机器学习

Machine learning

第三章 线性分类

Linear Classifier

授课人：周晓飞
zhouxiaofei@iie.ac.cn
2020-10-15

第三章 线性分类

3.1 概述

3.2 基础知识

3.3 感知机

3.4 线性鉴别分析

3.5 logistic 模型

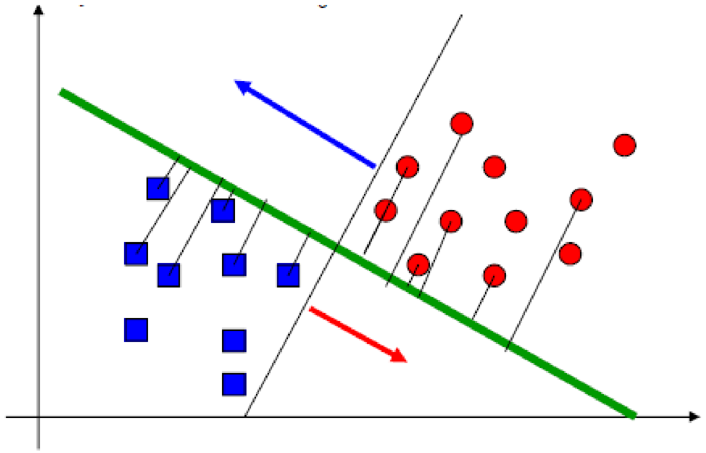
线性鉴别分析

基本思想

求线性变换

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

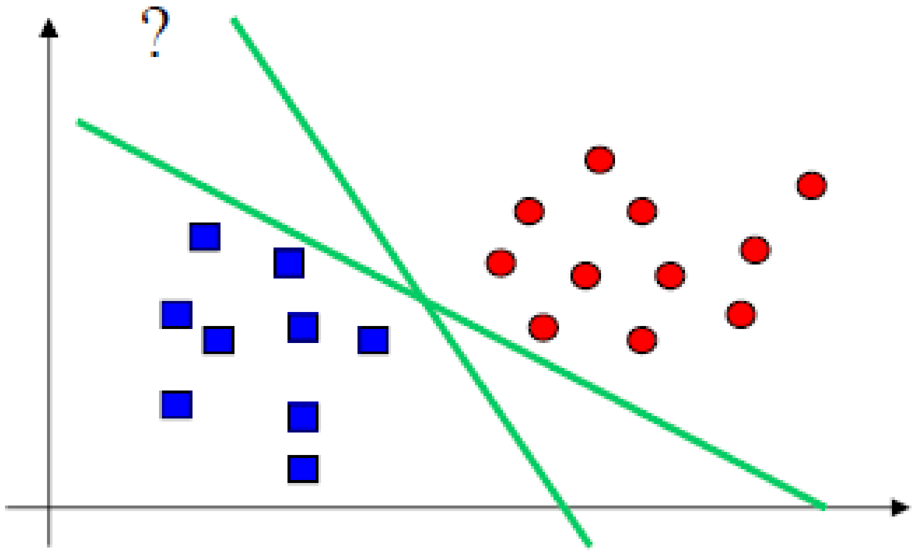
使得样本集$\{x_i\}$线性变换成一维变量$\{y_i\}$后，类别间距大，类内间距小，



线性鉴别分析

基本思想

怎么找到这个方向?

