模式识别引论

An Introduction to Pattern Recognition

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

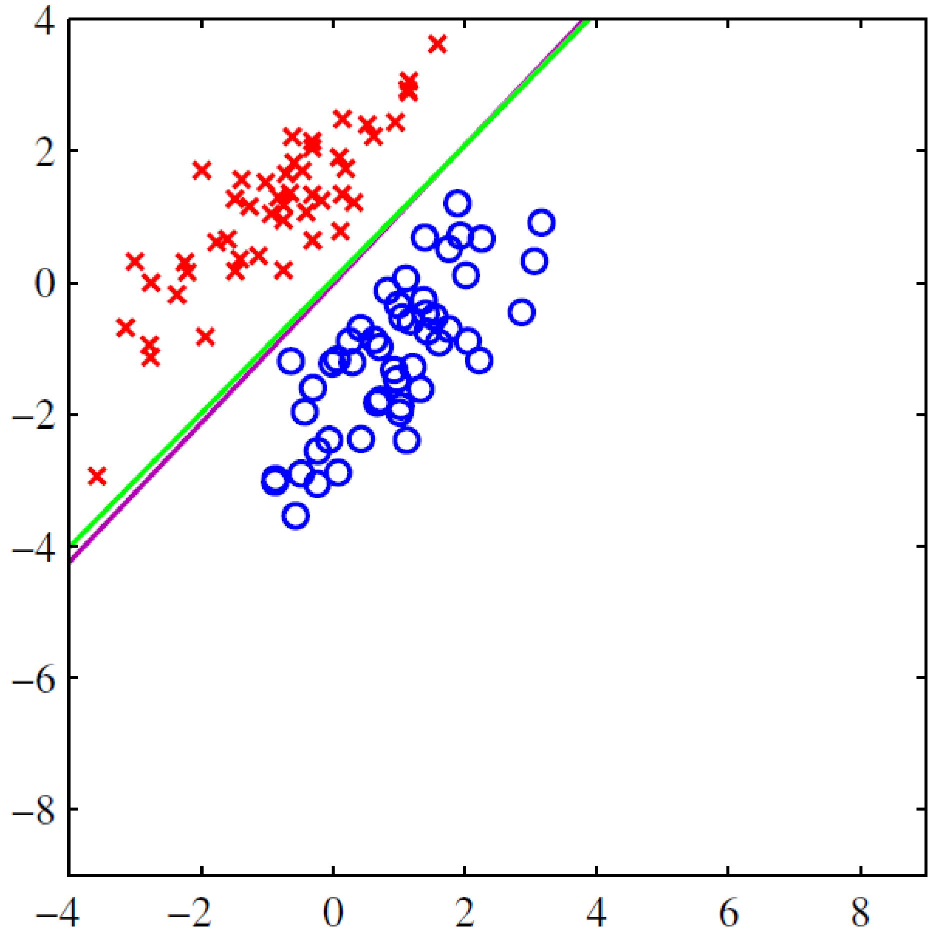
网络搜索教研中心 信息与通信工程学院 北京邮电大学

用于分类的线性模型 内容提要

- 引子: 2个类别的分类问题
- 鉴别函数
  - 最小二乘法用于分类
  - **Fisher**准则
  - 感知器法则
- 概率生成模型
  - 连续型输入数据
  - 离散型输入数据
- 概率鉴别模型
  - 逻辑斯蒂回归
  - 多类逻辑斯蒂回归
  - 贝叶斯逻辑斯蒂回归

2个类别的分类问题

- 考虑2个类别的分类问题
  - 如果在分类面的正侧，认为是类别1
  - 如果在分类面的负侧，认为是类别2
    - 分类面如何定义?



用于分类的线性模型 内容提要

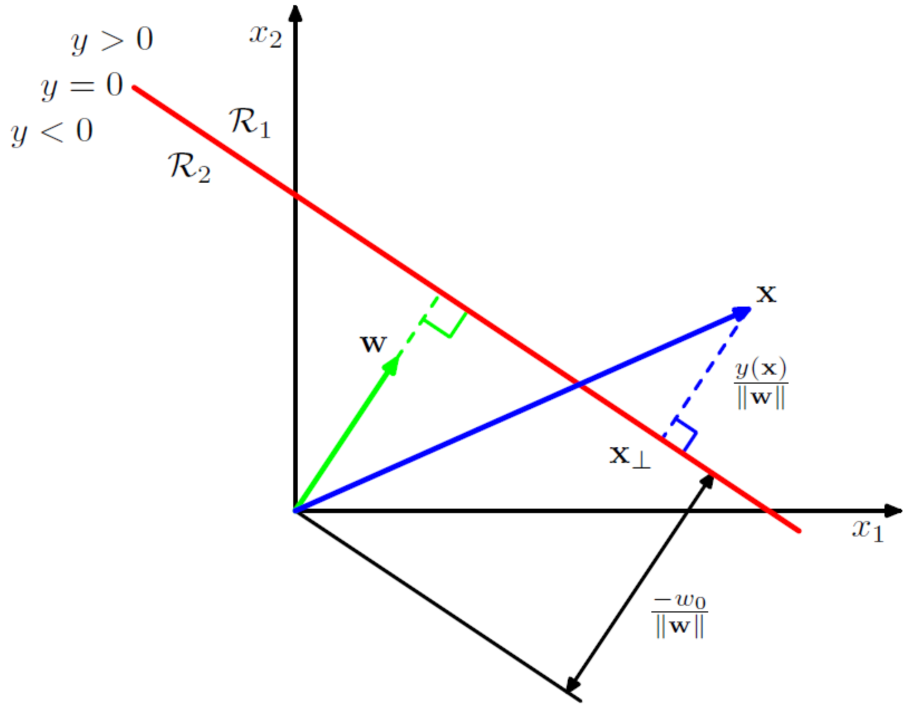
- 引子: 2个类别的分类问题
- 鉴别函数
  - 最小二乘法用于分类
  - Fisher准则
  - 感知器法则
- 概率生成模型
  - 连续型输入数据
  - 离散型输入数据
- 概率鉴别模型
  - 逻辑斯蒂回归
  - 多类逻辑斯蒂回归
  - 贝叶斯逻辑斯蒂回归

线性判别函数

- 定义线性判别函数为

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

- 几何意义:
  - 数据点x到分类超平面的代数距离
    - 提示: 借助点到超平面的距离来计算…
  - 如何估计w?



最小二乘法

- 给定训练数据集{x_n, t_n}，借助回归问题的方法

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}$$

其中 $\tilde{\mathbf{w}} = \begin{pmatrix} \mathbf{w} \\ w_0 \end{pmatrix}, \quad \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$

  – 写成矩阵向量形式 $y(\mathbf{x}) = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}$

  – 代价函数为 $E_D(\tilde{\mathbf{w}}) = \frac{1}{2}\mathrm{tr}\left\{\left(\tilde{X}^T\tilde{\mathbf{w}} - \mathbf{t}\right)\left(\tilde{X}^T\tilde{\mathbf{w}} - \mathbf{t}\right)^T\right\}$

  – 求导，令导数为**0**，则得到

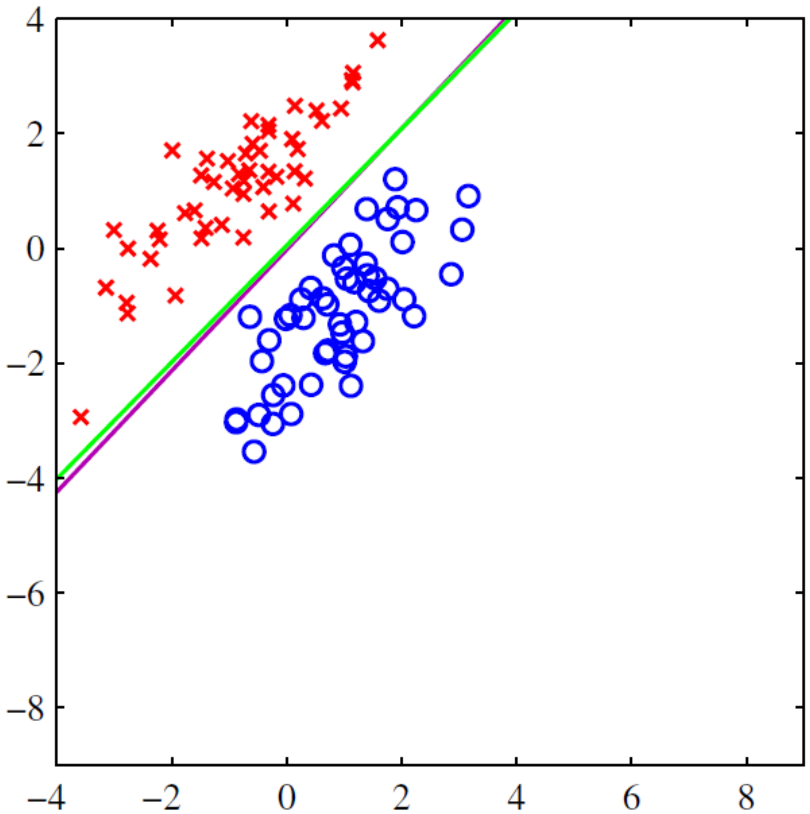
$$\tilde{\mathbf{w}} = \left(\tilde{X}\tilde{X}^T\right)^{-1}\tilde{X}\cdot\mathbf{t} = \tilde{X}^\dagger\cdot\mathbf{t}$$

  – 从而

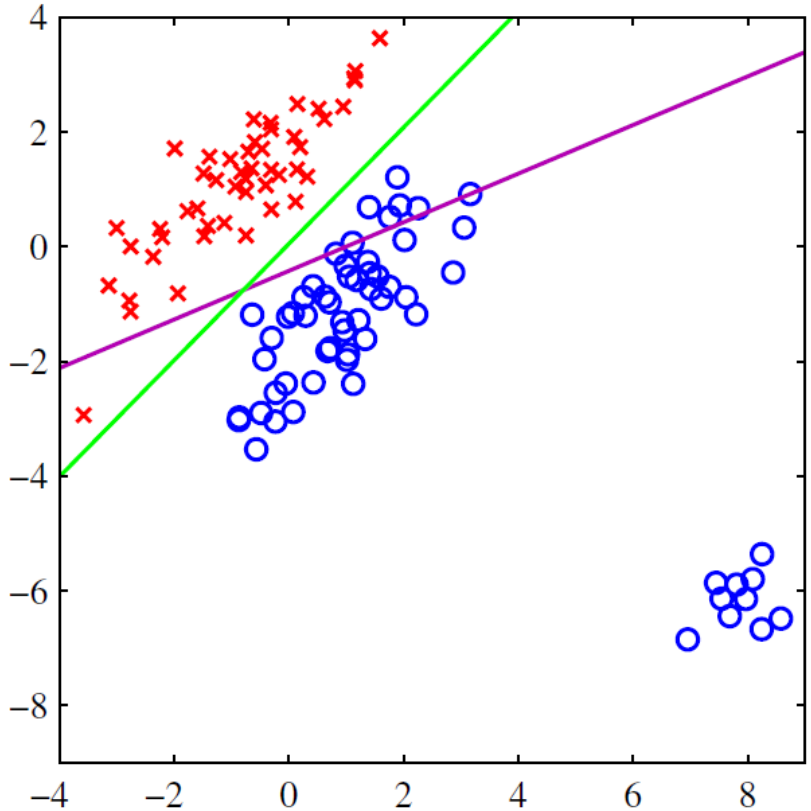
$$y(\mathbf{x}) = \mathbf{t}^T\tilde{X}^T\left(\tilde{X}\tilde{X}^T\right)^{-1}\tilde{\mathbf{x}}$$

