf{w}^{*\mathrm{T}}\mathbf{x}$比较进行决策**

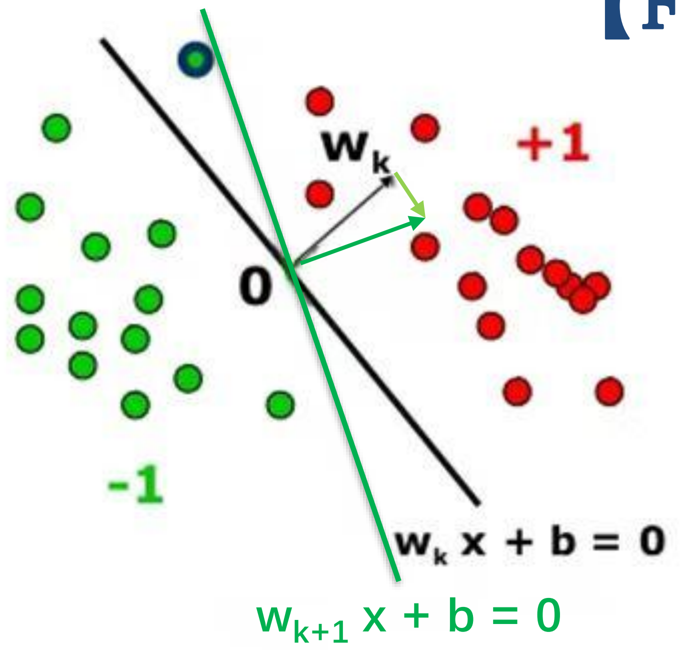
3.5 感知机准则

- 感知机准则的概念
- 感知机准则的求解

感知机准则

- 感知准则是一种自学习判别函数生成方法，由于 Rosenblatt 试图将其用于脑模型**感知机**，因此得名
- 该方法对随意给定的判别函数初始值，通过样本分类训练过程逐步对其修正直至最终确定

【F. Rosenblatt, 1957】



几个基本概念

- **线性可分性**

  - 一组容量为N的样本集$\mathbf{y}_1, \mathbf{y}_2, \dots \mathbf{y}_N$，其中$\mathbf{y}_n$为$\hat{d}$维增广样本向量，分别来自$w_1$类和$w_2$类，如果存在权向量$\mathbf{a}$，使得对于任何$\mathbf{y} \in w_1$，都有$\mathbf{a}^{\mathrm{T}}\mathbf{y} > 0$，而对于任何$\mathbf{y} \in w_2$，都有$\mathbf{a}^{\mathrm{T}}\mathbf{y} < 0$，则称这组样本为线性可分的，反之亦然

- **样本的规范化**

$$\begin{cases} \mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{i}} > 0, \mathbf{y}_{\mathrm{i}} \in w_1 \\ \boxed{\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{j}} < 0, \mathbf{y}_{\mathrm{j}} \in w_2} \end{cases}$$

$$\Downarrow$$

$$\mathbf{y}_{\mathrm{n}}' = \begin{cases} \mathbf{y}_{\mathrm{i}} & , \mathbf{y}_{\mathrm{i}} \in w_1 \\ -\mathbf{y}_{\mathrm{j}} & , \mathbf{y}_{\mathrm{j}} \in w_2 \end{cases}$$

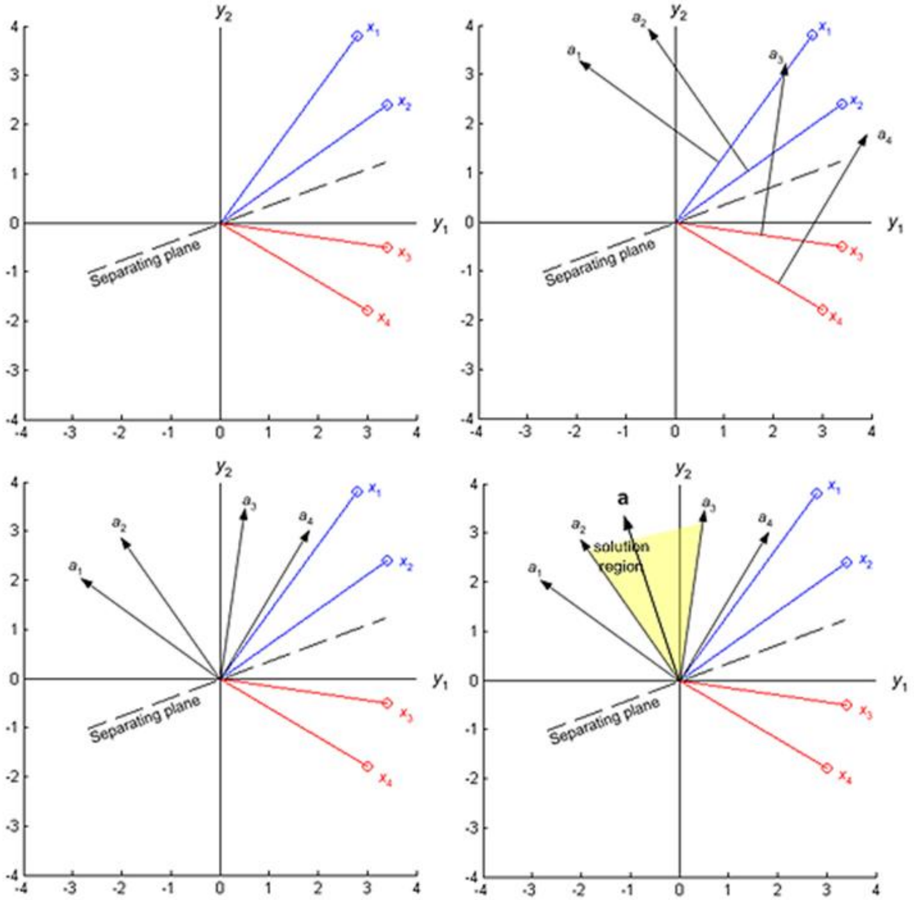
规范化增广样本向量

$$\Downarrow$$

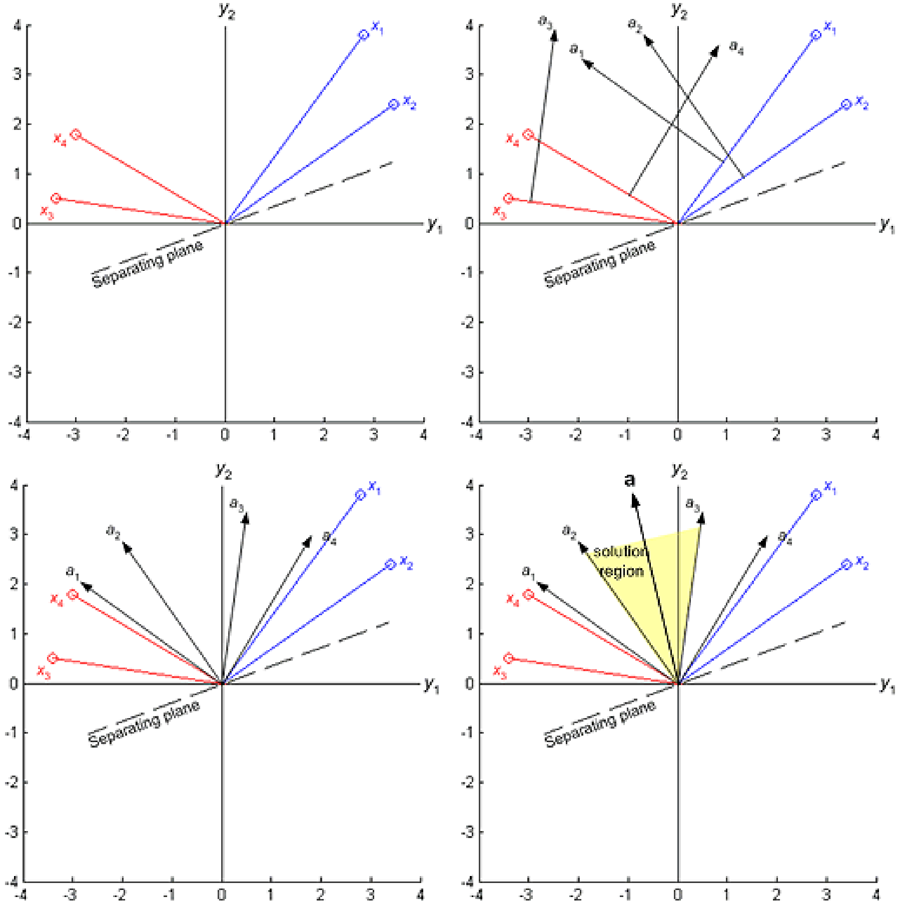
$$\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{i}}' > 0, n = 1,2, \dots, N$$