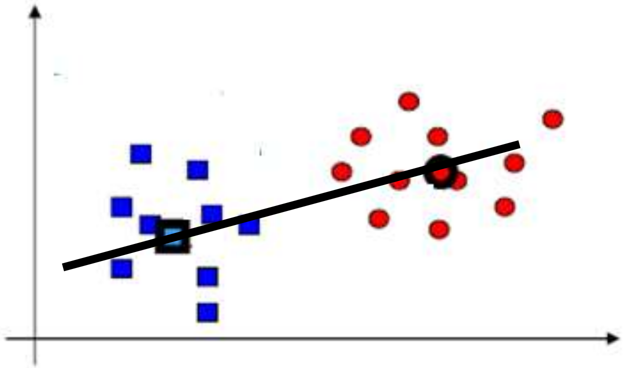


线性鉴别分析

基本思想

假设：　如果用各类的均值代表类别，类别间最大的方向

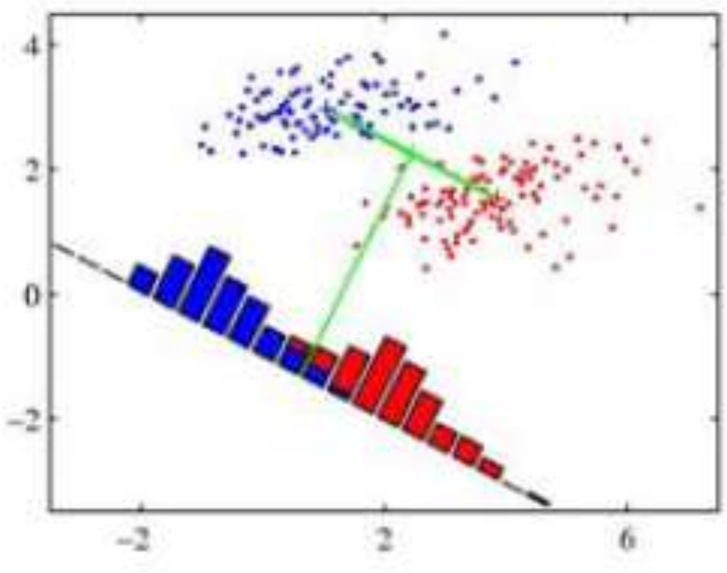
$$\boldsymbol{u}_1 = \frac{1}{N_1}\sum_{i\in C_1}^{N_1}\boldsymbol{x}_i, \qquad \boldsymbol{u}_2 = \frac{1}{N_2}\sum_{i\in C_2}^{N_2}\boldsymbol{x}_i$$

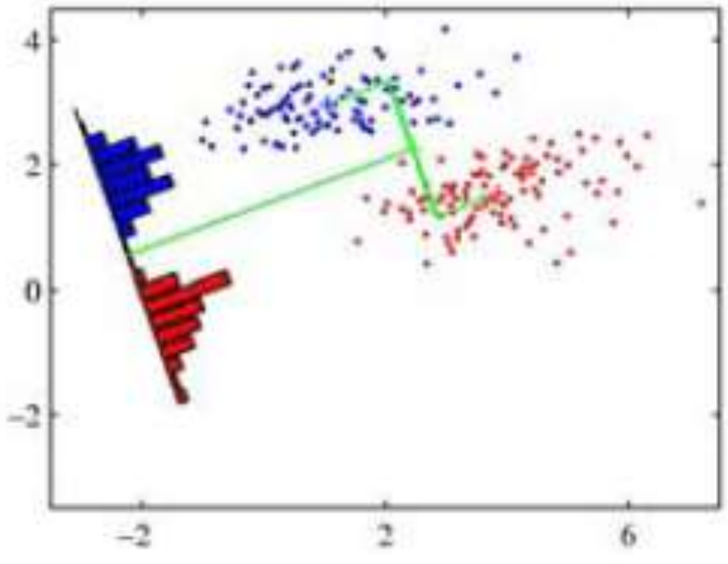


类间最大方向

基本思想

问题：只考虑类间，有可能线性不可分





目标函数（Fisher Criterion）

$$\max J(\boldsymbol{w}) = \frac{(\boldsymbol{m}_1 - \boldsymbol{m}_2)^2}{\boldsymbol{S}_1^2 + \boldsymbol{S}_2^2}$$

$$\boldsymbol{J} = \frac{\text{类别间距}}{\text{类内的平均距离}}$$

线性鉴别分析

类别间距离

样本投影后的类别间距离：$(\boldsymbol{m}_1 - \boldsymbol{m}_2)^2$； 其中，$\boldsymbol{m}_\mathrm{i}$表示第 i 类样本投影后的均值。

第 **k** 类样本平均值（类心）：

$$\boldsymbol{u}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \boldsymbol{x}_i$$

线性鉴别分析

类别间距离

样本投影后的类别间距离：$(m_1 - m_2)^2$； 其中，m_i表示第 i 类样本投影后的均值。

第 k 类样本平均值（类心）：

$$\boldsymbol{u}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \boldsymbol{x}_i$$

两个类别的类心：

$$\boldsymbol{u}_1 = \frac{1}{|C_1|} \sum_{\boldsymbol{x}_i \in C_1} \boldsymbol{x}_i \quad , \quad \boldsymbol{u}_2 = \frac{1}{|C_2|} \sum_{\boldsymbol{x}_i \in C_2} \boldsymbol{x}_i$$