示例：最小二乘法得到的分类超平面

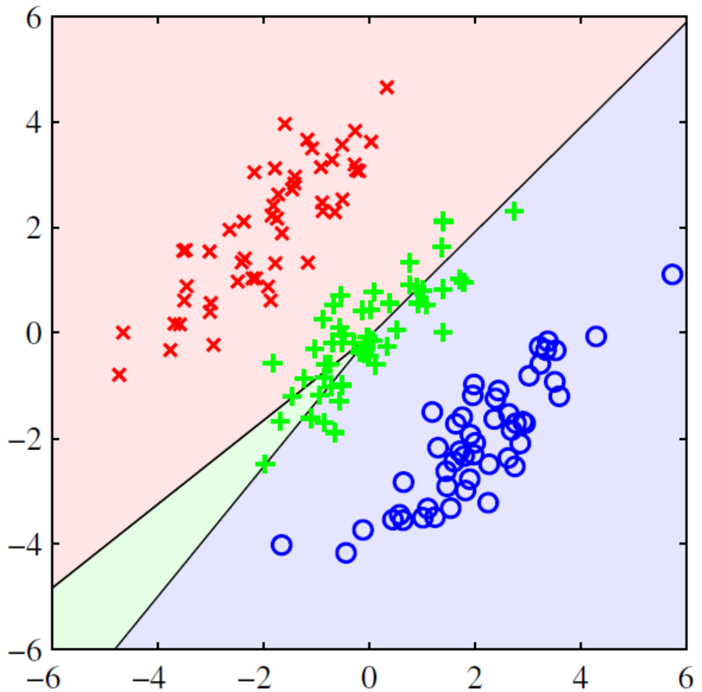
-



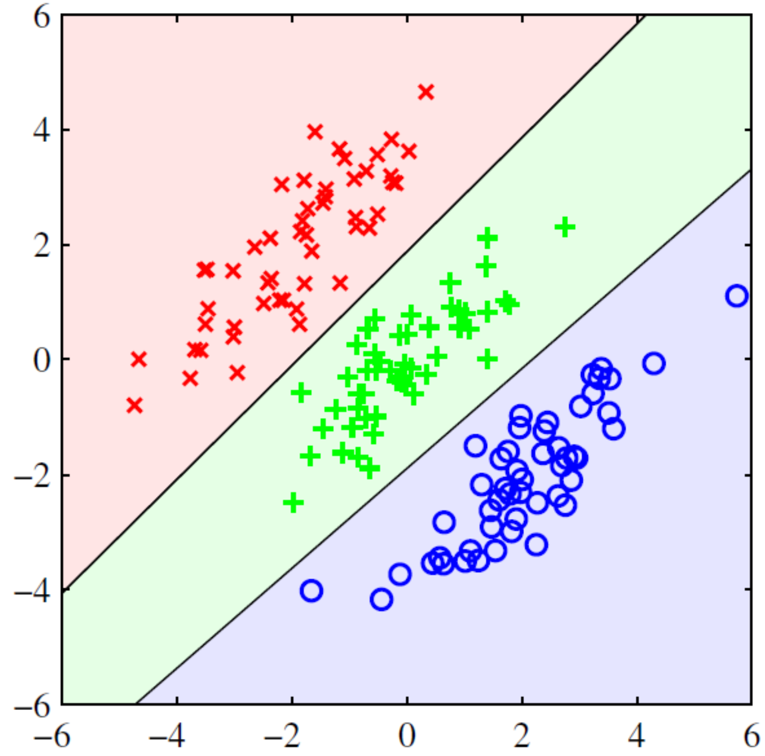
如果数据中存在异常值(outliers)则分类超平面出现严重偏移!



例：如果有3个类别，怎么样?

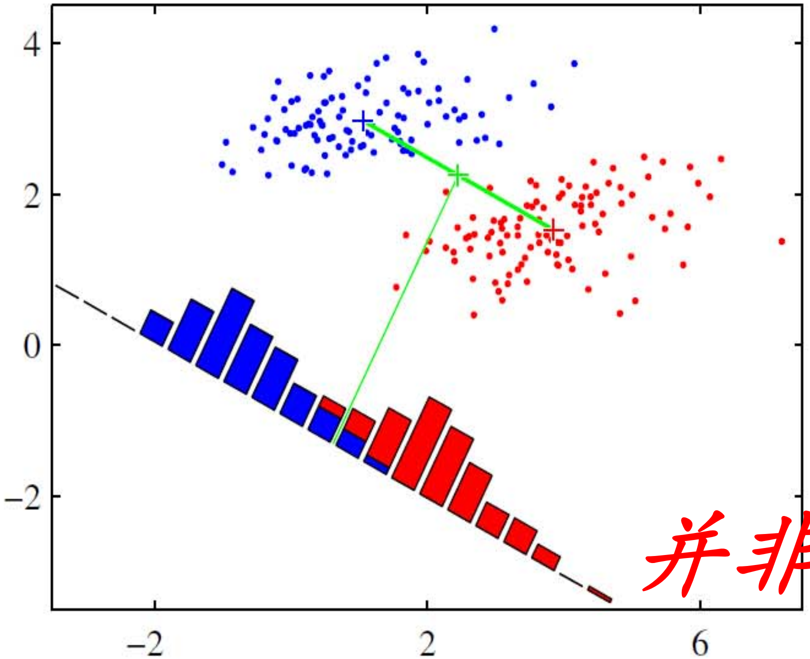
- 

?



Fisher 线性鉴别分析

- 基本思路: $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$
  - 寻找一个投影方向，使得两个类别的样本最大可能地分离开
    - 如果向类别中心的连线投影，则如左图所示：

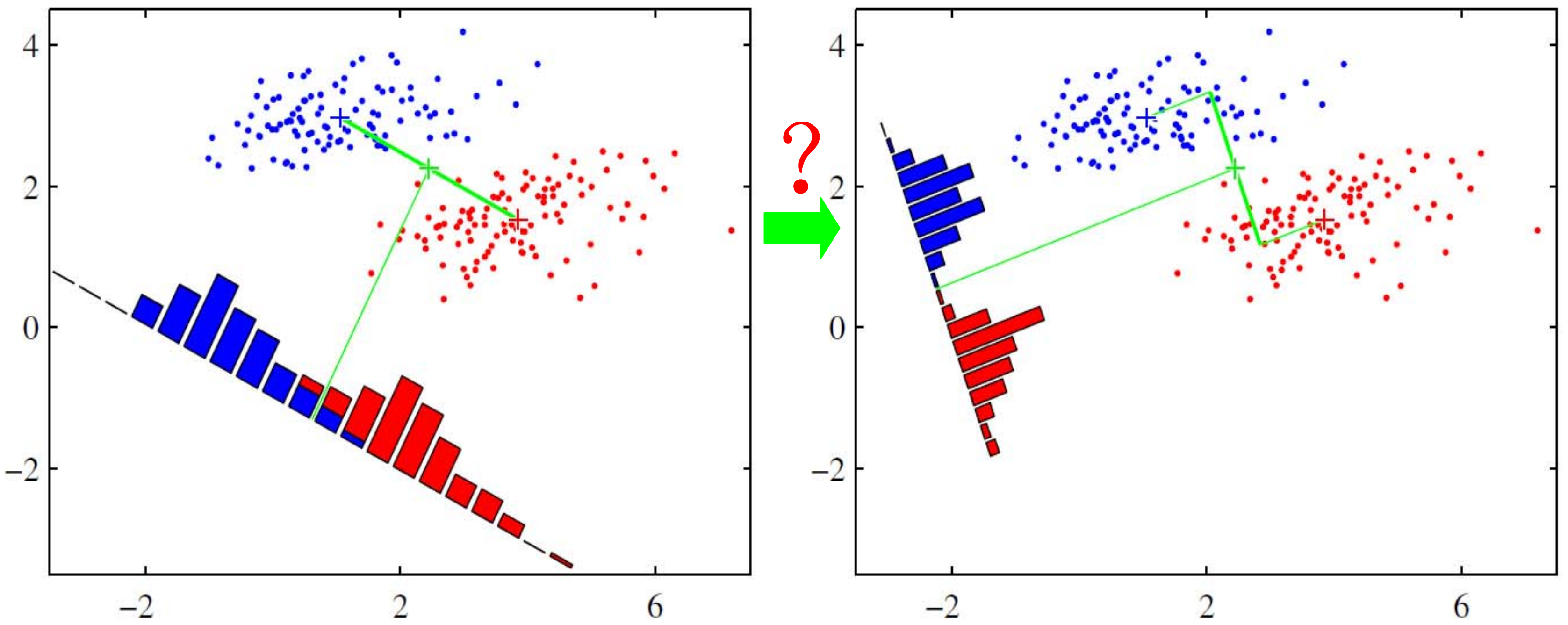


并非最优！

Fisher 线性鉴别分析

- 基本思路:
  - 寻找一个投影方向，使得两个类别的样本最大可能地分离

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$



Fisher 线性鉴别分析

- 基本思路: $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$
  - 寻找一个投影方向，使得两个类别的样本最大可能地分离
    - 最大化J(w) $J(\mathbf{w}) = \dfrac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$
    - 其中 $m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k$ $\quad s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$
    - 改写为使用类内和类间协方差矩阵的形式

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \qquad \mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}$$

$$\Longrightarrow \quad \mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

多类别Fisher鉴别分析

- 对于多个类别，使用目标函数

$$J(\mathbf{W}) = \mathrm{Tr}\left\{\mathrm{s}_{\mathrm{W}}^{-1}\mathrm{s}_{\mathrm{B}}\right\}$$

寻找最大化目标函数J(W)的投影矩阵W

  – 其中

$$\mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$$

$$\mathrm{s}_{\mathrm{W}} = \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\mathbf{y}_n - \boldsymbol{\mu}_k)(\mathbf{y}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\mathrm{s}_{\mathrm{B}} = \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathrm{T}}$$

