南昌大学
NANCHANG UNIVERSITY

统计机器学习

主讲人：彭振华

数学与计算机学院

2026年

目录

CONTENTS

●多元线性回归模型

- 线性模型一般形式

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

$\boldsymbol{x} = (x_1; x_2; \ldots; x_d)$是由属性描述的示例，其中$x_i$是$\boldsymbol{x}$在第$i$个属性上的取值

- 向量形式

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$$

其中$\boldsymbol{w} = (w_1; w_2; \ldots; w_d)$

线性回归（linear regression）**目的**

- 学得一个线性模型以尽可能准确地预测实值输出标记

线 性 模 型

- 形式简单、易于建模
- 可解释性
- 非线性模型的基础
  - 引入层级结构或高维映射
- 一个例子
  - 综合考虑色泽、根蒂和敲声来判断西瓜好不好
  - 其中根蒂的系数最大，表明根蒂最要紧；而敲声的系数比色泽大，说明敲声比色泽更重要

$$f_{\text{好瓜}}(\boldsymbol{x}) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.3 \cdot x_{\text{敲声}} + 1$$

- 给定数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

其中 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ $y_i \in \mathbb{R}$

单一属性的线性回归目标

$$f(x_i) = wx_i + b \quad \text{使得} \quad f(x_i) \simeq y_i$$

参数/模型估计：最小二乘法（least square method）

$$\begin{aligned}(w^*, b^*) &= \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \\ &= \arg\min_{(w,b)} \sum_{i=1}^{m} (y_i - wx_i - b)^2\end{aligned}$$

$$E_{(w,b)} = \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

- 分别对 w 和 b 求导，可得

$$\frac{\partial E_{(w,b)}}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)\,x_i\right)$$

$$\frac{\partial E_{(w,b)}}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)$$

- 得到闭式（closed-form）解

$$w = \frac{\sum_{i=1}^{m} y_i\,(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2} \qquad b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i) \qquad \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

•多元线性回归

- 给定数据集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$

$\boldsymbol{x}_i=(x_{i1};x_{i2};\ldots;x_{id})$ $\quad y_i\in\mathbb{R}$

- 多元线性回归目标

$$f(\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b \text{ 使得} f(\boldsymbol{x}_i)\simeq y_i$$

- 把$\boldsymbol{w}$和 b 吸收入向量形式 $\hat{\boldsymbol{w}}=(\boldsymbol{w};b)$，数据集表示为

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix}=\begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1 \end{pmatrix} \qquad \boldsymbol{y}=(y_1;y_2;\ldots;y_m)$$

- 最小二乘法（least square method）

$$\hat{\boldsymbol{w}}^* = \arg\min_{\hat{w}} \left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)^{\mathrm{T}} \left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)$$

令 $E_{\hat{w}} = \left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)^{\mathrm{T}} \left(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}}\right)$ ，对 $\hat{\boldsymbol{w}}$ 求导得到

$$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}} = 2\mathbf{X}^{\mathrm{T}} \left(\mathbf{X}\hat{\boldsymbol{w}} - \boldsymbol{y}\right)$$

令上式为零可得 $\hat{w}$ 最优解的闭式解： $\mathbf{X}^T\mathbf{X}\hat{w} = \mathbf{X}^T\boldsymbol{y}$

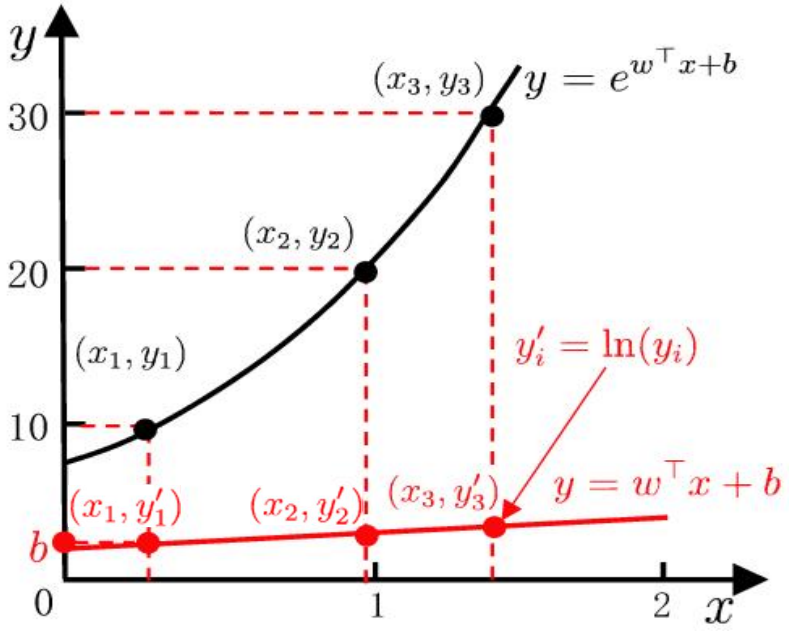
- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 是满秩矩阵或正定矩阵，则 $\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$