几个基本概念

- **解向量和解区**



未规范化

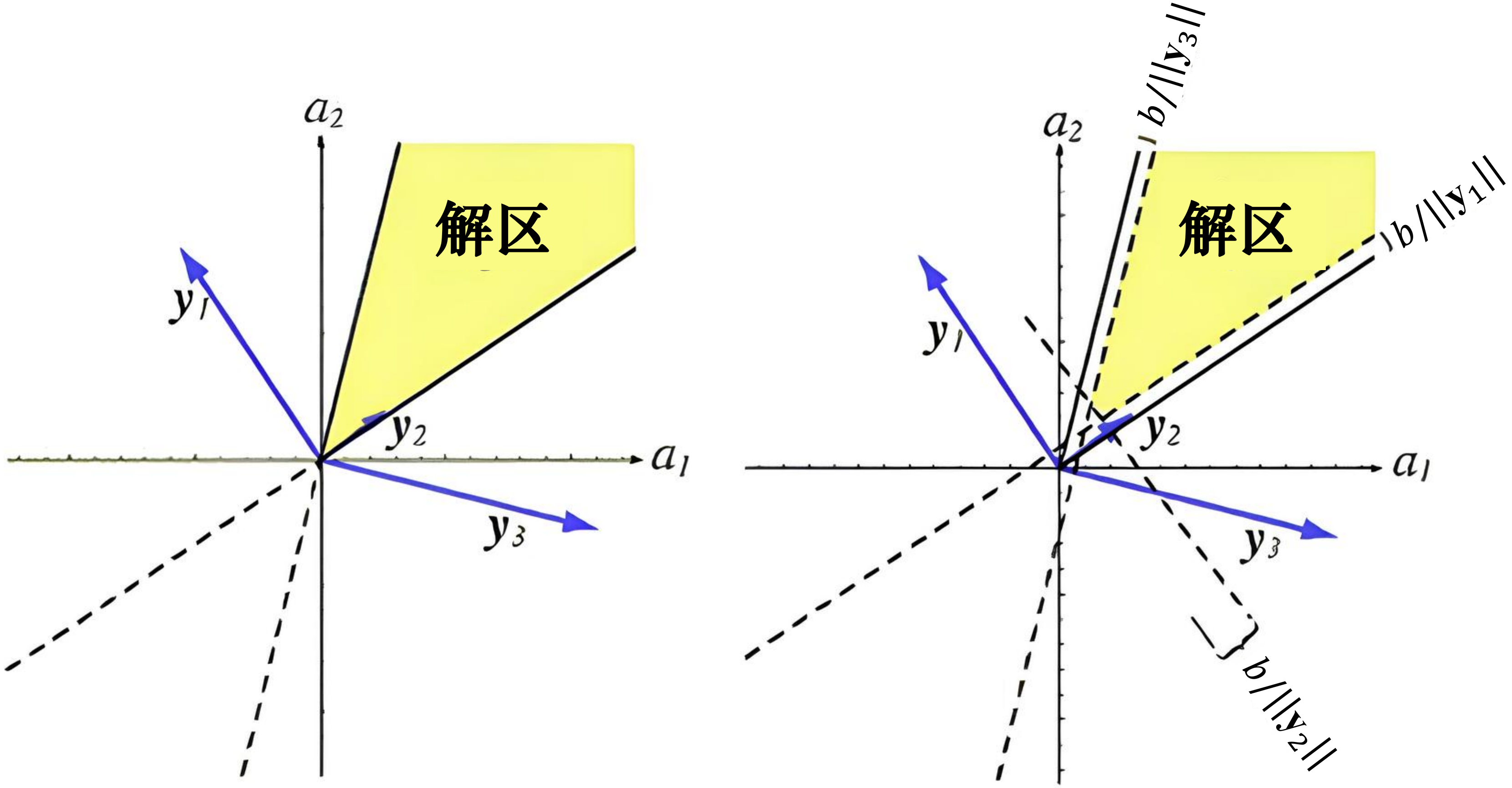


规范化

规范化以后，解向量均满足$a^T y_n > 0$，解区并没有变化

几个基本概念

- **对解区的限制**
  - **使解向量更可靠$\mathbf{a}^T\mathbf{y}_n \geq b > 0$，避免解收敛到解区边界的某点上**



感知机准则

- **寻找解向量$\mathbf{a}^*$**
  - 对于一组样本$\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_N$, 其中$\mathbf{y}_n$是规范化增广样本向量, 使得:

$$\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} > 0, n = 1,2, \ldots, N$$

  - 对于线性可分问题，构造准则函数 $J_P(\mathbf{a}) = \sum_{y \in \eta^k} (-\mathbf{a}^{\mathrm{T}}\mathbf{y})$

    η^k是被权向量a错分的样本集合，即当$\mathbf{y}$被错分时，就有 $\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} \leq 0$

    因此$J_P(\mathbf{a}) \geq 0$，仅当$\mathbf{a}$为解向量或在解区边界时$J_P(\mathbf{a}) = 0$

    也就是说，当且仅当η_k为空集时$J_P^*(\mathbf{a}) = \min J_P(\mathbf{a}) = 0$

    此时无错分样本，这时的$\mathbf{a}$就是解向量$\mathbf{a}^*$

感知机准则函数的求解

- 求使$J_P(\mathbf{a})$达到最小值的$\mathbf{a}^*$
  - 采用梯度下降法求解

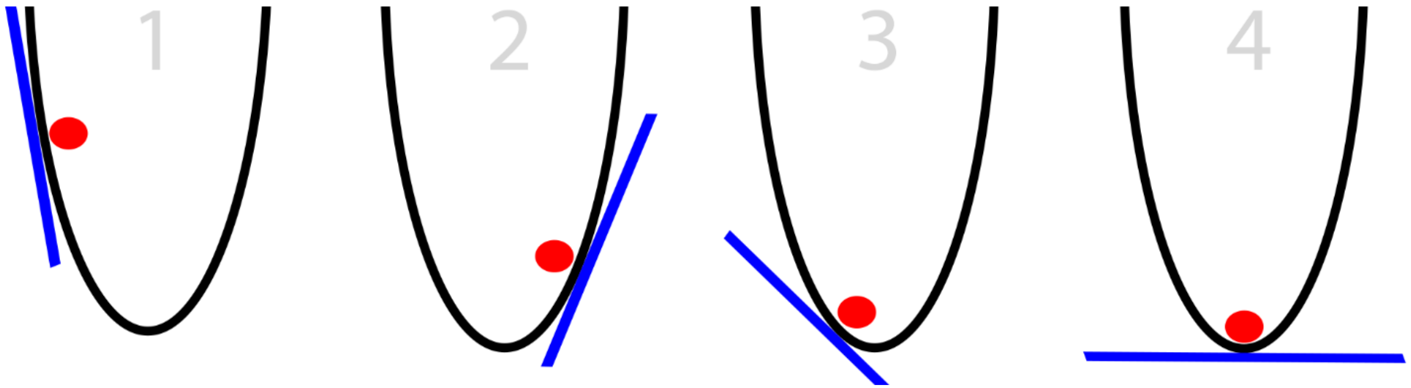
$$J_P(\mathbf{a}) = \sum_{\mathbf{y}\in\eta_k} (-\mathbf{a}^{\mathrm{T}}\mathbf{y})$$

$$\nabla J_P(\mathbf{a}) = \frac{\partial J_P(\mathbf{a})}{\partial \mathbf{a}} = \sum_{\mathbf{y}\in\eta_k} (-\mathbf{y})$$

$$\mathbf{a}(k+1) = \mathbf{a}(k) - \rho_k \nabla J$$

梯度下降法迭代公式

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \rho_k \sum_{\mathbf{y}\in\eta_k} \mathbf{y}$$