线性鉴别分析

类别间距离

样本投影后的类别间距离：$(m_1 - m_2)^2$； 其中，m_i表示第 i 类样本投影后的均值。

样本 x_i 投影到 w 方向后，为 y_i： $y_i = \boldsymbol{w}^T \boldsymbol{x}_i$

投影后的类心：

$$
\begin{aligned}
m_k &= \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} y_i \\
&= \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \boldsymbol{w}^T \boldsymbol{x}_i \\
&= \boldsymbol{w}^T \Big(\frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \boldsymbol{x}_i \Big) \\
&= \boldsymbol{w}^T \boldsymbol{u}_k
\end{aligned}
$$

类别间距离

样本投影后的类别间距离：$(m_1 - m_2)^2$； 其中，m_i表示第 i 类样本投影后的均值。

- 投影后两类的类心：

$$m_1 = \mathbf{w}^T \mathbf{u}_1 \quad , \quad m_2 = \mathbf{w}^T \mathbf{u}_2$$

- $\mathbf{w}$ 方向投影后，类间距 ： $m_1 - m_2 = \mathbf{w}^T (\mathbf{u}_1 - \mathbf{u}_2)$

$$\begin{aligned}(m_1 - m_2)^2 &= (m_1 - m_2)(m_1 - m_2)^T \\ &= \mathbf{w}^T (\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T \mathbf{w} \\ &= \mathbf{w}^T S_b \mathbf{w}\end{aligned}$$

其中，类间散度矩阵：$S_b = (\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T$

若考虑先验可以定义： $S_b = p(\omega_1) p(\omega_2)(\mathbf{u}_1 - \mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)^T$

线性鉴别分析

类别内的距离

样本投影后的类别内距离：投影后的各类样本方差 $S_1{}^2 + S_2{}^2$

- 样本 x_i 投影到 w 方向后为 y_i： $y_i = \boldsymbol{w}^T \boldsymbol{x}_i$
- 在投影方向 w 上，第 k 类别内，样本距离

$$S_k{}^2 = \sum_{x_i \in C_k} (y_i - m_k)^2 = \sum_{x_i \in C_k} (\boldsymbol{w}^T (\boldsymbol{x}_i - \boldsymbol{u}_k))^2 = \sum_{x_i \in C_k} (\boldsymbol{w}^T \tilde{\boldsymbol{x}}_i)^2$$

$$= \sum_{x_i \in C_k} \left(\boldsymbol{w}^T \tilde{\boldsymbol{x}}_i\right)\left(\boldsymbol{w}^T \tilde{\boldsymbol{x}}_i\right)^T = \sum_{x_i \in C_k} \boldsymbol{w}^T \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i{}^T \boldsymbol{w} = \boldsymbol{w}^T \left(\sum_{x_i \in C_k} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i{}^T \right) \boldsymbol{w}$$

$$= \boldsymbol{w}^T (\boldsymbol{X}_{\mathrm{k}} \boldsymbol{X}_{\mathrm{k}}{}^T) \boldsymbol{w}$$

其中，$\boldsymbol{X}_k = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots \tilde{\boldsymbol{x}}_i, \ldots]_{C_k}$ 第 k 类样本矩阵（$\tilde{\boldsymbol{x}}_i$ 是列向量）

线性鉴别分析

类别内的距离

样本投影后的类别内距离：投影后的各类样本方差 $S_1^2+S_2^2$

- 在投影方向 $\boldsymbol{w}$ 上，类别内距离

$$S_1^2+S_2^2=\boldsymbol{w}^T(\boldsymbol{X}_1\boldsymbol{X}_1^T)\boldsymbol{w}+\boldsymbol{w}^T(\boldsymbol{X}_2\boldsymbol{X}_2^T)\boldsymbol{w}$$

$$=\boldsymbol{w}^T(\boldsymbol{X}_1\boldsymbol{X}_1^T+\boldsymbol{X}_2\boldsymbol{X}_2^T)\boldsymbol{w}$$

$$=\boldsymbol{w}^T S_w\boldsymbol{w}$$

其中，类内散度矩阵：

$$\boldsymbol{S}_w=\boldsymbol{X}_1\boldsymbol{X}_1^T+\boldsymbol{X}_2\boldsymbol{X}_2^T$$