其中 $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ 是 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 的逆矩阵，线性回归模型为

$$f\left(\hat{\boldsymbol{x}}_i\right) = \hat{\boldsymbol{x}}_i^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$$

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 不是满秩矩阵
  - 根据归纳偏好选择解
  - 引入正则化

- 输出标记的对数为线性模型逼近的目标



- 广义线性模型形式

$$y = g^{-1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right)$$

- $g(\cdot)$ 称为联系函数（link function）
  - 单调可微函数
- 对数线性回归是 $g(\cdot) = \ln(\cdot)$ 时广义线性模型的特例

$$\ln y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$\Updownarrow$$

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

•二分类任务

- 预测值与输出标记 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ $\quad\quad$ $y \in \{0, 1\}$
- 寻找函数将分类标记与线性回归模型输出联系起来
- 最理想的函数：单位阶跃函数

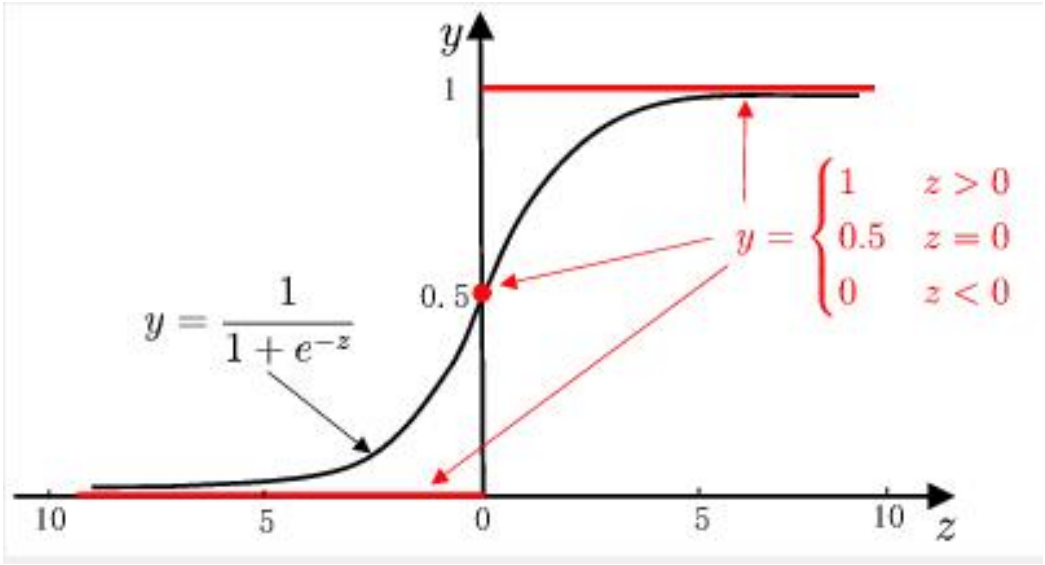
$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

- 预测值大于零就判为正例，小于零则判为反例，预测值为临界值零则可任意判别

- 单位阶跃函数缺点
  - 不连续
- 替代函数：对数几率函数（logistic function）
  - 单调可微、任意阶可导

$$y = \frac{1}{1 + e^{-z}}$$

单位阶跃函数与对数几率函数的比较



- 运用对数几率函数

$$y = \frac{1}{1+e^{-z}} \quad \text{变为} \quad y = \frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