感知机准则-示例

- 二分情况下的样本集

➢ 类别1：

$$\mathbf{x}_1 = \begin{pmatrix} -2 \\ 2 \end{pmatrix},\ \mathbf{x}_2 = \begin{pmatrix} -2 \\ -2 \end{pmatrix} \qquad \mathbf{y}_1 = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}$$

➢ 类别2：

$$\mathbf{x}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix},\ \mathbf{x}_4 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \qquad \mathbf{y}_3 = \begin{pmatrix} -1 \\ -2 \\ -1 \end{pmatrix},\ \mathbf{y}_4 = \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}$$

➢ 初始化权重向量：

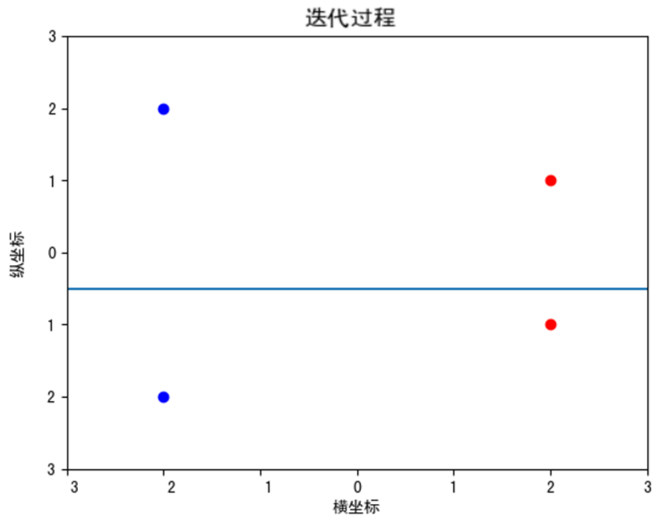
$$\mathbf{a}(1) = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} \qquad \mathbf{a}(1)^{\mathrm{T}} = \begin{pmatrix} 0 & 2 & 1 \end{pmatrix}$$

感知机准则-示例

- 迭代过程
  - (1) $\mathbf{y}^{(1)\mathrm{T}} = (1 \quad -2 \quad 2)$

$$\mathbf{a}(1)^{\mathrm{T}}\mathbf{y}^{(1)} = (0 \quad 2 \quad 1)\begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} = -2 < 0$$

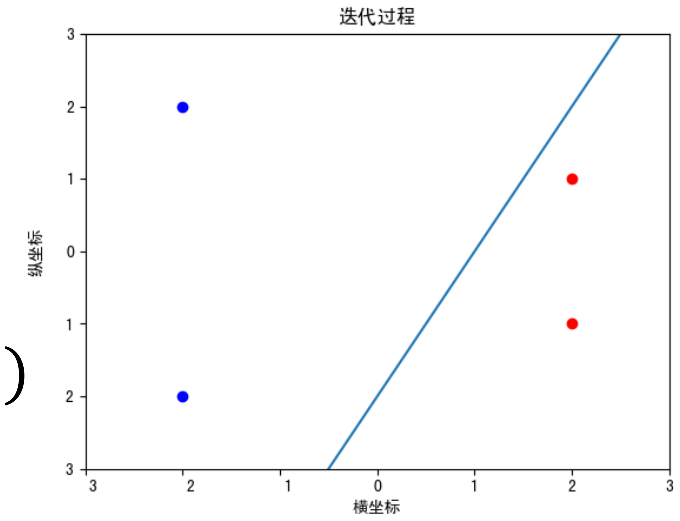
$$\mathbf{a}(2)^{\mathrm{T}} = (0 \quad 2 \quad 1) + (1 \quad -2 \quad 2) = (1 \quad 0 \quad 3)$$



  - (2) $\mathbf{y}^{(2)\mathrm{T}} = (1 \quad -2 \quad -2)$

$$\mathbf{a}(2)^{\mathrm{T}}\mathbf{y}^{(2)} = (1 \quad 0 \quad 3)\begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix} = -5 < 0$$

$$\mathbf{a}(3)^{\mathrm{T}} = (1 \quad 0 \quad 3) + (1 \quad -2 \quad -2) = (2 \quad -2 \quad 1)$$

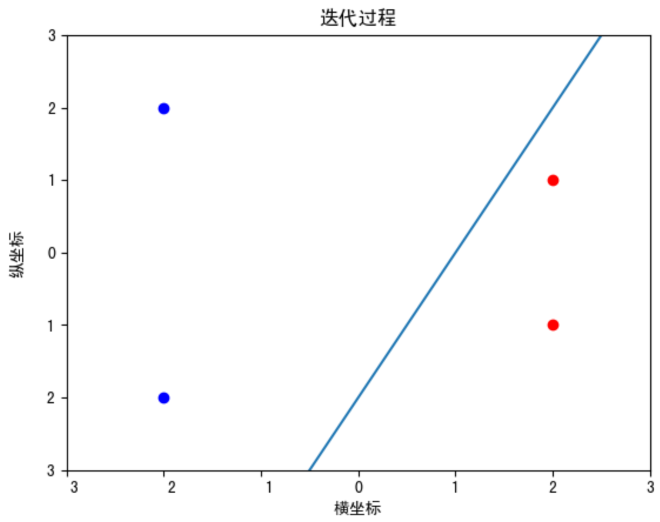


感知机准则-示例

- **迭代过程**
  - **(3)** $\mathbf{y}^{(3)\mathrm{T}} = (-1 \quad -2 \quad -1)$

$$\mathbf{a}(3)^{\mathrm{T}}\mathbf{y}^{(3)} = (2 \quad -2 \quad 1)\begin{pmatrix}-1\\-2\\-1\end{pmatrix} = 1 > 0$$

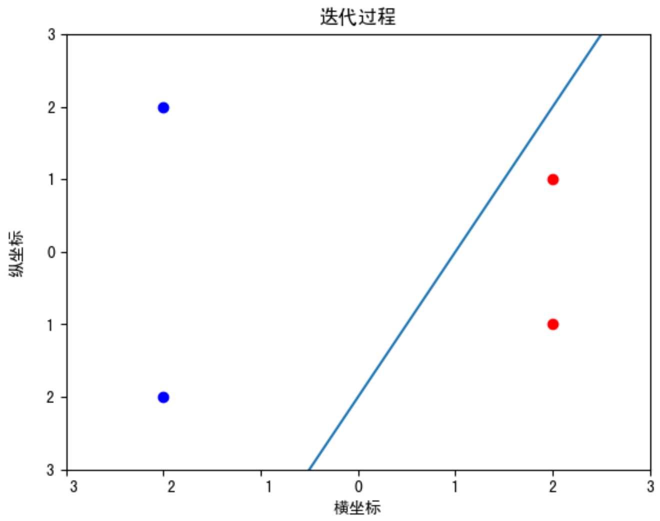
$$\mathbf{a}(4)^{\mathrm{T}} = (2 \quad -2 \quad 1) \quad \text{（没有变化）}$$



  - **(4)** $\mathbf{y}^{(4)\mathrm{T}} = (2 \quad -2 \quad 1)$

$$\mathbf{a}(4)^{\mathrm{T}}\mathbf{y}^{(4)} = (2 \quad -2 \quad 1)\begin{pmatrix}-1\\-2\\1\end{pmatrix} = 3 > 0$$

$$\mathbf{a}(5)^{\mathrm{T}} = (2 \quad -2 \quad 1) \quad \text{（没有变化）}$$



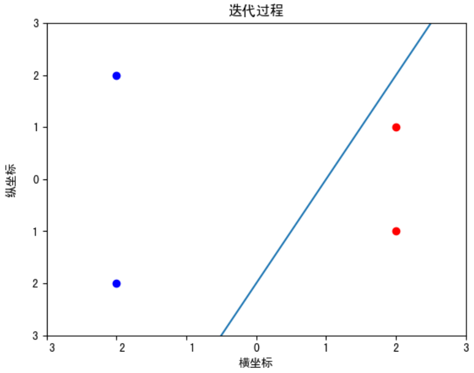
感知机准则-示例

迭代过程

- **(5)** $\mathbf{y}^{(5)\mathrm{T}} = (2 \quad -2 \quad 1)$

  $$\mathbf{a}(5)^{\mathrm{T}}\mathbf{y}^{(5)} = (2 \quad -2 \quad 1)\begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} = 8 > 0$$