若考虑先验可以定义： $\boldsymbol{S}_w=p(\omega_1)\boldsymbol{X}_1\boldsymbol{X}_1^T+p(\omega_2)\boldsymbol{X}_2\boldsymbol{X}_2^T$

求解过程

$$\max \mathrm{J}(\boldsymbol{w})=\frac{(\boldsymbol{m}_1-\boldsymbol{m}_2)^2}{S_1^2+S_2^2}=\frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}$$

- 广义的 Rayleigh 商，可用 Lagrange 乘子求解， 假设：$\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}=c$

$$L(\boldsymbol{w},\lambda)=\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}-\lambda(\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}-c)$$

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial \boldsymbol{w}}=2\boldsymbol{S}_b\boldsymbol{w}-2\lambda\boldsymbol{S}_w\boldsymbol{w}=0$$

$$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w}=\lambda\boldsymbol{w}$$

- 最优解 $\boldsymbol{w}$ 是 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 的特征向量

线性鉴别分析

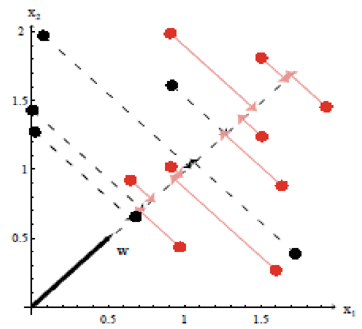
求解过程

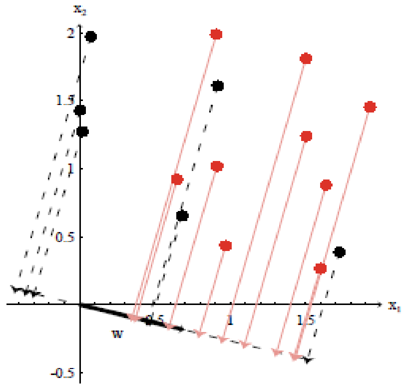
实际并没有求特征值，因为 $S_b w$ 在 u_1-u_2 方向上

$$S_b w=(u_1-u_2)(u_1-u_2)^T w=\beta(u_1-u_2)$$

$$S_w^{-1}S_b w=\lambda w \quad \Longrightarrow \quad S_w^{-1}\beta(u_1-u_2)=\lambda w$$

$$w=S_w^{-1}(u_1-u_2)$$





Have a break!

第三章 线性分类

3.1 概述

3.2 基础知识

3.3 感知机

3.4 线性鉴别分析

3.5 logistic 模型

logistic 模型

基本思想

假设 likelihood ratio 的对数为线性判别函数

$$\log\left(\frac{p(\boldsymbol{x}|\omega_i)}{p(\boldsymbol{x}|\omega_M)}\right)=\beta_{i,0}+\boldsymbol{\beta}_i^T\boldsymbol{x},\quad i=1,2,...,M\text{-}1$$

$$\Rightarrow\quad \log\left(\frac{p(\omega_i|\boldsymbol{x})}{p(\omega_M|\boldsymbol{x})}\right)=w_{i,0}+\boldsymbol{w}_i^T\boldsymbol{x},\quad i=1,2,...,M\text{-}1$$

logistic 模型

基本思想

多类问题

$$\ln\left(\frac{p(\omega_i \mid \boldsymbol{x})}{p(\omega_M \mid \boldsymbol{x})}\right) = w_{i,0} + \boldsymbol{w}_i^T \boldsymbol{x}, \quad i = 1, \ldots, M-1$$

$$\sum_{i=1}^{M} p(\omega_i \mid \boldsymbol{x}) = 1$$

$$\Rightarrow \begin{cases} p(\omega_M \mid \boldsymbol{x}) = \dfrac{1}{1 + \sum_{i=1}^{M-1} \exp(w_{i,0} + \boldsymbol{w}_i^T \boldsymbol{x})} & (1) \\ p(\omega_i \mid \boldsymbol{x}) = \dfrac{\exp(w_{i,0} + \boldsymbol{w}_i^T \boldsymbol{x})}{1 + \sum_{i=1}^{M-1} \exp(w_{i,0} + \boldsymbol{w}_i^T \boldsymbol{x})}, i = 1, \ldots, M-1 & (2) \end{cases}$$

logistic 模型

基本思想

两类问题：

$$\begin{cases} p(\omega_2 \mid \boldsymbol{x}) = \dfrac{1}{1+\exp(w_0+\boldsymbol{w}^T\boldsymbol{x})} \\ p(\omega_1 \mid \boldsymbol{x}) = \dfrac{\exp(w_0+\boldsymbol{w}^T\boldsymbol{x})}{1+\exp(w_0+\boldsymbol{w}^T\boldsymbol{x})} = \dfrac{1}{1+\exp(-(w_0+\boldsymbol{w}^T\boldsymbol{x}))} \end{cases}$$