- 对数几率（log odds / logit）
  - 样本作为正例的相对可能性的对数

$$\ln\frac{y}{1-y} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

- 对数几率回归优点
  - 无需事先假设数据分布
  - 可得到“类别”的近似概率预测
  - 可直接应用现有数值优化算法求取最优解

- 对数几率

$$\ln\frac{y}{1-y} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \xrightarrow{y\sim p(y=1|\boldsymbol{x})} \ln\frac{p\left(y=1\mid\boldsymbol{x}\right)}{p\left(y=0\mid\boldsymbol{x}\right)} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$$

显然有

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

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

- 对数似然函数

$$L(w)=\sum_{i=1}^{N}[y_i\log\pi(x_i)+(1-y_i)\log(1-\pi(x_i))]$$

$$=\sum_{i=1}^{N}\left[y_i\log\frac{\pi(x_i)}{1-\pi(x_i)}+\log(1-\pi(x_i))\right]$$

$$=\sum_{i=1}^{N}[y_i(w\cdot x_i)-\log(1+\exp(w\cdot x_i)]$$

- 对L(w)求极大值，得到w的估计值。

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{m}\left(-y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i+\ln\left(1+e^{\beta^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right)$$

- 通常采用梯度下降法及拟牛顿法，学到的模型：

$$P(Y=1|x)=\frac{\exp(\hat{w}\cdot x)}{1+\exp(\hat{w}\cdot x)}\qquad P(Y=0|x)=\frac{1}{1+\exp(\hat{w}\cdot x)}$$

作业二

1. 设计逻辑回归的梯度下降算法

2. 函数是否为凸函数，为什么?

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i + \ln\left(1 + e^{\beta^{\mathrm{T}} \hat{\boldsymbol{x}}_i}\right) \right)$$

3. 考虑如下数据集，试建立最小二乘模型，并设计对应的牛顿算法

$$\left((1,2),3\right),\left((2,1),2\right),\left((3,3),5\right)$$

- 求解得

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

- 牛顿法第t+1轮迭代解的更新公式

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

其中关于β 的一阶、二阶导数分别为

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \left(y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) \right)$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}} p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) \left(1 - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) \right)$$

高阶可导连续凸函数，梯度下降法/牛顿法 [Boyd and Vandenberghe, 2004]

➢ 逻辑回归与最大熵模型

最大熵模型(Maximum Entropy Model)由最大熵原理推导实现。

最大熵原理:

学习概率模型时，在所有可能的概率模型(分布)中，熵最大的模型是最好的模型，表述为在满足约束条件的模型集合中选取熵最大的模型。

假设离散随机变量X的概率分布是P(X)，

熵:

$$H(P)=-\sum_{x} P(x)\log P(x)$$

$$0 \leqslant H(P) \leqslant \log |X|$$

且:

|X|是X的取值个数，X均匀分布时右边等号成立。

➢ 逻辑回归与最大熵模型

例题： 假设随机变量X有5个取值{A,B,C,D,E},估计各个值的概率。

解：满足 $P(A)+P(B)+P(C)+P(D)+P(E)=1$

等概率估计：$$P(A)=P(B)=P(C)=P(D)=P(E)=\frac{1}{5}$$

加入一些先验：$$P(A)+P(B)=\frac{3}{10}$$

$$P(A)+P(B)+P(C)+P(D)+P(E)=1$$

于是：$$P(A)=P(B)=\frac{3}{20}$$

$$P(C)=P(D)=P(E)=\frac{7}{30}$$

再加入约束：$$P(A)+P(C)=\frac{1}{2}$$

$$P(A)+P(B)=\frac{3}{10}$$

$$P(A)+P(B)+P(C)+P(D)+P(E)=1$$

➢ 逻辑回归与最大熵模型

X和Y分别是输入和输出的集合，这个模型表示的是对于给定的输入X，以条件概率P(Y|X)输出Y.