- 目标函数等价于

$$J(\mathbf{w}) = \mathrm{Tr}\left\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\right\}$$

感知器算法

- 假设有两个类别C_1和C_2

$$t=+1 \text{ for class } \mathcal{C}_1 \quad t=-1 \text{ for class } \mathcal{C}_2$$

$$y(\mathbf{x})=f\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})\right) \qquad f(a)=\begin{cases}+1, & a \geqslant 0\\ -1, & a<0\end{cases}$$

  – 定义感知器错分代价

$$E_{\mathrm{P}}(\mathbf{w})=-\sum_{n\in\mathcal{M}}\mathbf{w}^{\mathrm{T}}\phi_n t_n$$

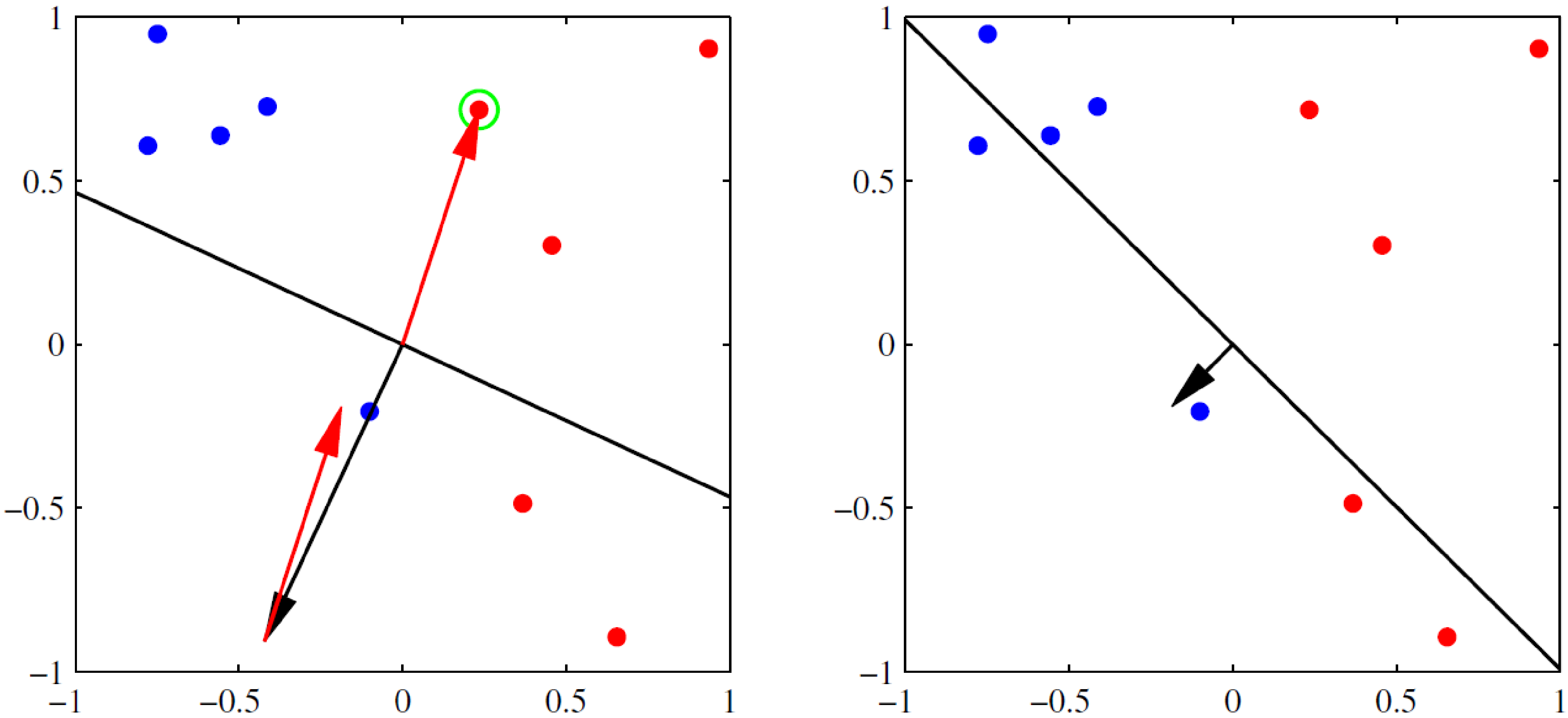
其中**M**为出现错分的样本下标集合

$$\boxed{\mathbf{w}^{(\tau+1)}=\mathbf{w}^{(\tau)}-\eta\nabla E_{\mathrm{P}}(\mathbf{w})=\mathbf{w}^{(\tau)}+\eta\phi_n t_n}$$

如果该样本被错分，则调整权值**w**

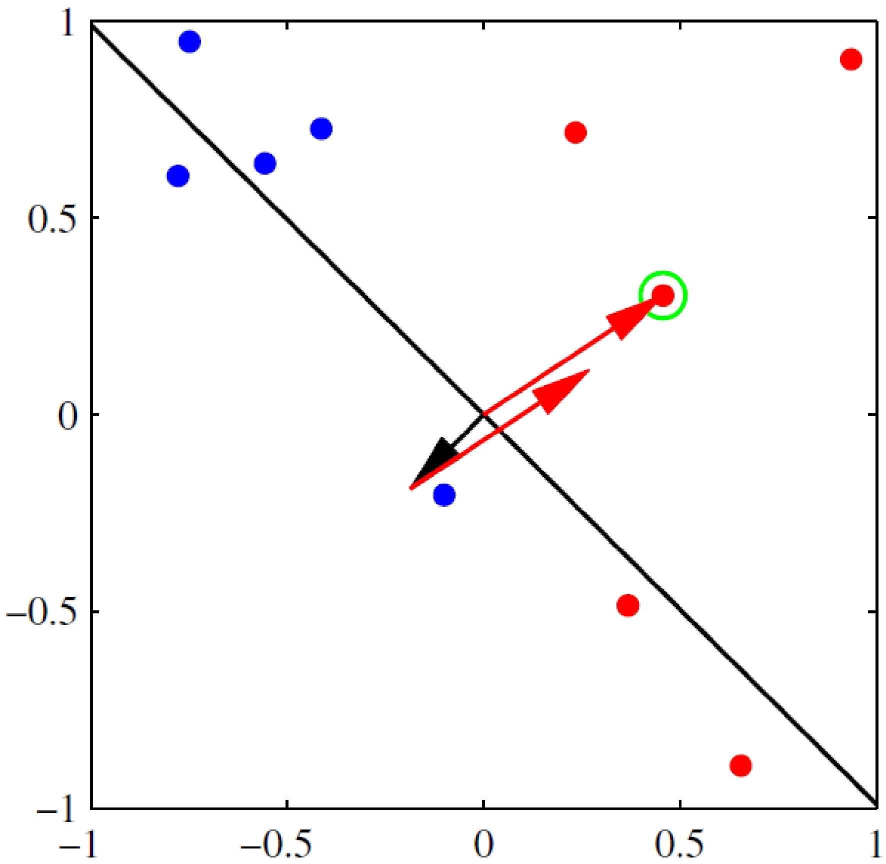
感知器算法示例

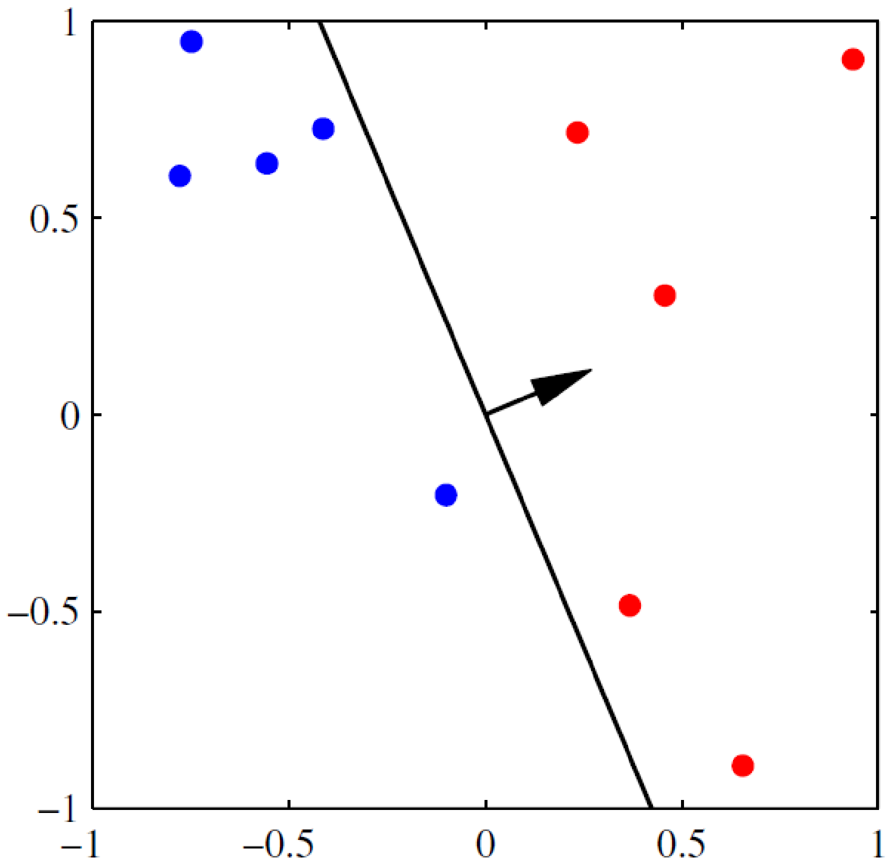
- 假设绿色圈的样本被错分:



感知器算法示例

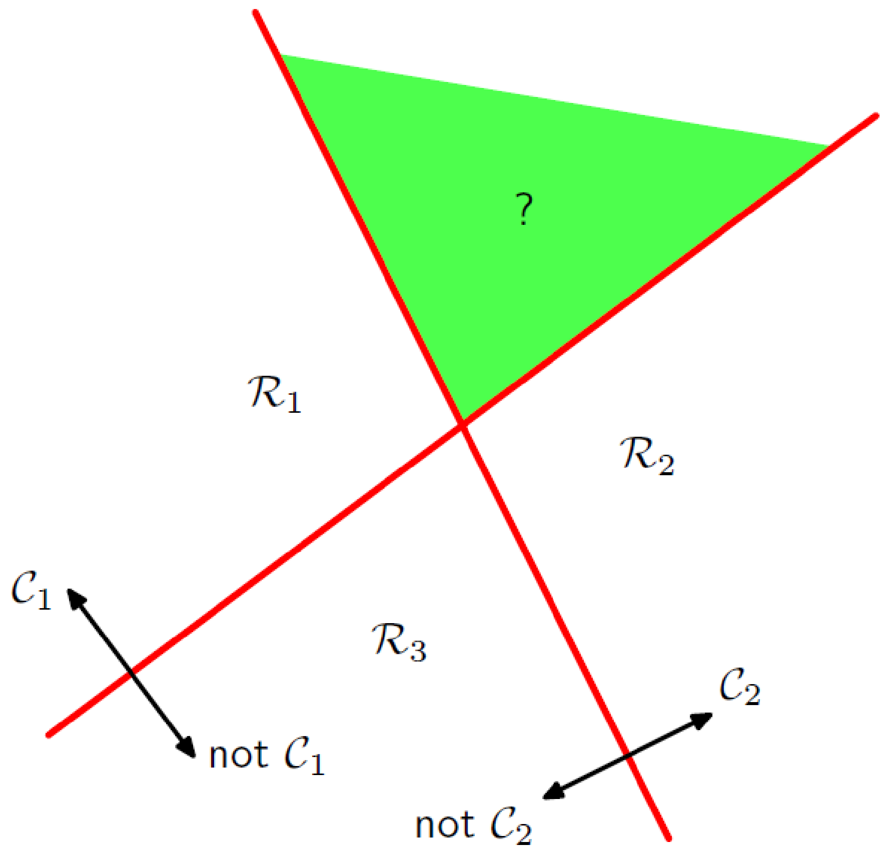
- 假设绿色圈的样本被错分:





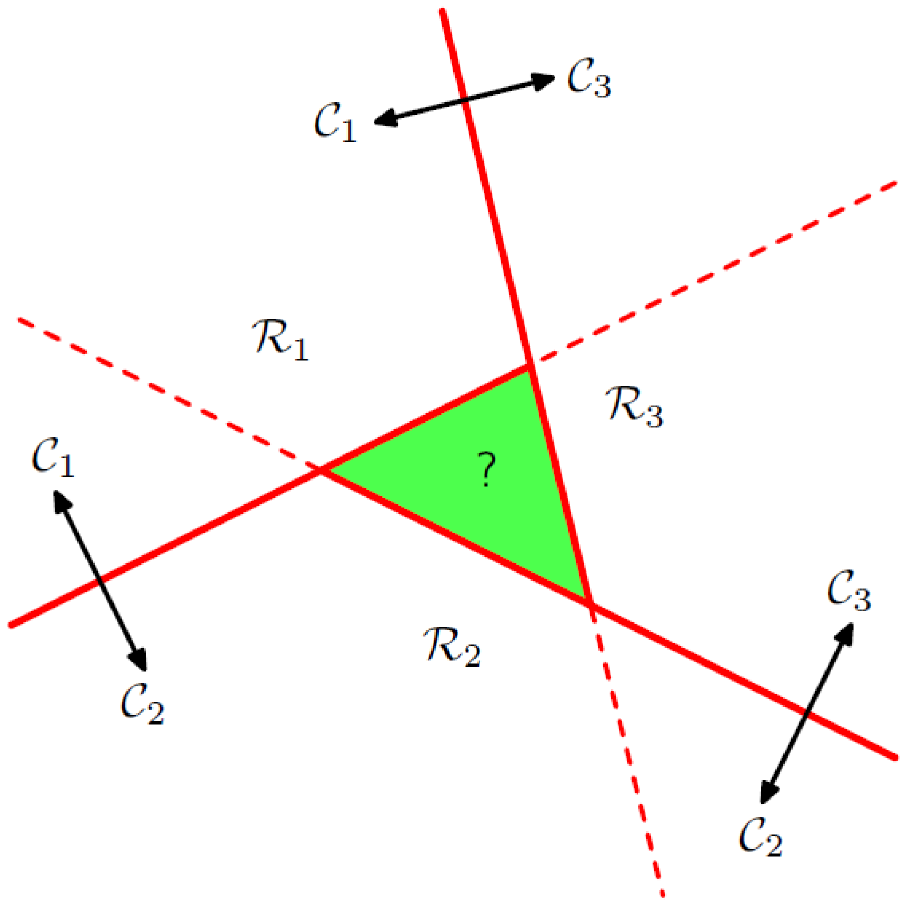
多个类别的分类问题

- one-versus-the-rest



多个类别的分类问题

- one-versus-one



Q / A

- Any Questions...

用于分类的线性模型 内容提要

- 引子: 2个类别的分类问题
- 鉴别函数
  - 最小二乘法用于分类
  - Fisher准则
  - 感知器法则
- 概率生成模型
  - 连续型输入数据
  - 离散型输入数据
- 概率鉴别模型
  - 逻辑斯蒂回归
  - 多类逻辑斯蒂回归
  - 贝叶斯逻辑斯蒂回归

2分类问题与Logistic激活函数

- 考虑一个2类别的分类问题

$$
\begin{aligned}
p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\
&= \frac{1}{1+\exp(-a)} = \sigma(a)
\end{aligned}
$$

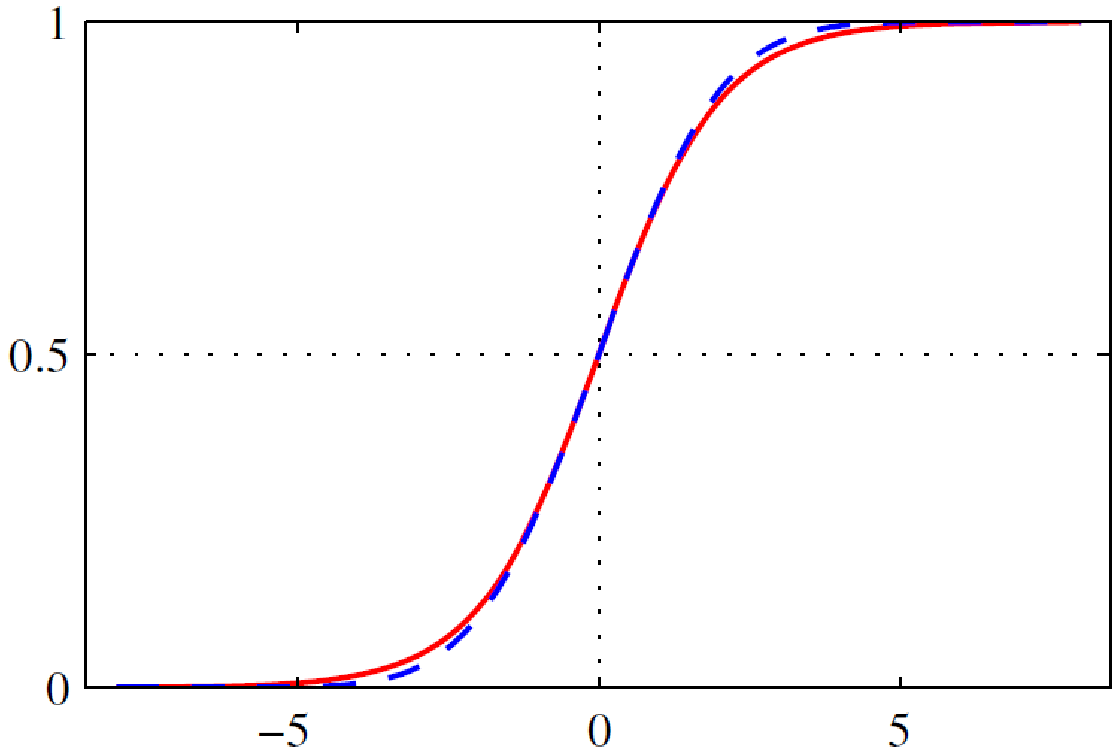
  – 其中

$$\sigma(a) = \frac{1}{1+\exp(-a)} \qquad\qquad a = \ln\left(\frac{\sigma}{1-\sigma}\right)$$

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

逻辑斯蒂(Logistic)激活函数

- Logistic 型Sigmoid函数



连续型输入数据：2个类别

- 假设:
  - **2个类别条件密度(class-conditional densities)均为Gaussian分布**
  - **2个类别共享同一个协方差矩阵(covariance matrix)**

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

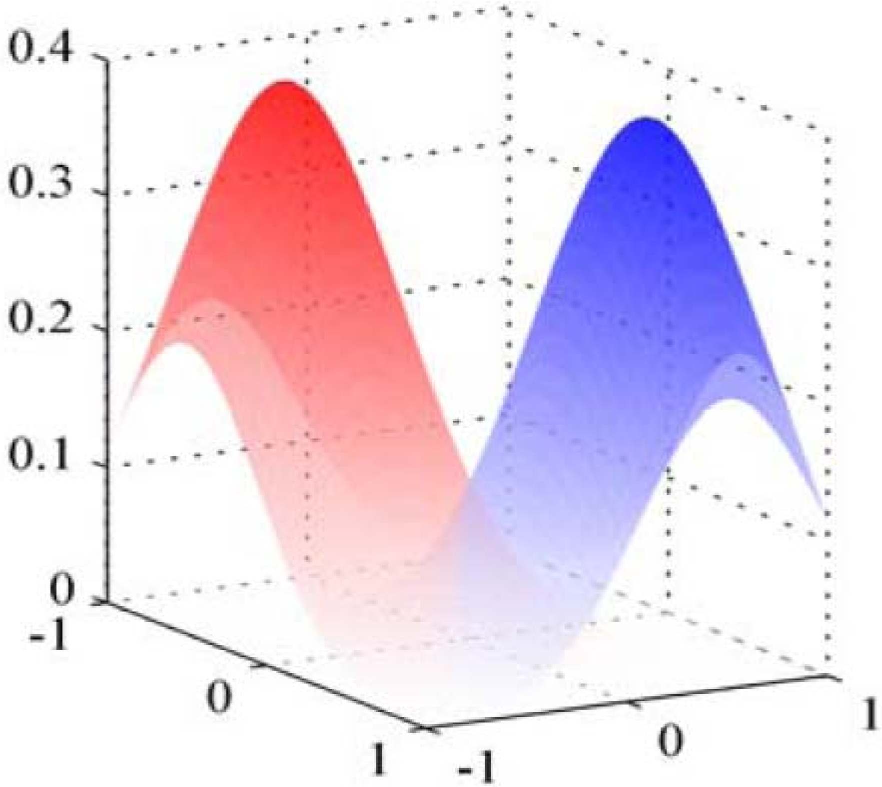
$$\Rightarrow \quad p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