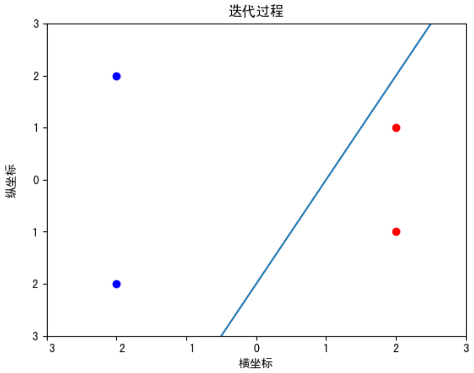
  $\mathbf{a}(6)^{\mathrm{T}} = (2 \quad -2 \quad 1)$ （没有变化）



- **(6)** $\mathbf{y}^{(6)\mathrm{T}} = (1 \quad -2 \quad 2)$

  $$\mathbf{a}(6)^{\mathrm{T}}\mathbf{y}^{(6)} = (2 \quad -2 \quad 1)\begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix} = 4 > 0$$

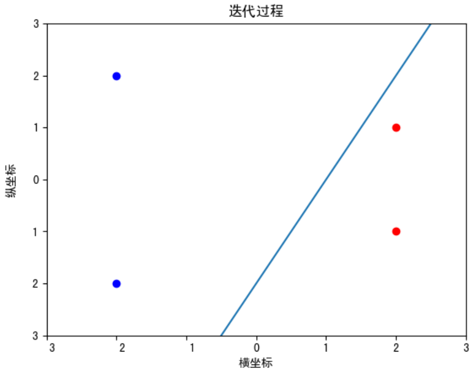
  $\mathbf{a}(7)^{\mathrm{T}} = (2 \quad -2 \quad 1)$ （没有变化）



- **(7)** $\mathbf{y}^{(7)\mathrm{T}} = (-1 \quad -2 \quad -1)$

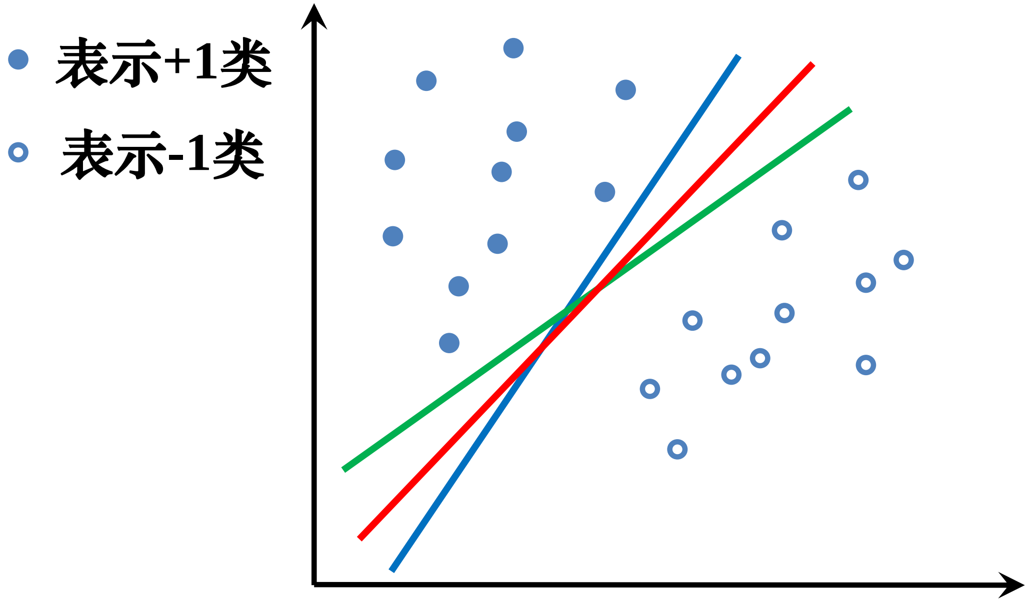
  $$\mathbf{a}(7)^{\mathrm{T}}\mathbf{y}^{(7)} = (2 \quad -2 \quad 1)\begin{pmatrix} -1 \\ -2 \\ -1 \end{pmatrix} = 1 > 0$$

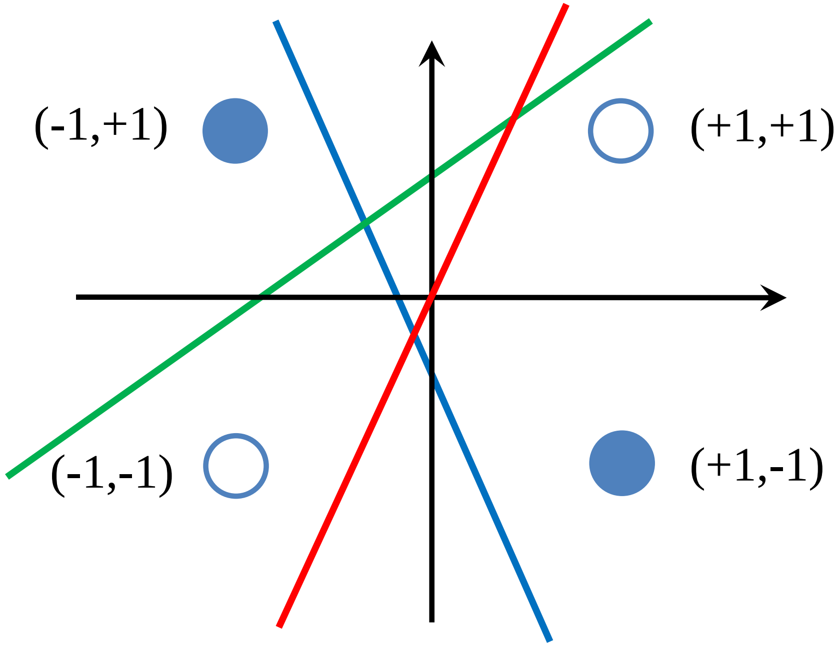
  $\mathbf{a}(8)^{\mathrm{T}} = (2 \quad -2 \quad 1)$ （没有变化）



感知机准则

- 在线性可分情形下，感知机准则一定收敛，但收敛结果不唯一
- 在线性不可分情况下不收敛





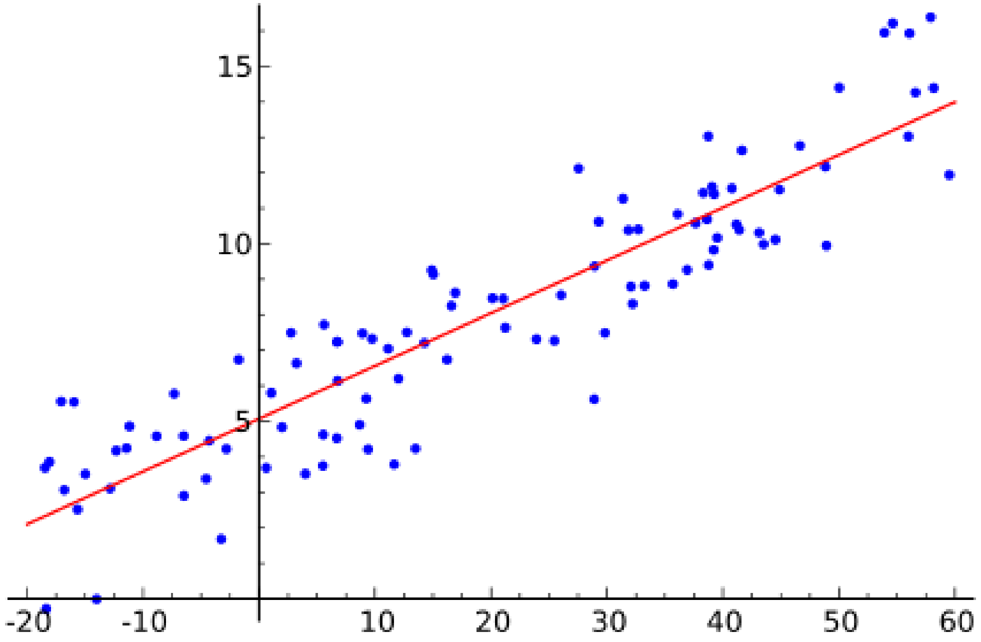
3.6 最小二乘准则

- 最小二乘准则的概念
- 最小二乘准则的求解

最小二乘准则

- 最小二乘法(最小平方误差法)通过最小化误差的平方和寻找数据的最佳函数匹配，即可以使求得的数据与实际数据之间误差的平方和最小

【A.-M. Legendre, 1806提出】
【C. Gauss, 1809提出, 1829证明】



最小二乘准则

- 目标：寻找最优投影方向$\mathbf{a}^*$

$$\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} > 0 \longrightarrow \mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} = b_n > 0$$