给定数据集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

联合分布P(Y|X)的经验分布，边缘分布P(X)的经验分布：

$$\tilde{P}(X,Y) \Rightarrow \tilde{P}(X=x,Y=y)=\frac{\nu(X=x,Y=y)}{N}$$

$$\tilde{P}(X) \Rightarrow \tilde{P}(X=x)=\frac{\nu(X=x)}{N}$$

特征函数：

$$f(x,y)=\begin{cases}1, & x与y满足某一事实\\ 0, & 否则\end{cases}$$

➢ 逻辑回归与最大熵模型

特征函数f(x,y)关于经验分布 $\tilde{P}(X,Y)$ 的期望值:

$$E_{\tilde{p}}(f)=\sum_{x,y}\tilde{P}(x,y)f(x,y)$$

特征函数f(x,y)关于模型P(Y|X)与经验分布$\tilde{P}(X)$的期望值:

$$E_{P}(f)=\sum_{x,y}\tilde{P}(x)P(y|x)f(x,y)$$

如果模型能够获取训练数据中的信息，那么就可以假设这两个期望值相等，即

$$E_{P}(f)=E_{\tilde{p}}(f) \Rightarrow \sum_{x,y}\tilde{P}(x)P(y|x)f(x,y)=\sum_{x,y}\tilde{P}(x,y)f(x,y)$$

假设有n个特征函数:

$$f_i(x,y),\quad i=1,2,\cdots,n$$

➢ 逻辑回归与最大熵模型

定义:

假设满足所有约束条件的模型集合为:

$$C \equiv \{P \in \mathcal{P} \mid E_P(f_i) = E_{\tilde{P}}(f_i),\ \ i = 1,2,\cdots,n\}$$

定义在条件概率分布P(Y | X)上的条件熵:

$$H(P) = -\sum_{x,y} \tilde{P}(x)P(y \mid x)\log P(y \mid x)$$

则模型集合C中条件熵H(P)最大的模型称为最大熵模型

➢ 逻辑回归与最大熵模型

最大熵模型的学习可以形式化为约束最优化问题。

对于给定的数据集以及特征函数： $f_i(x,y)$

最大熵模型的学习等价于约束最优化问题:

$$\max_{P\in C} \quad H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

$$\text{s.t.} \quad E_P(f_i)=E_{\tilde{P}}(f_i),\quad i=1,2,\cdots,n$$

$$\sum_y P(y|x)=1$$

$$\min_{P\in C} \quad -H(P)=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

$$\text{s.t.} \quad E_P(f_i)-E_{\tilde{P}}(f_i)=0,\quad i=1,2,\cdots,n$$

$$\sum_y P(y|x)=1$$

➢ 逻辑回归与最大熵模型

这里，将约束最优化的原始问题转换为无约束最优化的对偶问题,通过求解对偶问题求解原始间题:

引进拉格朗日乘子，定义拉格朗日函数:

$$L(P,w) \equiv -H(P) + w_0\left(1-\sum_y P(y|x)\right) + \sum_{i=1}^{n} w_i\left(E_{\tilde{p}}(f_i) - E_p(f_i)\right)$$

$$= \sum_{x,y} \tilde{P}(x)P(y|x)\log P(y|x) + w_0\left(1-\sum_y P(y|x)\right)$$

$$+ \sum_{i=1}^{n} w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y) - \sum_{x,y}\tilde{P}(x)P(y|x)f_i(x,y)\right)$$

最优化原始问题 到 对偶问题:

$$\min_{P\in C}\max_{w} L(P,w) \Rightarrow \max_{w}\min_{P\in C} L(P,w)$$

➢ 逻辑回归与最大熵模型

最优化原始问题 到 对偶问题：

$$\min_{P\in C}\max_{w} L(P,w) \Rightarrow \max_{w}\min_{P\in C} L(P,w)$$

L(P,w)是P的凸函数，解的等价性（证明部分在SVM部分介绍）

先求极小化问题：$\min_{P\in C} L(P,w)$ 是w的函数，

$$\Psi(w)=\min_{P\in C} L(P,w)=L(P_w,w)$$

$$P_w=\arg\min_{P\in C} L(P,w)=P_w(y\mid x)$$

➢ 逻辑回归与最大熵模型

求L(P,w)对P(y | x)的偏导数：

$$\frac{\partial L(P,w)}{\partial P(y|x)} = \sum_{x,y}\tilde{P}(x)\left(\log P(y|x)+1\right)-\sum_{y}w_0-\sum_{x,y}\left(\tilde{P}(x)\sum_{i=1}^{n}w_i f_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x)\left(\log P(y|x)+1-w_0-\sum_{i=1}^{n}w_i f_i(x,y)\right)$$