令 $v=\boldsymbol{w}^T\boldsymbol{x}+w_0$，则

$$\begin{cases} p(\omega_2|\boldsymbol{x})=\dfrac{1}{1+\exp(v)} \\ p(\omega_1|\boldsymbol{x})=\dfrac{1}{1+\exp(-v)} \end{cases}$$

logistic 模型

基本思想

两类问题:

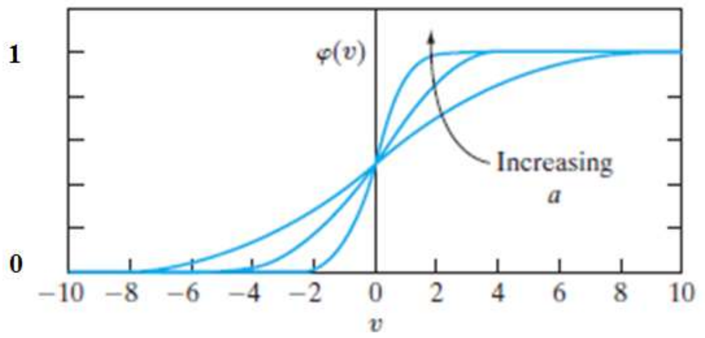
$$\begin{cases} p(\omega_2 \mid \mathbf{x}) = \dfrac{1}{1+\exp(w_0 + \mathbf{w}^T\mathbf{x})} \\ p(\omega_1 \mid \mathbf{x}) = \dfrac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1+\exp(w_0 + \mathbf{w}^T\mathbf{x})} = \dfrac{1}{1+\exp(-(w_0 + \mathbf{w}^T\mathbf{x}))} \end{cases}$$

令 $v=\mathbf{w}^T\mathbf{x}+w_0$，则

$$\begin{cases} p(\omega_2|\mathbf{x})=\dfrac{1}{1+\exp(v)} \\ p(\omega_1|\mathbf{x})=\dfrac{1}{1+\exp(-v)} \end{cases}$$

Logistic 函数（Sigmoid 函数）

$$\varphi(v) = \frac{1}{1+\exp(-av)}$$



logistic 模型

基本思想

两类问题:

$$
\begin{cases}
p(\omega_2 \mid \boldsymbol{x}) = \dfrac{1}{1+\exp(w_0 + \boldsymbol{w}^T \boldsymbol{x})} \\
p(\omega_1 \mid \boldsymbol{x}) = \dfrac{\exp(w_0 + \boldsymbol{w}^T \boldsymbol{x})}{1+\exp(w_0 + \boldsymbol{w}^T \boldsymbol{x})} = \dfrac{1}{1+\exp(-(w_0 + \boldsymbol{w}^T \boldsymbol{x}))}
\end{cases}
$$

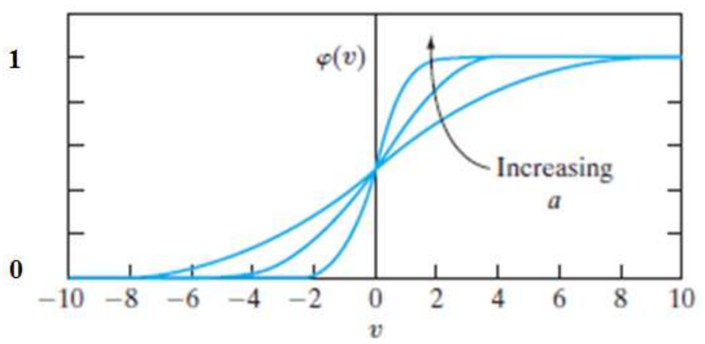
令 $v = \boldsymbol{w}^T \boldsymbol{x} + w_0$，则