b_n是任意给定的正常数

➢ 方程组形式： $\mathbf{Ya} = \mathbf{b}$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^{\mathrm{T}} \\ \mathbf{y}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_{\mathrm{N}}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1\hat{d}} \\ y_{21} & y_{22} & \cdots & y_{2\hat{d}} \\ \cdots & \cdots & \cdots & \cdots \\ y_{N1} & y_{N2} & \cdots & y_{N\hat{d}} \end{bmatrix} \qquad \mathbf{b} = [b_1 \quad b_2 \quad \cdots \quad b_N]$$

$\mathbf{y}^{\mathrm{n}}$是规范化增广向量样本；$\mathbf{Y}$是$N \times \hat{d}$维矩阵，通常$N > \hat{d}$，一般为列满秩阵

$\mathbf{b}$是N维向量，$b_n > 0$， n=1, 2, …, N

➢ 方程数多于未知数的矛盾方程组通常没有精确解

➢ 定义误差向量：$e = \mathbf{Ya} - \mathbf{b}$以及平方误差准则函数

$$J_s(\mathbf{a}) = \| \mathbf{e} \|^2 = \| \mathbf{Ya} - \mathbf{b} \|^2 = \sum_{n-1}^{N} (\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} - \mathbf{b}_{\mathrm{n}})^2$$

最小二乘准则

- 求使 $J_s(\mathbf{a})$ 最小的$\mathbf{a}^*$ (最小二乘近似解/伪逆解/均方误差(MSE)解)

  - 采用解析法求伪逆解

$$J_s(\mathbf{a}) = \|\mathbf{e}\|^2 = \|\mathbf{Ya}-\mathbf{b}\|^2 = \sum_{n-1}^{N} (\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} - \mathbf{b}_{\mathrm{n}})^2$$

$$\nabla J_s(\mathbf{a}) = \sum_{n-1}^{N} 2(\mathbf{a}^{\mathrm{T}}\mathbf{y}_{\mathrm{n}} - \mathbf{b}_{\mathrm{n}})\mathbf{y}_{\mathrm{n}} = 2\mathbf{Y}^{\mathrm{T}}(\mathbf{Ya}-\mathbf{b})$$

令 $\nabla J_s(\mathbf{a}) = 0$ 得 $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\mathbf{a}^* = \mathbf{Y}^{\mathrm{T}}\mathbf{b}$

矩阵$\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$是$\hat{d} \times \hat{d}$方阵，一般非奇异

唯一解 $\mathbf{a}^* = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{b} = \mathbf{Y}^{+}\mathbf{b}$

其中$\hat{d} \times N$矩阵$\mathbf{Y}^{+} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}$是Y的左逆矩阵

  - 如何选b?

$$\mathbf{b} = \begin{array}{cl} \begin{bmatrix} N/N_1 \\ \cdots \\ N/N_1 \\ N/N_2 \\ \cdots \\ N/N_2 \end{bmatrix} & \begin{array}{l} N_1\text{个} \\ \\ \\ \\ N_2\text{个} \end{array} \end{array}$$

$\mathbf{a}^*$等价于Fisher解

$$g_0(x) = P(w_1|\mathbf{x}) - P(w_2|\mathbf{x})$$

$N \to \infty, \mathbf{b} = [1, 1, \cdots, 1]^{\mathrm{T}}$

以最小均方误差逼近贝叶斯判别函数

最小二乘准则

- 求使 $J_s(\mathbf{a})$ 最小的$\mathbf{a}^*$ (最小二乘近似解/伏逆解/均方误差(MSE)解)

$$\mathbf{a}^* = \mathbf{Y}^+\mathbf{b} \qquad \mathbf{Y}^+ = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$$

  - 问题：①要求$\mathbf{Y}^T\mathbf{Y}$ 非奇异；②求$\mathbf{Y}^+$ 计算量大同时可能引入较大误差
  - 采用梯度下降法求解

$$J_s(\mathbf{a}) = \|\mathbf{e}\|^2 = \|\mathbf{Ya}-\mathbf{b}\|^2 = \sum_{n-1}^{N}(\mathbf{a}^T\mathbf{y}_n - \mathbf{b}_n)^2$$

$$\nabla J_s(\mathbf{a}) = 2\mathbf{Y}^T(\mathbf{Ya}-\mathbf{b})$$

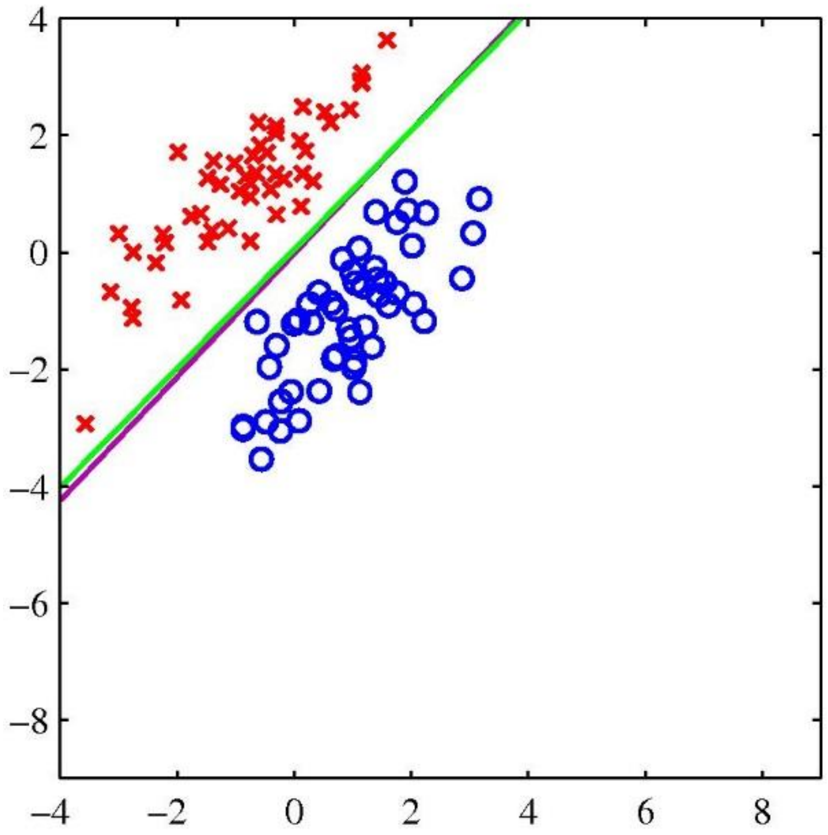
$$\begin{cases} \mathbf{a}(1), \text{随机值} \\ \mathbf{a}(k+1) = \mathbf{a}(k) - \rho_k \mathbf{Y}^T(\mathbf{Ya}-\mathbf{b}) \end{cases}$$