得：

$$P(y|x)=\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)+w_0-1\right)=\frac{\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)}{\exp(1-w_0)}$$

➢ 逻辑回归与最大熵模型

由：$$\sum_{y} P(y|x)=1$$

得：$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right) \quad (6.22)$$

规范化因子：$$Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right) \quad (6.23)$$

模型 $P_w = P_w(y|x)$ 就是最大熵模型

求解对偶问题外部的极大化问题：

$$\max_{w} \Psi(w) \qquad w^* = \arg\max_{w} \Psi(w) \qquad P^* = P_{w^*} = P_{w^*}(y|x)$$

➢ 逻辑回归与最大熵模型

例题： 原例子中的最大熵模型 $\min \ -H(P)=\sum_{i=1}^{5}P(y_i)\log P(y_i)$

$$\text{s.t.} \quad P(y_1)+P(y_2)=\tilde{P}(y_1)+\tilde{P}(y_2)=\frac{3}{10}$$

$$\sum_{i=1}^{5}P(y_i)=\sum_{i=1}^{5}\tilde{P}(y_i)=1$$

$$L(P,w)=\sum_{i=1}^{5}P(y_i)\log P(y_i)+w_1\left(P(y_1)+P(y_2)-\frac{3}{10}\right)+w_0\left(\sum_{i=1}^{5}P(y_i)-1\right)$$

$$\max_{w}\min_{P}L(P,w)$$

➢ 逻辑回归与最大熵模型

$$\frac{\partial L(P,w)}{\partial P(y_1)} = 1 + \log P(y_1) + w_1 + w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_2)} = 1 + \log P(y_2) + w_1 + w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_3)} = 1 + \log P(y_3) + w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_4)} = 1 + \log P(y_4) + w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_5)} = 1 + \log P(y_5) + w_0$$

解得：

$$P(y_1) = P(y_2) = e^{-w_1 - w_0 - 1}$$

$$P(y_3) = P(y_4) = P(y_5) = e^{-w_0 - 1}$$

➢ 逻辑回归与最大熵模型

$$\min_{P} L(P,w) = L(P_w,w) = -2e^{-w_1-w_0-1} - 3e^{-w_0-1} - \frac{3}{10}w_1 - w_0$$

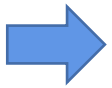
得：

$$\max_{w} L(P_w,w) = -2e^{-w_1-w_0-1} - 3e^{-w_0-1} - \frac{3}{10}w_1 - w_0$$

对wi求偏导并令为0：

$$e^{-w_1-w_0-1} = \frac{3}{20}$$

$$e^{-w_0-1} = \frac{7}{30}$$



$$P(y_1) = P(y_2) = \frac{3}{20}$$

$$P(y_3) = P(y_4) = P(y_5) = \frac{7}{30}$$

➢ 逻辑回归与最大熵模型

最大熵模型就是(6.22),(6.23)表示的条件概率分布，

证明：对偶函数的极大化等价于最大熵模型的极大似然估计.

已知训练数据的经验概率分布 $\tilde{P}(X,Y)$，条件概率分布P(Y|X)的对数似然函数表示为：

$$L_{\tilde{p}}(P_w)=\log\prod_{x,y}P(y|x)^{\tilde{P}(x,y)}=\sum_{x,y}\tilde{P}(x,y)\log P(y|x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-\sum_{x,y}\tilde{P}(x,y)\log Z_w(x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x)$$

➢ 逻辑回归与最大熵模型

而：

$$\Psi(w)=\sum_{x,y}\tilde{P}(x)P_w(y|x)\log P_w(y|x)$$

$$+\sum_{i=1}^{n}w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P_w(y|x)f_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)+\sum_{x,y}\tilde{P}(x)P_w(y|x)\left(\log P_w(y|x)-\sum_{i=1}^{n}w_if_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)-\sum_{-x,y}\tilde{P}(x)P_w(y|x)\log Z_w(x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x)$$