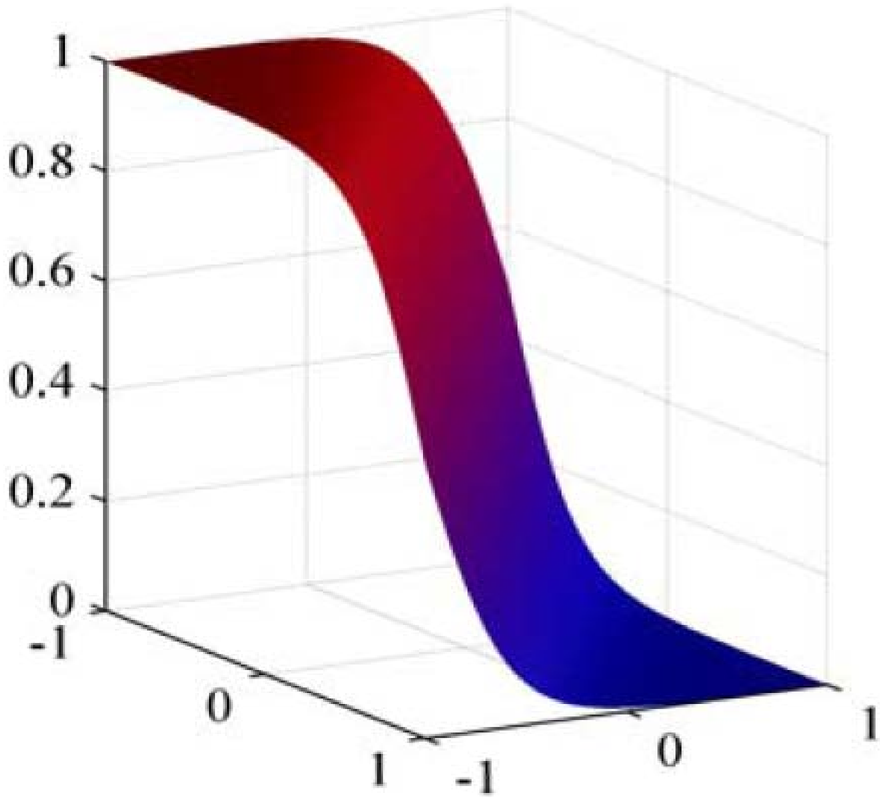
未知参数如何估计?

其中
$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$
$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

图示：2个高斯分布类别

- 联合分布 & 决策曲面





参数的最大似然估计(MLE)

- 给定了类别条件密度 $p(\mathbf{x}\,|\,C_k)$ 和类别先验概率 $p(C_k)$，我们可以基于训练数据估计模型中的参数
- 设训练数据为$\{\mathbf{x}_n, t_n\}_{n=1}^{N}$，其中 $p(\mathcal{C}_1)=\pi \quad p(\mathcal{C}_2)=1-\pi$

$$t_n=\begin{cases}1 & \text{class } C_1\\ 0 & \text{class } C_2\end{cases}$$

- 则对每个数据点x_n，我们有:

$$p(\mathbf{x}_n,\mathcal{C}_1)=p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1)=\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$$

$$p(\mathbf{x}_n,\mathcal{C}_2)=p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2)=(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$$

  - 似然函数为：

$$p(\mathbf{t}|\pi,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma})=\prod_{n=1}^{N}\left[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})\right]^{t_n}\left[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})\right]^{1-t_n}$$

  - 取对数似然函数，求导数

$$\Rightarrow \quad \pi=\frac{1}{N}\sum_{n=1}^{N}t_n=\frac{N_1}{N}=\frac{N_1}{N_1+N_2}$$

参数的最大似然估计(MLE)

- 类似地，根据似然函数

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

  – 取对数似然函数，然后对两个均值分别求偏导

$$\Rightarrow \quad \boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n \qquad \boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n=1}^{N} (1-t_n)\mathbf{x}_n$$

参数的最大似然估计(MLE)

- 类似地，根据似然函数

$$p(\mathbf{t}|\pi,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1,\boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2,\boldsymbol{\Sigma})]^{1-t_n}$$

  – 取对数似然函数，整理得出： $-\frac{N}{2}\ln|\Sigma| - \frac{N}{2}\mathrm{tr}\left(\Sigma^{-1}S\right)$

  – 然后对协方差矩阵求导，利用 $\frac{\partial}{\partial A}\ln|A| = \left(A^{-1}\right)^T$

$\Rightarrow$ $\Sigma = S$ 其中

$$\mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2$$

$$\mathbf{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}(\mathbf{x}_n-\boldsymbol{\mu}_1)(\mathbf{x}_n-\boldsymbol{\mu}_1)^{\mathrm{T}}$$

$$\mathbf{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}(\mathbf{x}_n-\boldsymbol{\mu}_2)(\mathbf{x}_n-\boldsymbol{\mu}_2)^{\mathrm{T}}$$

对于多类别的分类问题

- 使用“软最大”(Soft-max)函数
  - 归范化 **(normalization)**

$$
\begin{aligned}
p(\mathcal{C}_k|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} \\
&= \frac{\exp(a_k)}{\sum_j \exp(a_j)}
\end{aligned}
$$

  - 其中 $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$

连续型输入数据： k个类别

- 假设:
  - **k个类别条件密度(class-conditional densities)均为Gaussian分布**
  - **所有类别共享同一个协方差矩阵(covariance matrix)**

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

$$\begin{aligned} \mathbf{w}_k &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \\ w_{k0} &= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k) \end{aligned}$$

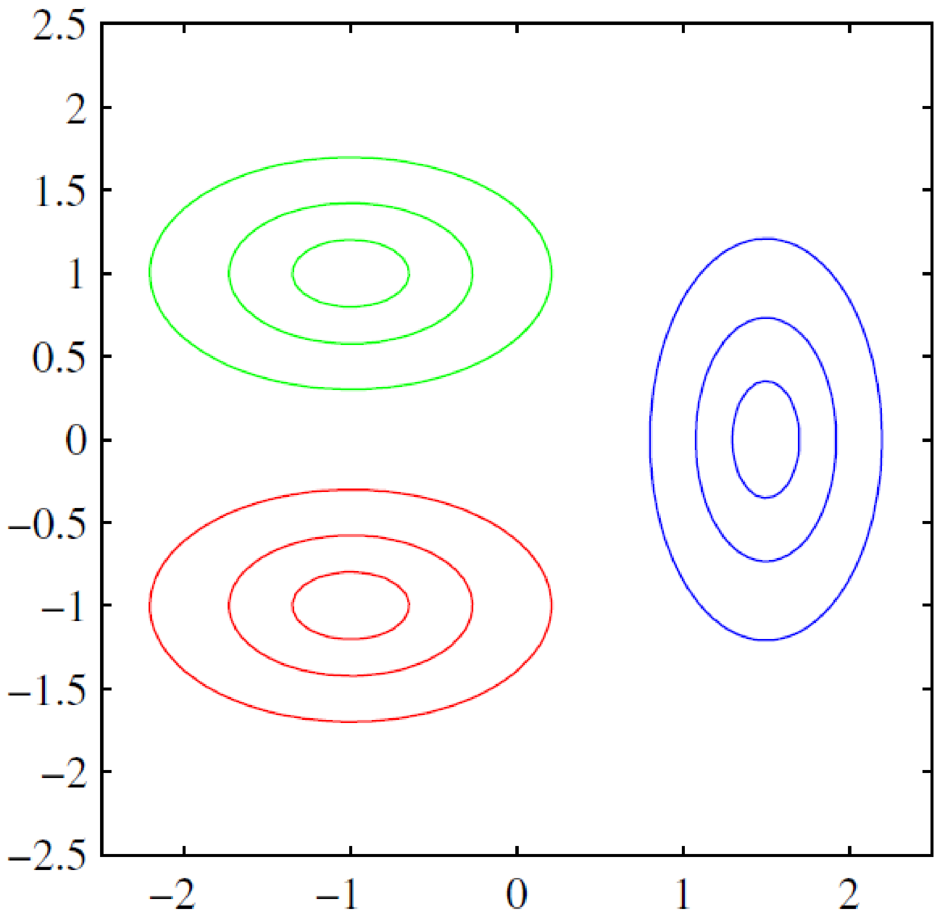
连续型输入数据：k个类别

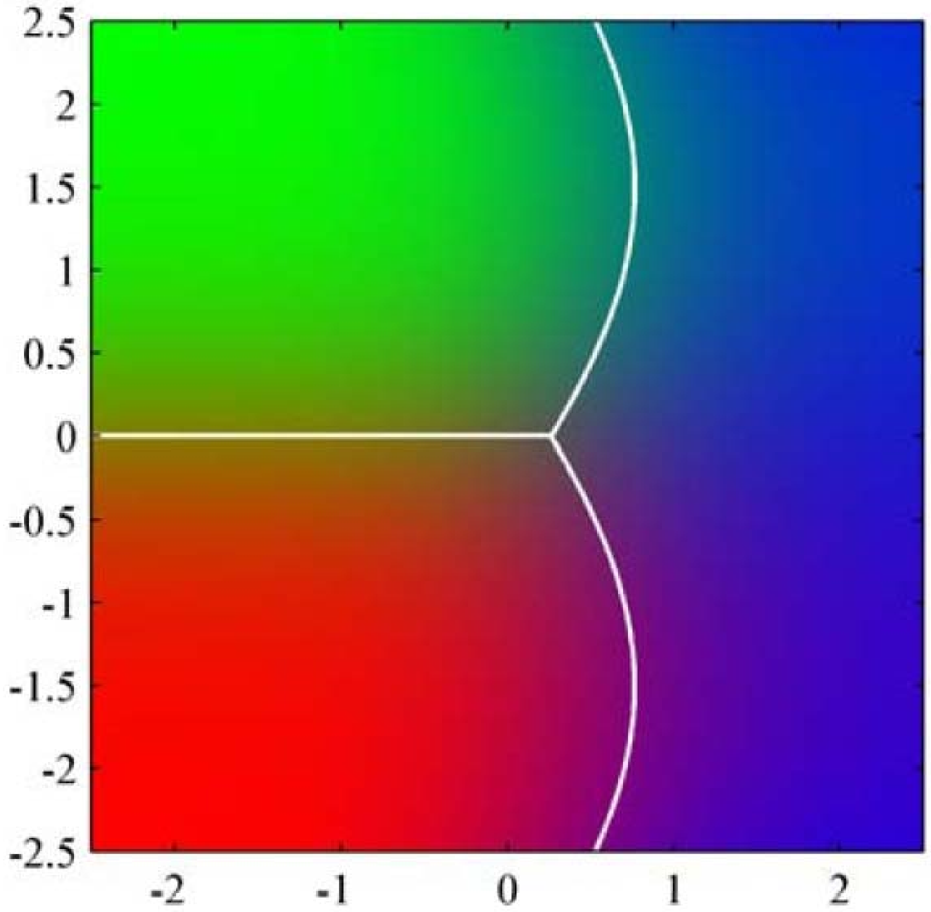
- 假设:
  - **k**个类别条件密度**(class-conditional densities)**均为**Gaussian**分布
  - 每个类别使用自己的协方差矩阵**(covariance matrix)**