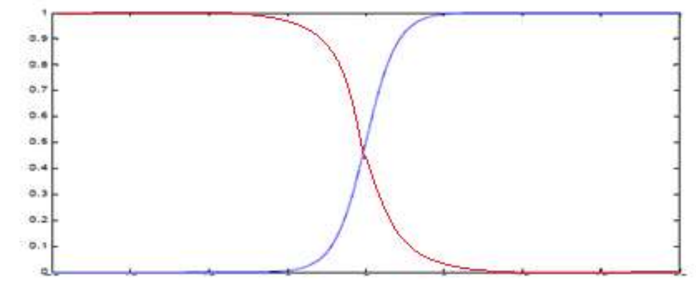
$$
\begin{cases}
p(\omega_2 \mid \boldsymbol{x}) = \dfrac{1}{1+\exp(v)} \\
p(\omega_1 \mid \boldsymbol{x}) = \dfrac{1}{1+\exp(-v)}
\end{cases}
$$

是两个对称函数

Logistic 函数（Sigmoid 函数）

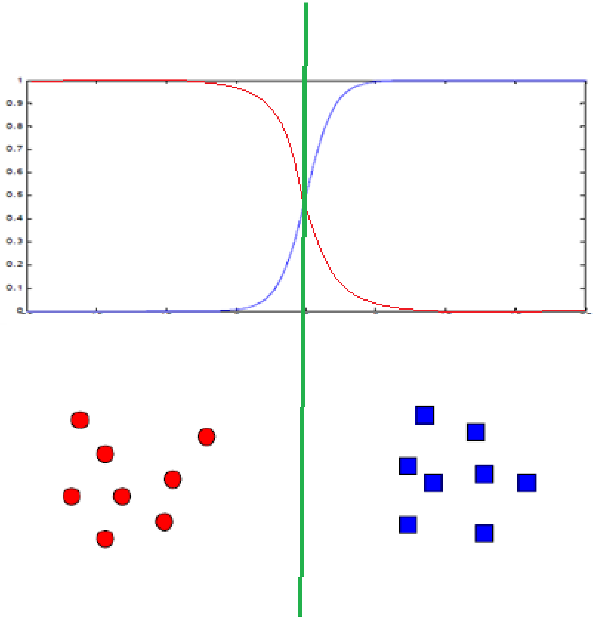
$$\varphi(v) = \frac{1}{1+\exp(-av)}$$





基本思想

求参数 $\boldsymbol{w}$ 和 w_0，相当于确定一个线性判别函数 $g(x)=\boldsymbol{w}^T\boldsymbol{x}+w_0$



$$\begin{cases} p(\omega_2 \mid \boldsymbol{x}) = \dfrac{1}{1+\exp(w_0+\boldsymbol{w}^T\boldsymbol{x})} \\ p(\omega_1 \mid \boldsymbol{x}) = \dfrac{\exp(w_0+\boldsymbol{w}^T\boldsymbol{x})}{1+\exp(w_0+\boldsymbol{w}^T\boldsymbol{x})} = \dfrac{1}{1+\exp(-(w_0+\boldsymbol{w}^T\boldsymbol{x}))} \end{cases}$$