➢ 逻辑回归与最大熵模型

- 最大熵模型与逻辑斯谛回归模型有类似的形式，它们又称为对数线性模型(log linear model). 模型学习就是在给定的训练数据条件下对模型进行极大似然估计或正则化的极大似然估计。
- 逻辑斯谛回归模型、最大熵模型学习归结为以似然函数为目标函数的最优化问题，通常通过迭代算法求解，它是光滑的凸函数，因此多种最优化的方法都适用。
- 常用的方法有：

  改进的迭代尺度法

  梯度下降法

  牛顿法

  拟牛顿法

➢ 逻辑回归与最大熵模型

改进的迭代尺度法(improved iterative scaling,IIS)

由最大熵模型

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right) \qquad Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)$$

对数似然函数

$$L(w)=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x)$$

求对数似然函数的极大值 $\hat{w}$

IIS思路：假设 $w=(w_1,w_2,\cdots,w_n)^{\mathrm{T}}$ 希望找到一个新的参数向量 $w+\delta=(w_1+\delta_1,w_2+\delta_2,\cdots,w_n+\delta_n)^{\mathrm{T}}$，使得模型的对数似然函数值增大，如果有参数向量更新方法，那么就可以重复使用这一方法，直至找到对数似然函数的最大值。

➢ 逻辑回归与最大熵模型

$$L(w+\delta)-L(w)=\sum_{x,y}\tilde{P}(x,y)\log P_{w+\delta}(y\mid x)-\sum_{x,y}\tilde{P}(x,y)\log P_w(y\mid x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)-\sum_{x}\tilde{P}(x)\log\frac{Z_{w+\delta}(x)}{Z_w(x)}$$

利用 $-\log\alpha\geqslant 1-\alpha,\quad \alpha>0$

$$L(w+\delta)-L(w)\geqslant\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\tilde{P}(x)\frac{Z_{w+\delta}(x)}{Z_w(x)}$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)\exp\sum_{i=1}^{n}\delta_i f_i(x,y)$$

$$A(\delta\mid w)=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)\exp\sum_{i=1}^{n}\delta_i f_i(x,y)$$

➢ 逻辑回归与最大熵模型

于是有 $$L(w+\delta)-L(w)\geqslant A(\delta|w)$$

如果能找到适当的δ使下界A(δ|w)提高，那么对数似然函数也会提高。

δ是一个向量，含多个变量，一次只优化一个变量δ_i

引进一个量$f^{\#}(x,y)$,

$$f^{\#}(x,y)=\sum_i f_i(x,y)$$

$f_i(x,y)$是二值函数， $f^{\#}(x,y)$表示所有特征在(x,y)出现的次数。

$$A(\delta|w)=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_x\tilde{P}(x)\sum_y P_w(y|x)\exp\left(f^{\#}(x,y)\sum_{i=1}^{n}\frac{\delta_i f_i(x,y)}{f^{\#}(x,y)}\right)$$

➢ 逻辑回归与最大熵模型

利用指数函数的凸性，以及 $\frac{f_i(x,y)}{f^{\#}(x,y)} \geqslant 0$ 且 $\sum_{i=1}^{n} \frac{f_i(x,y)}{f^{\#}(x,y)} = 1$

根据Jensen不等式：

$$\exp\left(\sum_{i=1}^{n} \frac{f_i(x,y)}{f^{\#}(x,y)} \delta_i f^{\#}(x,y)\right) \leqslant \sum_{i=1}^{n} \frac{f_i(x,y)}{f^{\#}(x,y)} \exp(\delta_i f^{\#}(x,y))$$

$$A(\delta \mid w) \geqslant \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} \delta_i f_i(x,y) + 1 - \sum_{x} \tilde{P}(x) \sum_{y} P_w(y \mid x) \sum_{i=1}^{n} \left(\frac{f_i(x,y)}{f^{\#}(x,y)}\right) \exp(\delta_i f^{\#}(x,y))$$

$$B(\delta \mid w) = \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} \delta_i f_i(x,y) + 1 - \sum_{x} \tilde{P}(x) \sum_{y} P_w(y \mid x) \sum_{i=1}^{n} \left(\frac{f_i(x,y)}{f^{\#}(x,y)}\right) \exp(\delta_i f^{\#}(x,y))$$