  - 则获得二次鉴别 **(quadratic discriminant)**函数

示例：线性&二次决策边界

- 共享协方差则获得线性决策边界；如果协方差矩阵不同，则获得二次决策边界





第2次 作业

- 考虑k个类别的分类问题
  - 假设K (K>2)个类别条件密度(class-conditional densities)均为Gaussian分布
  - 每个类别使用自己的协方差矩阵(covariance matrix) Σ_k

  则获得二次鉴别 (quadratic discriminant)函数
  - 请完成:
    - 推导该二次鉴别函数
    - 给定训练数据集$\{\mathbf{x}_n, \mathbf{t}_n\}$，其中n=1,…,N. 试估计上述二次鉴别函数模型中的几个参数(k类别先验概率$\{\pi_k\}$、k个均值$\{\boldsymbol{\mu}_k\}$和k个协方差矩阵$\{\Sigma_k\}$)
    - 对你所获得的结果给出简要分析与讨论

离散型输入数据

- 考虑D维的离散型特征，其中每一个维为二值变量，即 $x_i \in \{0,1\}$
- 假设给定类别C_k条件下，各维特征相互独立，则类条件分布为

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

  – 根据定义 $a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$
  – 得出

$$a_k(\mathbf{x}) = \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1-x_i)\ln(1-\mu_{ki})\} + \ln p(\mathcal{C}_k)$$

指数族(exponential family)

- 类别k的条件密度为

$$p(\mathbf{x}|\boldsymbol{\lambda}_k) = h(\mathbf{x})g(\boldsymbol{\lambda}_k)\exp\left\{\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

  – 增加尺度参数，则

$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s}h\left(\frac{1}{s}\mathbf{x}\right)g(\boldsymbol{\lambda}_k)\exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}$$

  – 得出

$$a(\mathbf{x}) = (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_1) - \ln g(\boldsymbol{\lambda}_2) + \ln p(\mathcal{C}_1) - \ln p(\mathcal{C}_2)$$

$$a_k(\mathbf{x}) = \boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x} + \ln g(\boldsymbol{\lambda}_k) + \ln p(\mathcal{C}_k)$$

Q / A

- Any Questions...

用于分类的线性模型 内容提要

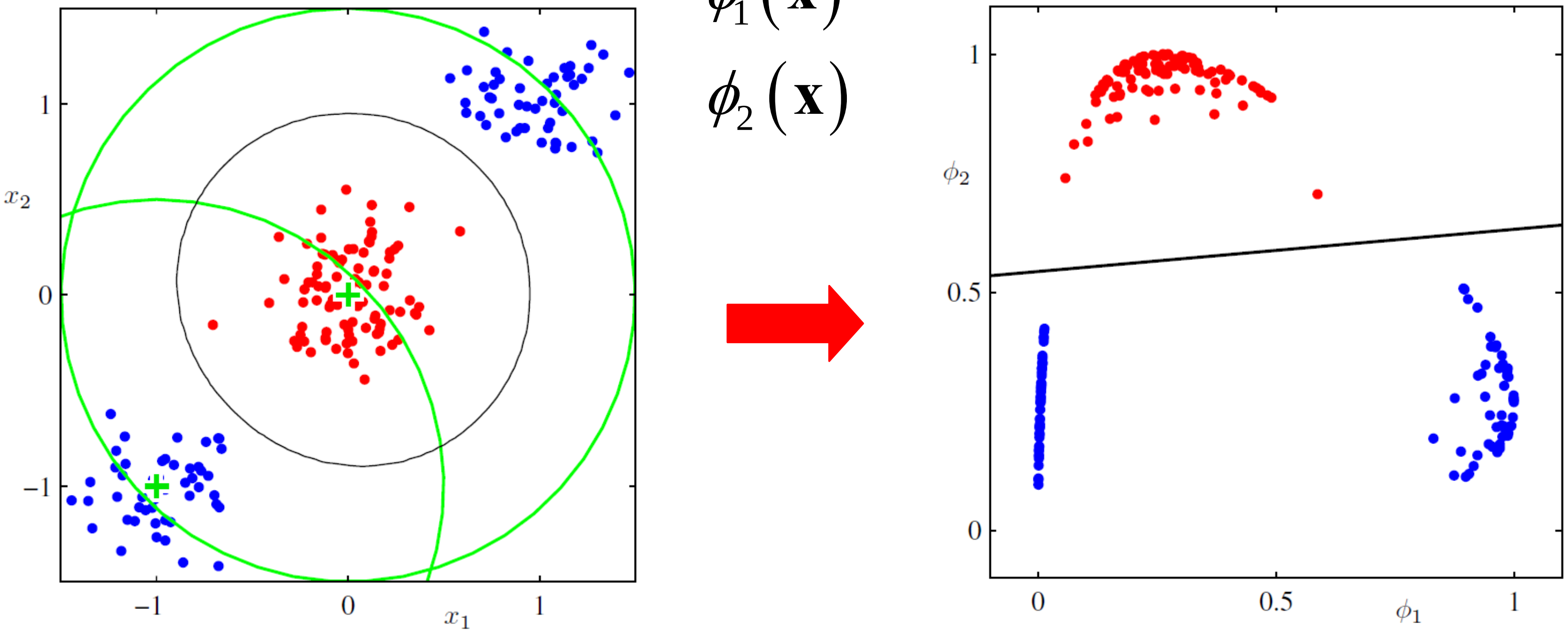
- 引子: 2个类别的分类问题
- 鉴别函数
  - 最小二乘法用于分类
  - **Fisher**准则
  - 感知器法则
- 概率生成模型
  - 连续型输入数据
  - 离散型输入数据
- 概率鉴别模型
  - 逻辑斯蒂回归
  - 多类逻辑斯蒂回归
  - 贝叶斯逻辑斯蒂回归

广义线性分类模型

- 使用非线性基函数对数据进行非线性变换
  - 比如使用**2**个高斯基函数

$$\phi_1(\mathbf{x})$$

$$\phi_2(\mathbf{x})$$



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

注意：下标从0开始

逻辑斯蒂(Logistic)回归

- 考虑一个2类别问题，则Logistic 回归模型为

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right)$$

  - 其中需要确定的参数数目为: **M**个
  - 如果使用**Gaussian**分布，则参数个数为:

    2M + M(M+1)/2

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

参数w的MLE估计

- 给定训练数据为 $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$ ，其中类别标签为1和0，即

$$t_n = \begin{cases} 1 & \text{class } C_1 \\ 0 & \text{class } C_2 \end{cases}$$

  - 根据 $p(\mathcal{C}_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right)$
  - 则对数据点$\mathbf{x}_\mathbf{n}$, 对于的输出为

$$y_n = y\left(\phi\left(\mathbf{x}_n\right)\right) = \sigma\left(\mathbf{w}^T\phi\left(\mathbf{x}_n\right)\right)$$

参数w的MLE估计

- 似然函数为

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n}$$

– 取负对数，则

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

– 计算梯度

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

其中使用到 $\frac{d\sigma}{da} = \sigma(1 - \sigma)$

没有closed form solution!

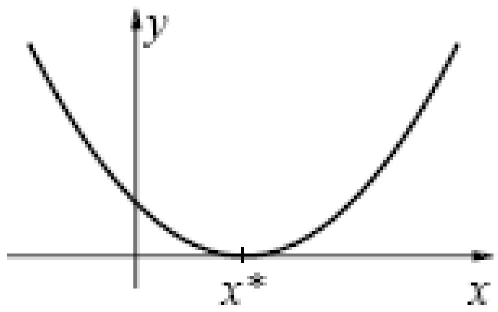
问题定义与基本概念

- 无约束最优化问题：

$$arg\,\mathbf{min}\quad f(x)$$

其中 $\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^T\in \boldsymbol{R}^n$， $\boldsymbol{f}:\boldsymbol{R}^n\mapsto \boldsymbol{R}$