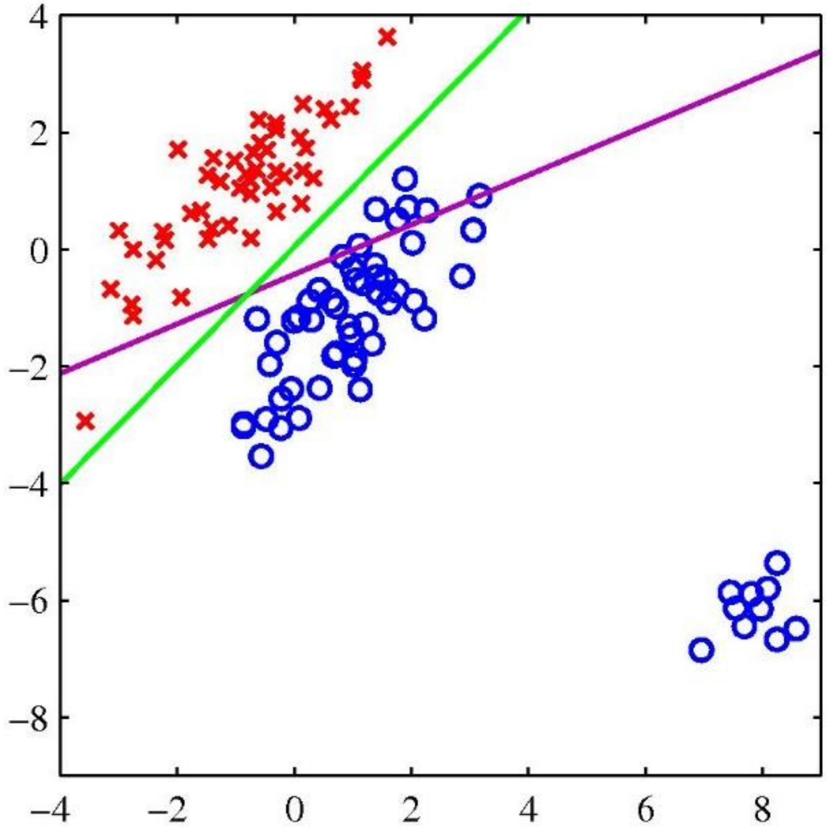
可以证明，选择$\rho_k = \frac{\rho_1}{k}$，$\rho_1$是任意常数，该算法权向量收敛于使$\nabla J_s(\mathbf{a}) = 2\mathbf{Y}^T(\mathbf{Ya}-\mathbf{b}) = 0$的权向量$\mathbf{a}^*$

  - 不要求$\mathbf{Y}^T\mathbf{Y}$ 奇异与否，只计算$\hat{d}\times\hat{d}$方阵$\mathbf{Y}^T\mathbf{Y}$ ，比$\hat{d}\times N$阵$\mathbf{Y}^+$计算量小

最小二乘准则的局限性

- 对于异常值(Outlier)非常敏感
  - ➢ 平方误差准则函数会惩罚那些“太正确”的预测，因为它们在决策边界的正确一侧距离太远





3.7 扩展与讨论

- 逻辑回归
- 二分类和多分类
- 生成式模型和判别式模型

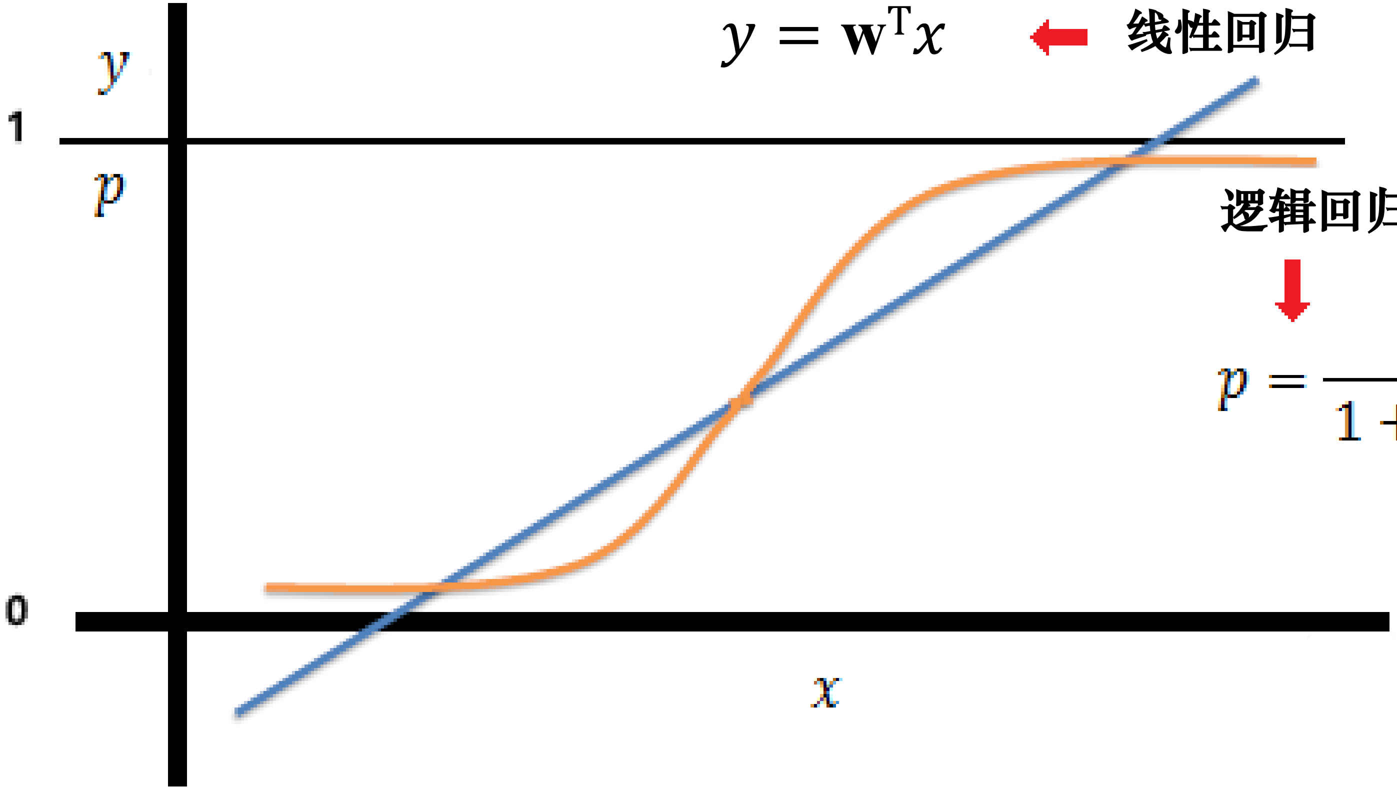
回归和分类

线性回归和逻辑回归

Sigmod函数

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

$y = \mathbf{w}^{\mathrm{T}}x$ ⬅ 线性回归



逻辑回归

⬇

$$p = \frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}}\boldsymbol{x}}}$$

逻辑回归 (Logistic Regression)