➢ 逻辑回归与最大熵模型

于是得到 $L(w+\delta)-L(w)\geqslant B(\delta|w)$

$B(\delta|w)$ 是对数似然函数改变量的一个新的下界

对δ_i 求偏导：

$$\frac{\partial B(\delta|w)}{\partial \delta_i}=\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x}\tilde{P}(x)\sum_{y}P_w(y|x)f_i(x,y)\exp(\delta_i f^{\#}(x,y))$$

令偏导数为0，得到：

$$\sum_{x,y}\tilde{P}(x)P_w(y|x)f_i(x,y)\exp(\delta_i f^{\#}(x,y))=E_{\tilde{p}}(f_i)$$

依次对δ_i 解方程。

➢ 逻辑回归与最大熵模型

算法

输入： 特征函数f1,f2…fn;经验分布 $\tilde{P}(X,Y)$,模型 $P_w(y|x)$

输出： 最优参数w_i*； 最优模型P_{w*}

（1）对所有$i\in\{1,2,\cdots,n\}$，取初值$w_i=0$

（2）对每一$i\in\{1,2,\cdots,n\}$：

（a）令δ_i是方程

$$\sum_{x,y}\tilde{P}(x)P(y|x)f_i(x,y)\exp(\delta_i f^{\#}(x,y))=E_{\tilde{p}}(f_i)$$

的解，这里 $f^{\#}(x,y)=\sum_{i=1}^{n}f_i(x,y)$

（b）更新w_i值：$w_i\leftarrow w_i+\delta_i$ 关键

（3）如果不是所有w_i都收敛，重复步 (2)

➢ 逻辑回归与最大熵模型

如果 $f^{\#}(x,y)$ 是常数 M

$$\delta_i = \frac{1}{M}\log\frac{E_{\tilde{P}}(f_i)}{E_P(f_i)}$$

如果 $f^{\#}(x,y)$ 不是常数　牛顿法

$$\sum_{x,y}\tilde{P}(x)P_w(y\mid x)f_i(x,y)\exp(\delta_i f^{\#}(x,y)) = E_{\tilde{P}}(f_i)$$

$$g(\delta_i) = 0$$

$$\delta_i^{(k+1)} = \delta_i^{(k)} - \frac{g(\delta_i^{(k)})}{g'(\delta_i^{(k)})}$$

➢ 逻辑回归与最大熵模型

最大熵模型：

$$P_w(y|x)=\frac{\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)}{\sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)}$$

目标函数：

$$\min_{w\in\mathbf{R}^n} f(w)=\sum_{x}\tilde{P}(x)\log\sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)-\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y)$$

梯度：

$$g(w)=\left(\frac{\partial f(w)}{\partial w_1},\frac{\partial f(w)}{\partial w_2},\cdots,\frac{\partial f(w)}{\partial w_n}\right)^{\mathrm{T}}$$

$$\frac{\partial f(w)}{\partial w_i}=\sum_{x,y}\tilde{P}(x)P_w(y|x)f_i(x,y)-E_{\tilde{p}}(f_i),\quad i=1,2,\cdots,n$$

➢ 逻辑回归与最大熵模型

输入：特征函数 $f_1, f_2, \cdots, f_n$；经验分布 $\tilde{P}(x,y)$

目标函数 $f(w)$，梯度 $g(w)=\nabla f(w)$，精度要求 ε

输出：最优参数值 w^*；最优模型 $P_{w^*}(y|x)$.

（1）选定初始点 $w^{(0)}$，取 B_0 为正定对称矩阵，置 $k=0$

（2）计算 $g_k=g(w^{(k)})$. 若 $\|g_k\|<\varepsilon$，则停止计算，

得 $w^*=w^{(k)}$；否则转 (3)