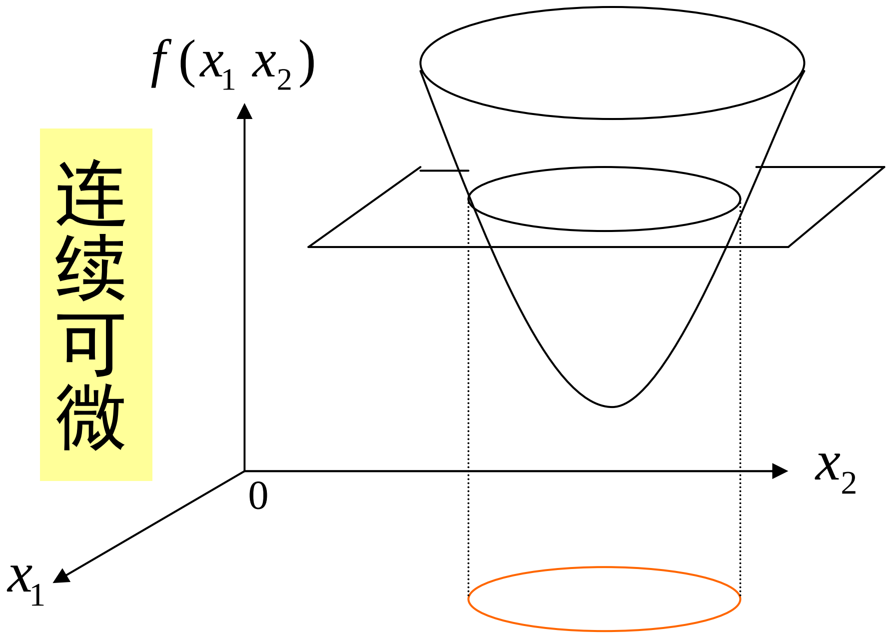
- 基本概念
  - 下降方向
  - 可行方向
  - 最优解条件

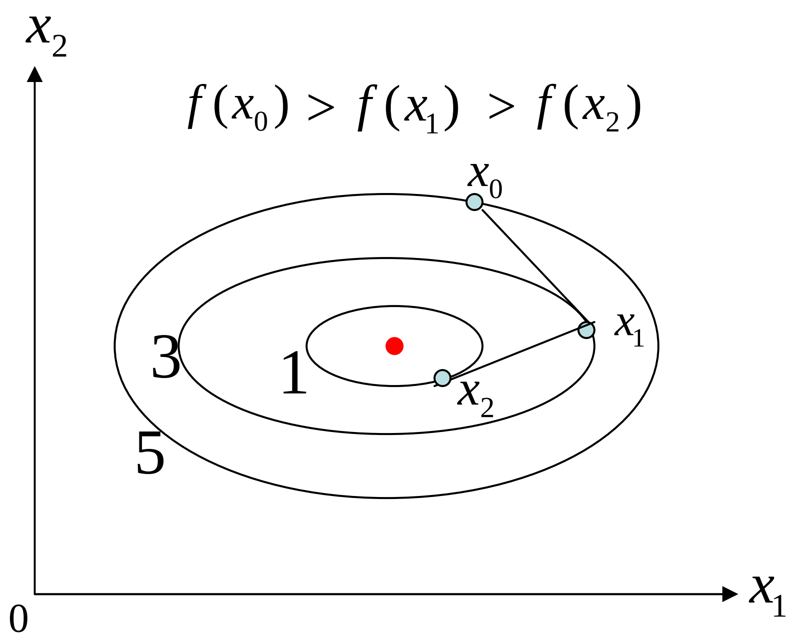


基本思想

- 迭代下降算法：
  - 寻找一个搜索方向 p，使得每次迭代时，

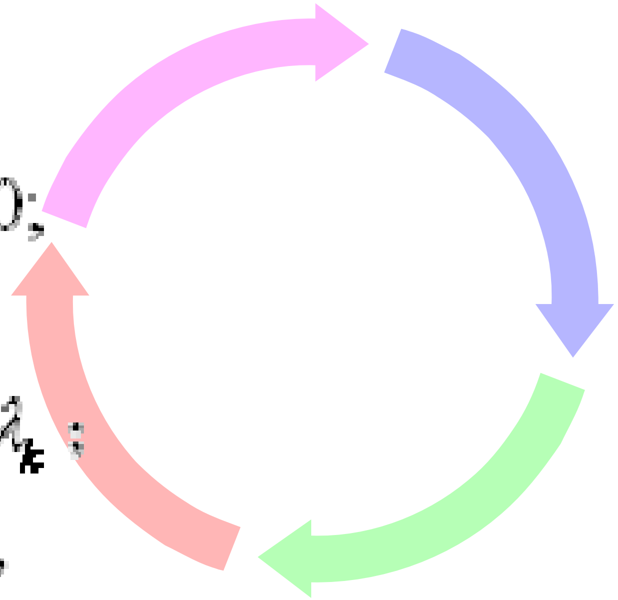
函数值减小，即 $x_{k+1} = x_k + p$ 时， $f\left(x_{k+1}\right) \le f\left(x_k\right)$





迭代下降法的基本步骤

- 4个基本步骤：



第 1 步　　选取初始点 $x^{(0)}$，k:=0;

第 2 步　　构造搜索方向 $d^{(k)}$；

第 3 步　　根据 $d^{(k)}$，确定步长 λ_k；

第 4 步　　令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$，

若 $x^{(k+1)}$ 已满足某种终止条件，停止迭代，输出近似解 $x^{(k+1)}$；否则令 k:=k+1，转回第 2 步。

迭代下降法的核心问题

- **搜索方向**的选择问题是迭代下降法的核心问题
  - 搜索方向的不同选择方式，对应着不同的算法
- **与目标函数的导数有关**, 比如:
  - 最速下降法
  - 牛顿法
  - 阻尼牛顿法、修正牛顿法
  - 拟牛顿法
  - 共轭梯度法

最速下降法

- 要求：目标函数 $f: R^n \mapsto R$ 一阶连续可微
  - 由柯西(Cauchy)在1847年提出的，是求无约束极值的最早的数值算法
- 步骤：

第 1 步　选取初始点 $x^{(0)}$，给定终止误差 $\varepsilon > 0$，令 $k := 0$；

第 2 步　计算 $\nabla f\left(x^{(k)}\right)$，若 $\left\|\nabla f\left(x^{(k)}\right)\right\| < \varepsilon$，停止迭代，输出 $x^{(k)}$，否则进行第 3 步；

第 3 步　取 $d^{(k)} = -\nabla f\left(x^{(k)}\right)$

第 4 步　进行一维搜索，求 λ_k，使得

$$f\left(x^{(k)} + \lambda_k d^{(k)}\right) = \min_{\lambda \ge 0} f\left(x^{(k)} + \lambda d^{(k)}\right)$$

令 $x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}$，$k := k+1$，转第 2 步。

牛顿法

- 设$f(x)$是二次可微的实函数，$x \in R^n$，又设$x^{(k)}$是$f(x)$的极小点的一个估计，我们把$f(x)$在$x^{(k)}$展开 Taylor 级数，并取二阶近似：

$$f(x) \approx \phi(x) = f\left(x^{(k)}\right) + \nabla f\left(x^{(k)}\right)^T \left(x - x^{(k)}\right)$$
$$+\frac{1}{2}\left(x - x^{(k)}\right)^T \nabla^2 f\left(x^{(k)}\right)\left(x - x^{(k)}\right)$$

其中$\nabla^2 f\left(x^{(k)}\right)$是$f(x)$在$x^{(k)}$处的 Hessian 矩阵. 为求$\phi(x)$的平稳点，令$\nabla\phi(x) = 0$，即$\phi(x) = \nabla f\left(x^{(k)}\right) + \nabla^2 f\left(x^{(k)}\right)\left(x - x^{(k)}\right) = 0$

设$\nabla^2 f\left(x^{(k)}\right)$可逆，那么可以得到牛顿法的迭代公式：

$$x^{(k+1)} = x^{(k)} - \nabla^2 f\left(x^{(k)}\right)^{-1} \nabla f\left(x^{(k)}\right),$$

其中$\nabla^2 f\left(x^{(k)}\right)^{-1}$是 Hessian 矩阵的逆矩阵

牛顿法

- 牛顿法
  - 目标函数要求二次可微
  - Taylor级数展开，取二阶近似
    - 确定函数的近似平稳点
  - 步骤

第 1 步 选定初始点 $x^{(0)} \in R^n$，给定允许误差 $\varepsilon > 0$，

令 k=0；

第 2 步 求 $\nabla f\left(x^{(k)}\right)$，$\left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1}$，检验：若 $\nabla f\left(x^{(k)}\right) < \varepsilon$

则停止迭代，$x^{(*)} = x^{(k)}$. 否则，转向(3)；

第 3 步 令 $d^{(k)} = -\left(\nabla^2 f\left(x^{(k)}\right)\right)^{-1} \nabla f\left(x^{(k)}\right)$（牛顿方向）；

第 4 步 $x^{(k+1)} = x^{(k)} + d^{(k)}$，$k \leftarrow k+1$，转回(2)

对比/回顾: 线性最小二乘

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

- 线性最小二乘:
  - 存在**closed-form solution**

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

  - 借助**Newton**法，一步迭代就得到最优点

$$\nabla E(\mathbf{w}) = \sum^{N} (\mathbf{w}^{\mathrm{T}} \phi_n - t_n) \phi_n = \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi} \mathbf{w} - \mathbf{\Phi}^{\mathrm{T}} \mathbf{t}$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{n=1}^{N} \phi_n \phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}$$

$$\boxed{\begin{aligned} \mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - (\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi})^{-1} \left\{ \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi} \mathbf{w}^{(\mathrm{old})} - \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \right\} \\ &= (\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi})^{-1} \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \end{aligned}}$$

迭代重加权最小二乘

- IRLS: Iterative Reweighted Least Squares
  - 逻辑斯蒂回归模型，虽无**closed-form solution**，但问题为**convex**的

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

  - 使用牛顿法

迭代重加权最小二乘

- IRLS: $E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$
  - 使用牛顿法

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}$$

R为对角矩阵，其中 $R_{nn} = y_n(1-y_n)$

$$\begin{aligned}
\mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\
&= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\mathbf{w}^{(\mathrm{old})} - \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\right\} \\
&= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}
\end{aligned}$$

迭代重加权最小二乘

- $$\mathbf{w}^{(\text{new})} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}$$

$$\begin{aligned}\mathbb{E}[t] &= \sigma(\mathbf{x}) = y \\ \operatorname{var}[t] &= \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \sigma(\mathbf{x}) - \sigma(\mathbf{x})^2 = y(1-y)\end{aligned}$$

$$\mathbf{z} = \mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y} - \mathbf{t})$$

多类Logistic回归

- 多类情况，使用Soft-max函数

$$p(\mathcal{C}_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \qquad a_k = \mathbf{w}_k^{\mathrm{T}}\phi$$

- 似然函数为

$$p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

- 交叉熵(cross-entropy)误差函数

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

迭代优化过程求解模型参数

- 在线更新(online update)
  - 序贯**(sequential)**方式

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\, \phi_n$$

- 批处理(Batch)方式

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N} y_{nk} (I_{kj} - y_{nj}) \phi_n \phi_n^{\mathrm{T}}.$$

拉普拉斯近似(Laplace Approximation)

- 考虑一个分布函数p(z)