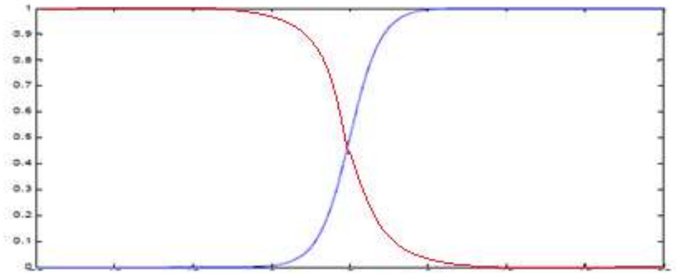
logistic 模型

学习过程

学习目标：

标签ω_1类的 x，$p(\omega_1|\mathbf{x})$越大，$p(\omega_2|\mathbf{x})$越小，

标签ω_2类的 x，$p(\omega_2|\mathbf{x})$越大，$p(\omega_1|\mathbf{x})$越小，



等价于

标签ω_1类的 x，$p(\omega_1|\mathbf{x})$越大，$1-p(\omega_1|\mathbf{x})$越小，

标签ω_1类的 x，$1-p(\omega_1|\mathbf{x})$越大，$p(\omega_1|\mathbf{x})$越小，

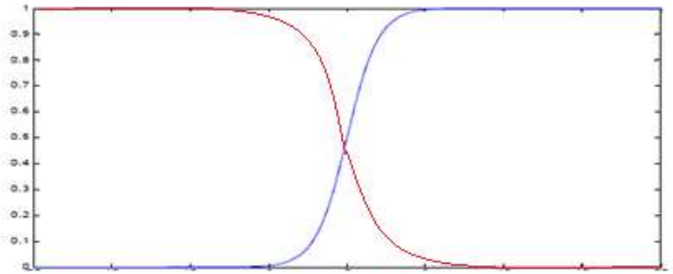
logistic 模型

学习过程

优化准则：

标签ω_1类的 x， $p(\omega_1|\mathbf{x})$越大

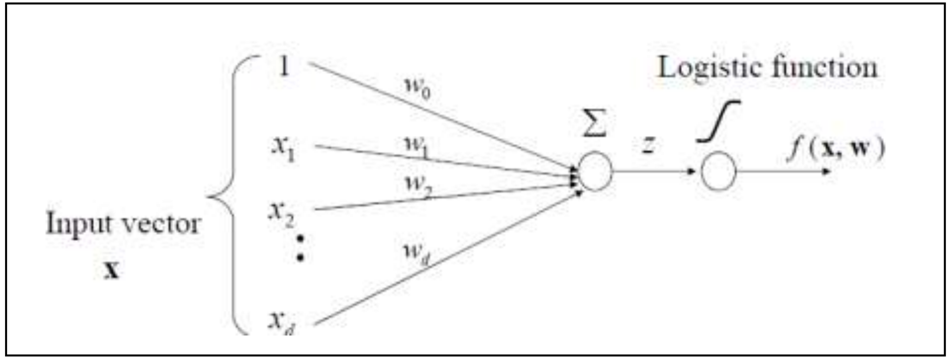
标签ω_2类的 x， $p(\omega_2|\mathbf{x})$越大



等价于

标签ω_1类的 x， $p(\omega_1|\mathbf{x})$越大

标签ω_1类的 x，$1-p(\omega_1|\mathbf{x})$越大



多分类的一般学习过程

最大似然求取参数 $\theta=\{w_i, w_{i,0}\}_{i=1,\ldots,M-1}$

$$L(\boldsymbol{\theta})=\ln\left\{\prod_{k=1}^{N_1} p(\boldsymbol{x}_k^{(1)}|\omega_1;\boldsymbol{\theta})\prod_{k=1}^{N_2} p(\boldsymbol{x}_k^{(2)}|\omega_2;\boldsymbol{\theta})\ldots\prod_{k=1}^{N_M} p(\boldsymbol{x}_k^{(M)}|\omega_M;\boldsymbol{\theta})\right\}$$

$$p(\boldsymbol{x}_k^{(m)}|\omega_m;\boldsymbol{\theta})=\frac{p(\boldsymbol{x}_k^{(m)})P(\omega_m|\boldsymbol{x}_k^{(m)};\boldsymbol{\theta})}{P(\omega_m)}$$

将(1)(2)带入后

$$L(\boldsymbol{\theta})=\sum_{k=1}^{N_1}\ln P(\omega_1|\boldsymbol{x}_k^{(1)})+\sum_{k=1}^{N_2}\ln P(\omega_2|\boldsymbol{x}_k^{(2)})+\ldots+\sum_{k=1}^{N_M}\ln P(\omega_M|\boldsymbol{x}_k^{(M)})+C$$

,

$$C=\ln\frac{\prod_{k=1}^{N} p(\boldsymbol{x}_k)}{\prod_{m=1}^{M} P(\omega_m)^{N_m}}$$

忽略先验的最大后验估计，就是最大似然估计

logistic 模型

学习过程

最大 $L(\theta)$ 问题转化为最小 $-L(\theta)$

求得 $\nabla L(\theta) = \frac{-\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$，采用梯度下降方法，

求解 $\boldsymbol{\theta} = \{\mathbf{w}_i, w_{i,0}\}_{i=1,\ldots,M-1}$； **m** 类与其他 **m-1** 类别的线性决策函数。

Have a break!