将回归问题转为分类问题

- 逻辑回归是**概率型非线性回归**，但其本质是线性回归，只是在特征到结果的映射中加入了一层函数映射，即先把特征线性求和，再使用sigmoid函数σ预测

- 将$p = \frac{1}{1+e^{-\mathrm{w}^{\mathrm{T}}x}}$转换形式为

对数几率（log odds，亦称logit）

$$\ln\frac{p}{1-p} = w^{\mathrm{T}}x$$

若将p视为样本x作为正例的可能性

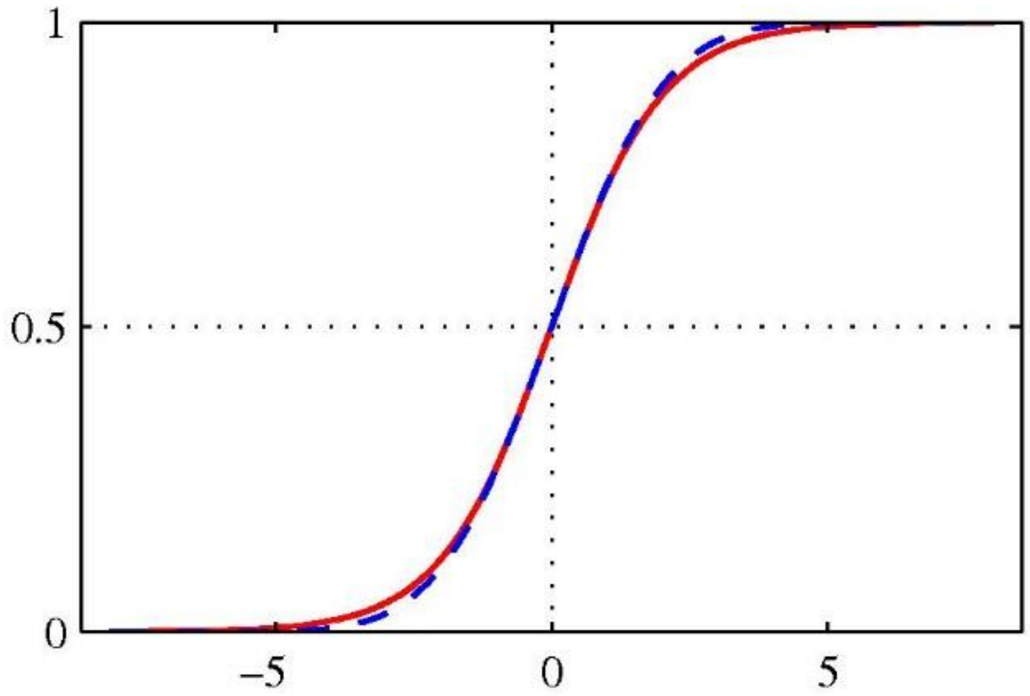
相对可能性，也称几率（odds）

- 实际上逻辑回归的表达式$p = \frac{1}{1+e^{-\mathrm{w}^{\mathrm{T}}x}}$是在**用线性回归模型的预测结果去逼近真实标记的对数几率**，故逻辑回归又称对数几率回归

逻辑回归 (Logistic Regression)

- **Sigmoid函数**
  - 形似“S”
  - 可以将连续值映射到0和1区间上
  - 对称性 $\sigma(-a) = 1 - \sigma(a)$
  - 反转性 $\boxed{a = \ln(\frac{\sigma}{1-\sigma})}$
  - 可导性 $\frac{d\sigma}{da} = \sigma(1-\sigma)$

Logit函数



逻辑回归 (Logistic Regression)

逻辑回归的求解

- 将后验概率表示为作用于变量x的线性函数的Logistic Sigmoid

$$p(C_1|x) = y(x) = \sigma(\mathbf{w}^{\mathrm{T}}x) \qquad p(C_2|x) = 1 - p(C_1|x)$$

- 数据集 $\{x_n, t_n\}, t_n \in \{0,1\}, n = 1,2, \dots, N$

- 似然函数 $p(t|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}, t = (t_1, \dots, t_N)^{\mathrm{T}}, y_n = p(C_1|x_n)$

- 误差函数 $E(\mathbf{w}) = -\ln p(t|\mathbf{w}) = \boxed{-\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}}$

$$y_n = \sigma(a_n), a_n = \mathbf{w}^{\mathrm{T}}x_n$$

$$\frac{d\sigma}{da} = \sigma(1-\sigma)$$

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)x_n$$

交叉熵(Cross-Entropy)误差函数

- 令导数等于零求解 $\nabla_{\mathbf{w}} \ln p(t|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}x_n\}x_n^{\mathrm{T}} = 0$

逻辑回归 (Logistic Regression)

- **将回归问题转为分类问题**

➢ **多类问题**

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln p(x|C_k)p(C_k)$$

归一化指数
Normalized Exponential

↓

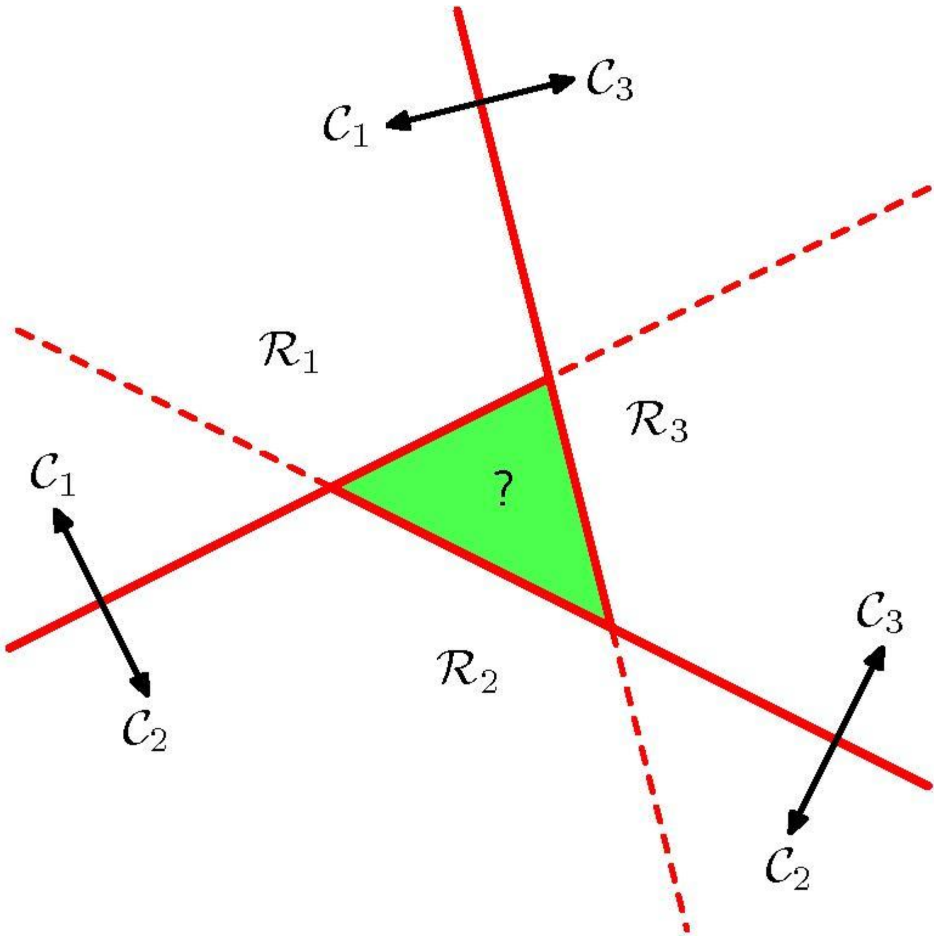
Softmax函数

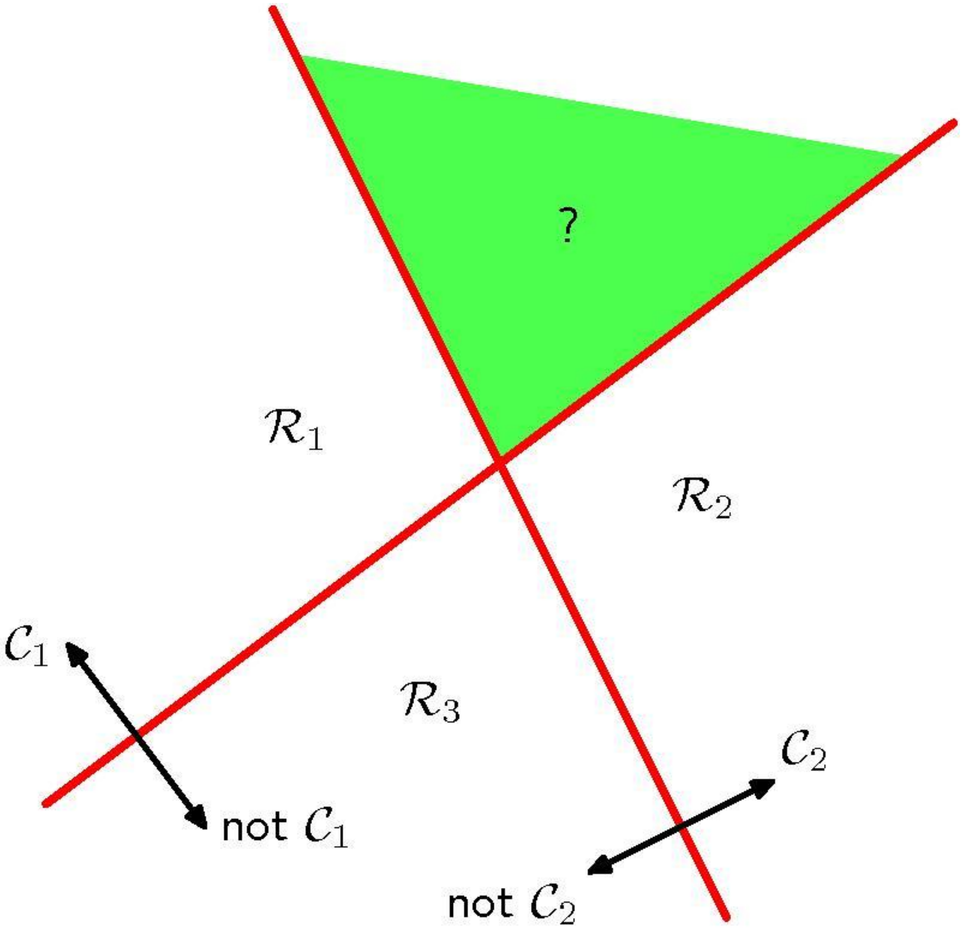
是一个平滑的max函数，对于所有的$k \neq j$**，如果**$a_k \gg a_j$**，有**$p(C_k|x) \simeq 1, p(C_j|x) \simeq 0$