$$p(z) = \frac{1}{Z} f(z)$$

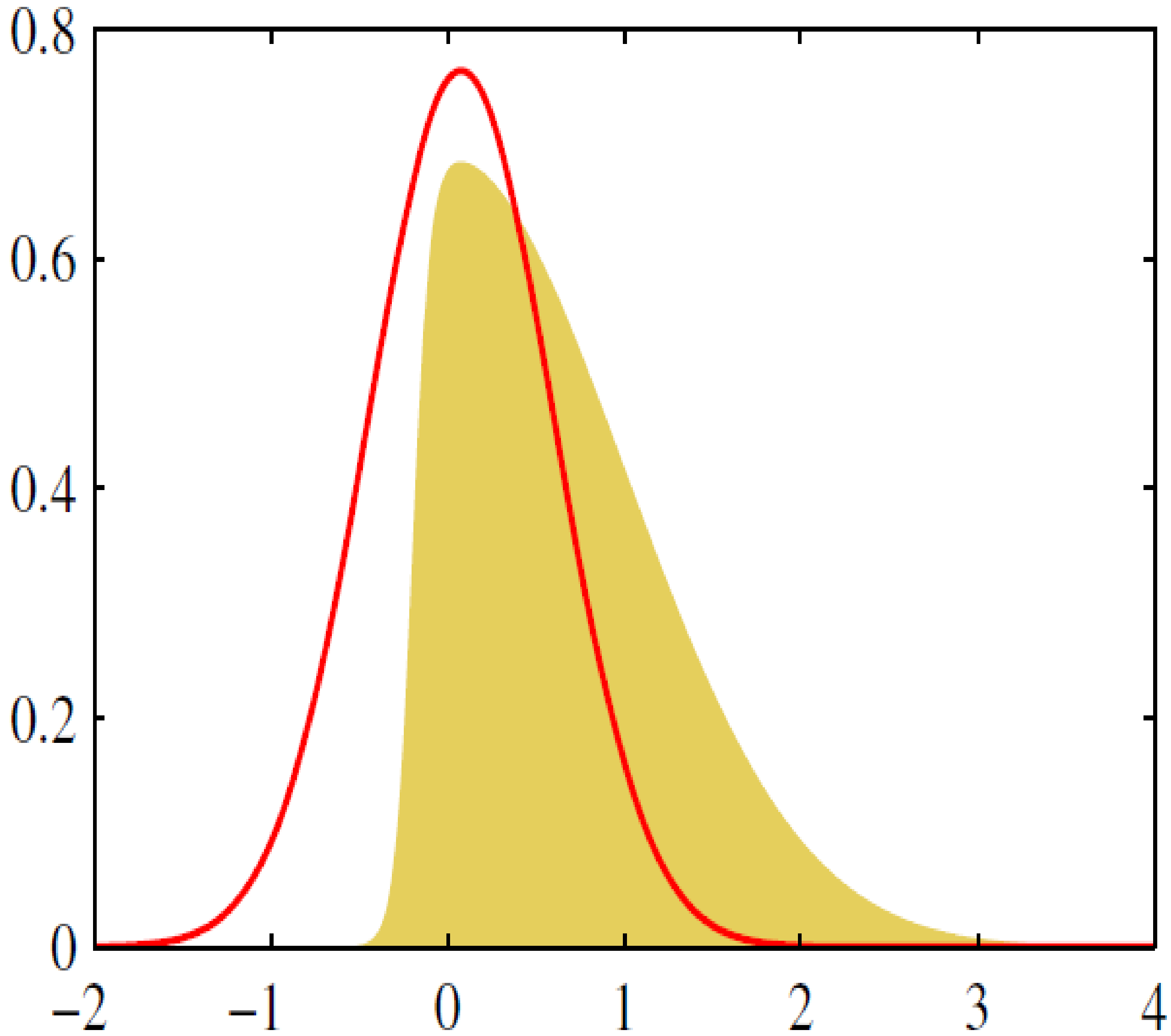
  – **寻找一个近似函数q(z), 其中心点z_0位于p(z)的模(mode)上，即**

$$\left.\frac{df(z)}{dz}\right|_{z=z_0} = 0$$

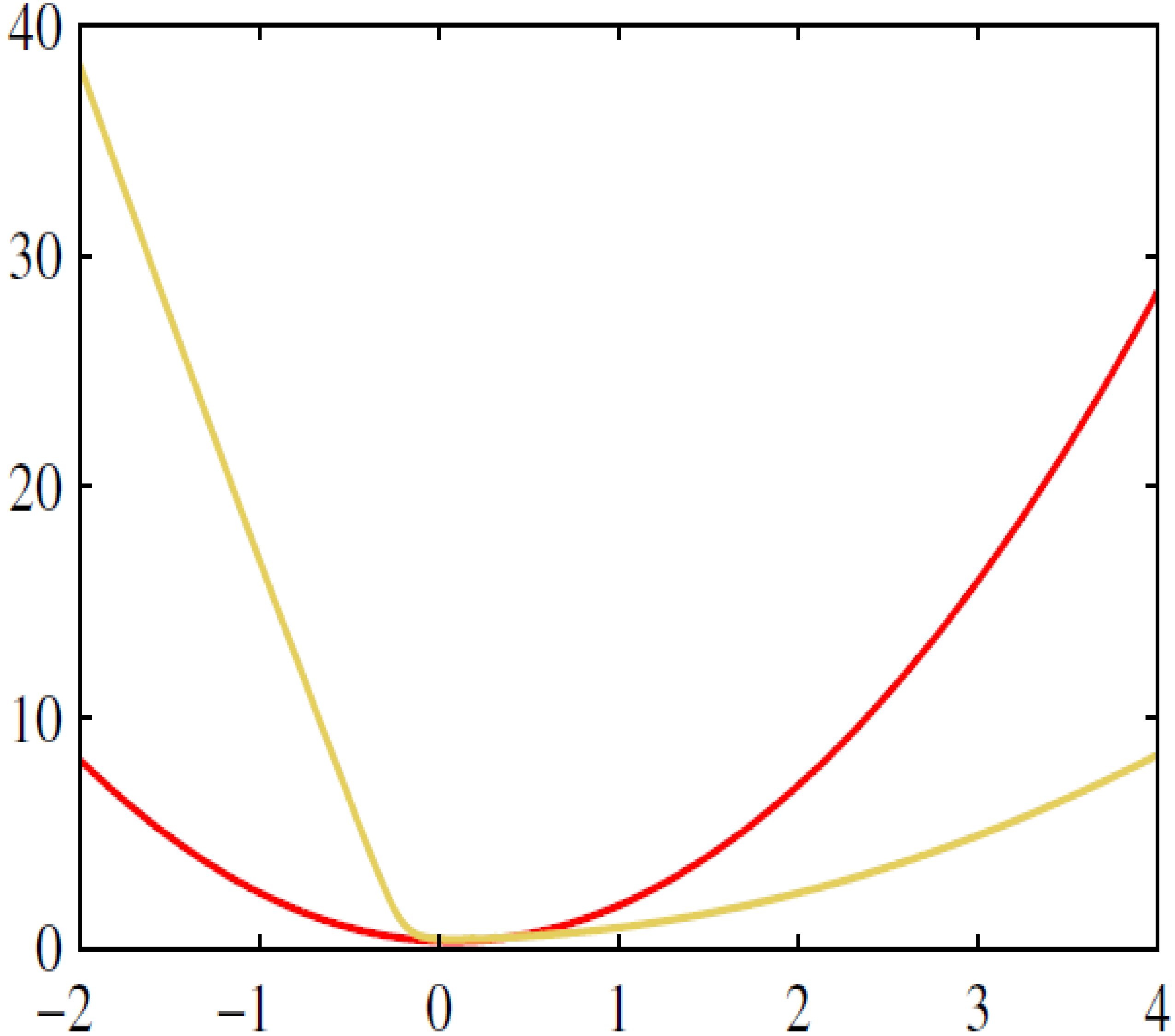
$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2 \qquad A = -\left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0}$$

$$\Rightarrow \quad q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

拉普拉斯近似



拉普拉斯近似



贝叶斯逻辑斯蒂回归

- Logistic回归的贝叶斯方式

- 与贝叶斯线性不同，由于Logistic函数的使用，在这里准确推理无法进行
  - 推理后验概率时，归一化常数**Z**的计算是**intractable**
  - 在计算预测性分布时，对参数**w**的积分是**intractable**
- 解决方法：
  - 使用**Laplace Approximation**

对后验分布做Laplace近似

- 先验分布: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$
- 似然函数: $p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$
  - 对数似然函数为:

$$\begin{aligned}\ln p(\mathbf{w}|\mathbf{t}) \quad = \quad & -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\ & +\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}\end{aligned}$$

对后验分布做Laplace近似

- 给定对数似然函数为:

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}$$

- 构造Laplace近似: q(w)
  - 寻找后验分布的模**(mode)**的位置: 即**MAP**解w_{MAP}
  - 计算负的对数似然函数的**Hessien**矩阵

$$\mathbf{S}_N^{-1} = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}$$

$$\Rightarrow \quad q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)$$

计算预测性分布

- 预测性分布为:

$$p(\mathcal{C}_1|\phi,\mathbf{t}) = \int p(\mathcal{C}_1|\phi,\mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\phi)q(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

– 变形技巧: $\sigma(\mathbf{w}^{\mathrm{T}}\phi) = \int \delta(a-\mathbf{w}^{\mathrm{T}}\phi)\sigma(a)\,\mathrm{d}a$

$$p(a) = \int \delta(a-\mathbf{w}^{\mathrm{T}}\phi)q(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

$$\Rightarrow \int \sigma(\mathbf{w}^{\mathrm{T}}\phi)q(\mathbf{w})\,\mathrm{d}\mathbf{w} = \int \sigma(a)p(a)\,\mathrm{d}a$$

$$\begin{aligned} p(\mathcal{C}_1|\phi,\mathbf{t}) &= \int p(\mathcal{C}_1|\phi,\mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\phi)q(\mathbf{w})\,\mathrm{d}\mathbf{w} \\ &= \int \sigma(a)\mathcal{N}(a|\mu_a,\sigma_a^2)\,\mathrm{d}a \end{aligned}$$

Q / A

- Any Questions...

用于分类的线性模型 内容小结

- 鉴别函数
  - 最小二乘法用于分类 /
  - **Fisher**准则
  - 感知器法则
- 概率生成模型
  - 连续型输入数据
  - 离散型输入数据
- 概率鉴别模型
  - 逻辑斯蒂回归
  - 多类逻辑斯蒂回归
  - 贝叶斯逻辑斯蒂回归
    - 拉普拉斯近似

Q / A

- Any Questions...

考核方式

- **平时成绩 40%**
  - 平时作业 **3-5**次
    - 编程实验 或 计算与推导
- **期末成绩 60%**
  - 在指定列表中选择题目完成课程论文**(**算法实现**/**性能评价**/**性能比较**/**实验分析**)**
  - 评价标准
    - 论文内容的完整程度
    - 工作量、创新点…
    - 格式是否达标
      - 采用标准论文格式**(**鼓励使用**LaTeX)**

期末论文题目列表

- 要求:
  - 任选**1-2**种算法完成**1-2**个特定任务
  - 提交**1**份课程报告，内容覆盖但不限于:
    - 对选定问题的分析、算法选择、算法的基本原理、算法实现、算法实现中遇到的问题、性能评价与比较、实验分析
- 数据集:
  - **USPS/MNIST**数据集上的**2**个类别分类任务
  - **USPS/MNIST**数据集上的**10**个类别分类任务
  - **USPS/MNIST**数据集上的奇数**(13579)**与偶数**(02468)**的分类任务
  - **Scene 15** 数据集上**2**类别分类任务
  - **Scene 15** 数据集上**15**类别分类任务
  - **Caltech101**上**2**个类别的分类任务
  - **Caltech101**上**10**个类别的分类任务
  - **Caltech101**上**101**个类别的分类任务
  - **20 News Group**上**2**类别分类任务
  - **20 News Group**上**10**类别分类任务
  - **20 News Group**上**20**类别分类任务

期末论文题目列表

- 备选算法列表
  - 线性回归模型
  - 多项式回归模型
  - 正则化的线性回归
  - 正则化的多项式回归
  - 贝叶斯线性回归
  - 贝叶斯多项式回归
  - Fisher线性鉴别分析
  - 线性支持向量机模型
  - 非线性支持向量机模型
  - 二次鉴别(Quadratic Discriminant)函数
  - 2类别Logistic 回归模型
  - 多类别Logistic回归模型
  - 贝叶斯Logistic回归
  - 单层感知器模型
  - 多层感知器模型
  - …