logistic 模型

模型理解

两类问题，

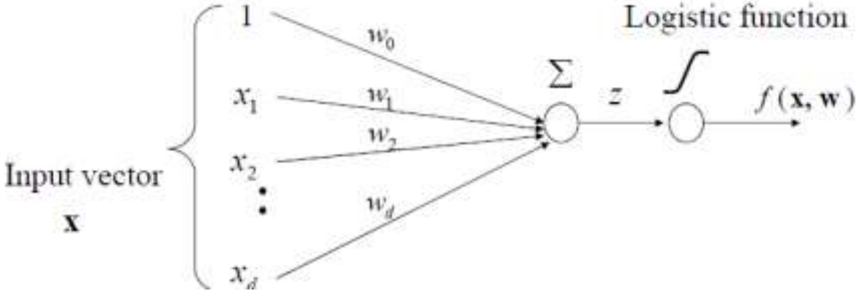
Discriminant functions:

$$g_1(\mathbf{x}) = g(\mathbf{w}^T\mathbf{x}) \qquad g_0(\mathbf{x}) = 1 - g(\mathbf{w}^T\mathbf{x})$$

Sigmoid function:

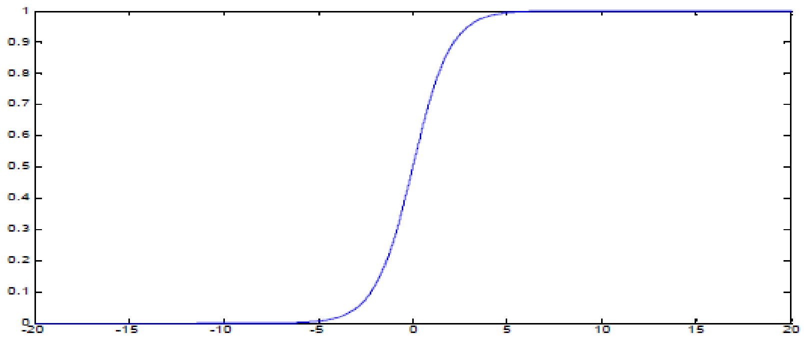
$$g(z) = 1/(1+e^{-z})$$

单层神经元、Logistic 激活函数



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

- Is also referred to as a **sigmoid function**
- Replaces the threshold function with smooth switching
- takes a real number and outputs the number in the interval [0,1]



logistic 模型

模型理解

– **Probabilistic interpretation**

$$f(\mathbf{x},\mathbf{w}) = p(y=1 \mid \mathbf{w},\mathbf{x}) = g_1(\mathbf{x}) = g(\mathbf{w}^T\mathbf{x})$$

$$p(y=0 \mid \mathbf{x},\mathbf{w}) = 1 - p(y=1 \mid \mathbf{x},\mathbf{w})$$

Decision boundary: $g_1(\mathbf{x}) = g_0(\mathbf{x})$

the boundary it must hold:

$$\log\frac{g_o(\mathbf{x})}{g_1(\mathbf{x})} = \log\frac{1-g(\mathbf{w}^T\mathbf{x})}{g(\mathbf{w}^T\mathbf{x})} = 0$$

logistic 模型

模型理解

线性决策界:

$$\log\frac{g_o(\mathbf{x})}{g_1(\mathbf{x})}=\log\frac{\dfrac{\exp-(\mathbf{w}^{\mathrm{T}}\mathbf{x})}{1+\exp-(\mathbf{w}^{\mathrm{T}}\mathbf{x})}}{\dfrac{1}{1+\exp-(\mathbf{w}^{\mathrm{T}}\mathbf{x})}}=\log\exp-(\mathbf{w}^{\mathrm{T}}\mathbf{x})=\mathbf{w}^{\mathrm{T}}\mathbf{x}=0$$

logistic 模型

简化公式的表达

$$p(y=1 \mid \boldsymbol{x}) = \frac{e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}} = \frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

$$p(y=0 \mid \boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}} = 1 - p(y=1 \mid \boldsymbol{x})$$

$$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1-y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})$$

或者 $$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1-y_i)(1-p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}))$$

最大化的似然估计：

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b)$$

小结

1. 掌握基础知识:

   线性模型的基本表达、向量相似计算、常用的统计量;

2.重点掌握线性分类模型：感知器、线性鉴别;

   了解logistic鉴别;

3. 掌握随机梯度下降优化方法.

参考文献

1. Pattern Recognition 2nd.《模式识别》(第二版)，边肇祺，张学工等,清华大学出版社，2000.1。
2. Pattern Classification, 2nd.模式分类，第二版。
3. 周志华，机器学习，清华大学出版社，2016.