Sigmoid函数可以看成是Softmax函数在$n = 2$**的一种特殊情形，前者经常用于二元回归/分类问题，而后者则可以应用于多元回归/分类**

二分类与多分类

- 多分类问题可以被分解为多个二分类问题进行求解，常用的分解策略有一对一(1 *vs.* 1)和一对其余(1 *vs.* (N-1))



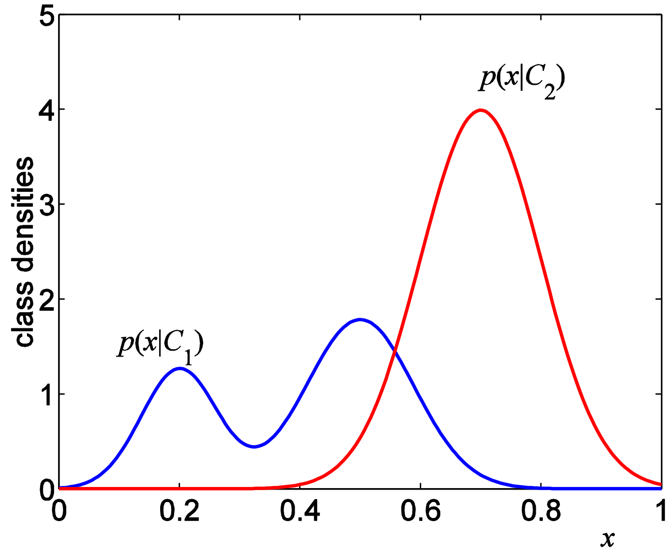


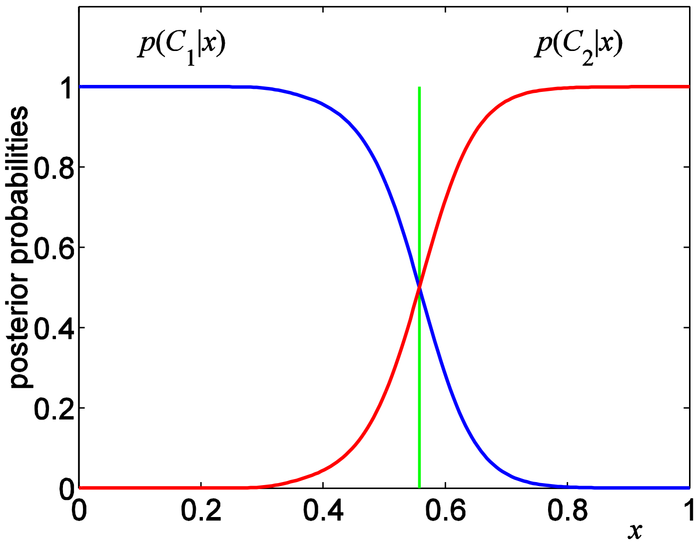
生成式模型和判别式模型

- **生成式模型 (Generative Model)**
  - 分别对各类的**类条件密度**$p(x|C_k)$和**先验概率**$p(C_k)$进行建模，之后利用贝叶斯定理计算**后验概率**
  - 或者直接对**联合分布**$p(x,C_k)$建模得到后验概率

$$p(C_k|x)=\frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)}$$

- **判别式模型 (Discriminative Model)**
  - 直接对**后验概率**$p(C_k|x)$建模



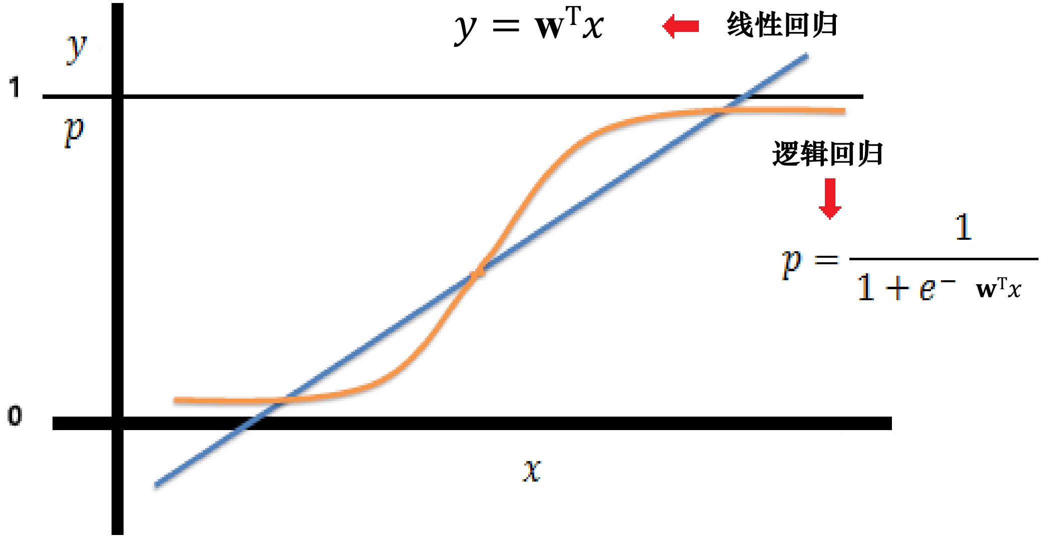


生成式模型和判别式模型

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

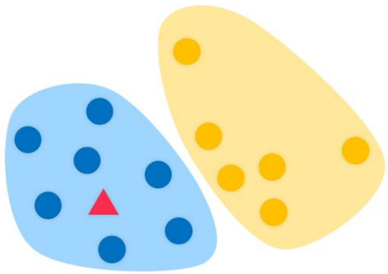
$$a = \ln\frac{p(C_1|x)}{p(C_2|x)} = \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$$

- **生成式模型**
  - 估计类条件密度并计算a
- **判别式模型**
  - 把a视为线性函数$a = \mathbf{w}^{\mathrm{T}}x + w_0$直接估计



生成式模型和判别式模型

● 生成式模型



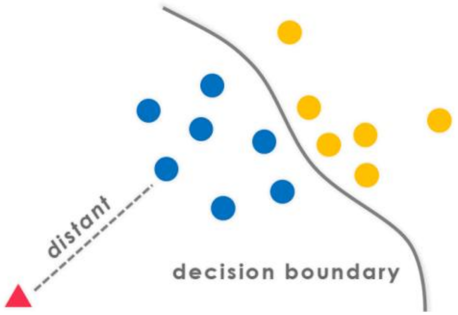
优点：

- ➢ 信息丰富
- ➢ 单类问题灵活性强
- ➢ 增量学习
- ➢ 合成缺失数据

缺点：

- ➢ 学习过程复杂
- ➢ 为分布牺牲分类性能

●判别式模型



优点：

- ➢ 类间差异清晰
- ➢ 分类边界灵活
- ➢ 学习简单
- ➢ 性能较好

缺点：

- ➢ 不能反映数据特性
- ➢ 需要全部数据进行学习

由生成模型可以得到判别模型
但由判别模型得不到生成模型

生成式模型和判别式模型

- **生成式模型代表算法：**
  - Naive Bayes
  - Mixtures of Gaussians
  - Hidden Markov Models
  - Bayesian Networks
  - Deep Belief Network

- **判别式模型代表算法：**
  - Linear & Logistic Regression
  - Support Vector Machine
  - Nearest Neighbor
  - Conditional Random Fields
  - Boosting