（3）由 $B_k p_k=-g_k$ 求出 p_k

（4）一维搜索：求 λ_k 使得

$$f(w^{(k)}+\lambda_k p_k)=\min_{\lambda\geq 0} f(w^{(k)}+\lambda p_k)$$

➢ 逻辑回归与最大熵模型

（5）置 $w^{(k+1)} = w^{(k)} + \lambda_k p_k$

（6）计算 $g_{k+1} = g(w^{(k+1)})$，若 $\| g_{k+1} \| < \varepsilon$，则停止计算

得 $w^* = w^{(k+1)}$；否则按下式求出 B_{k+1}：

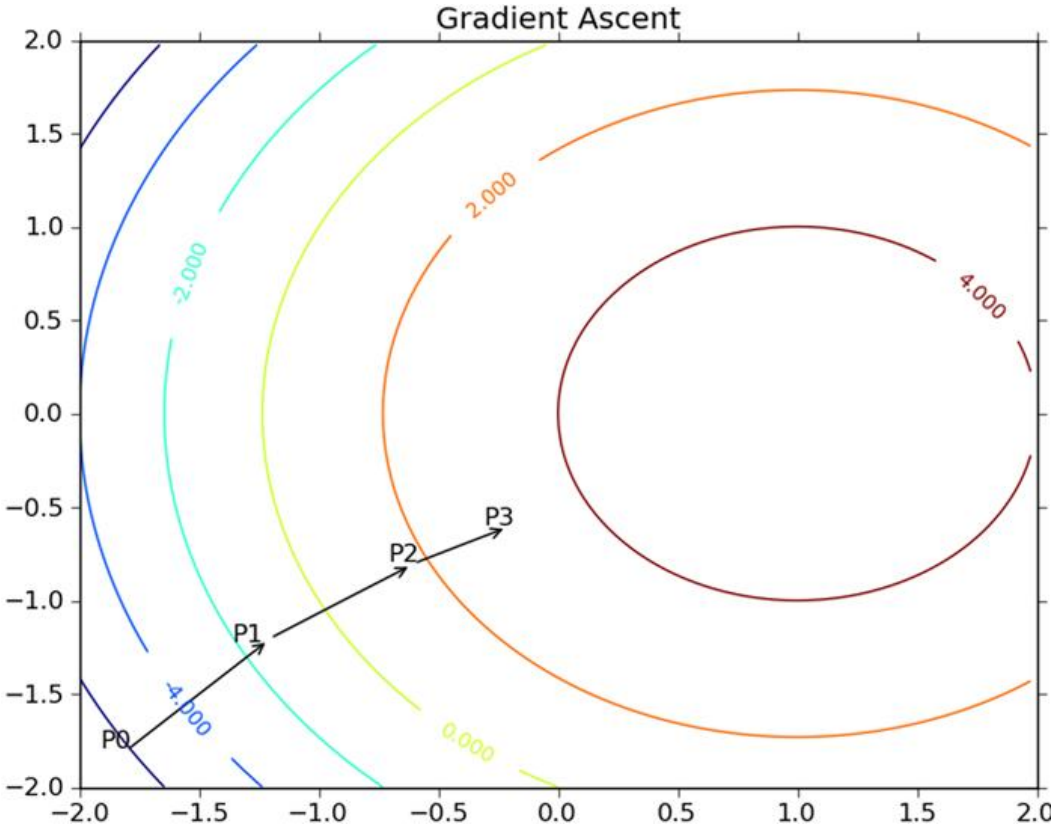
$$B_{k+1} = B_k + \frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}} \delta_k} - \frac{B_k \delta_k \delta_k^{\mathrm{T}} B_k}{\delta_k^{\mathrm{T}} B_k \delta_k}$$

$$y_k = g_{k+1} - g_k, \quad \delta_k = w^{(k+1)} - w^{(k)}$$

（7）置 $k = k+1$，转 (3)

➢ 逻辑回归与最大熵模型

要找到某函数的最大值，最好的方法是沿着该函数的梯度方向探寻。



➢ 逻辑回归与最大熵模型

函数f(x,y)的梯度：前提是函数f(x,y)必须在待计算的点上有定义并且可微。

$$\nabla f(x,y)=\begin{pmatrix}\dfrac{\partial f(x,y)}{\partial x}\\ \dfrac{\partial f(x,y)}{\partial y}\end{pmatrix}$$

$$w:=w+\alpha\nabla_w f(w)$$

若移动的步长记作a

迭代停止条件：

迭代次数达到某个指定的值

算法达到某个可以允许的误差范围

➢ 逻辑回归与最大熵模型

即求最大值：

$$\sum_{i=1}^{m}\left(y_i \bullet x_i - \ln(1+e^{x_i})\right)$$

由：

$$x_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots + w_n x_{in}$$

且：

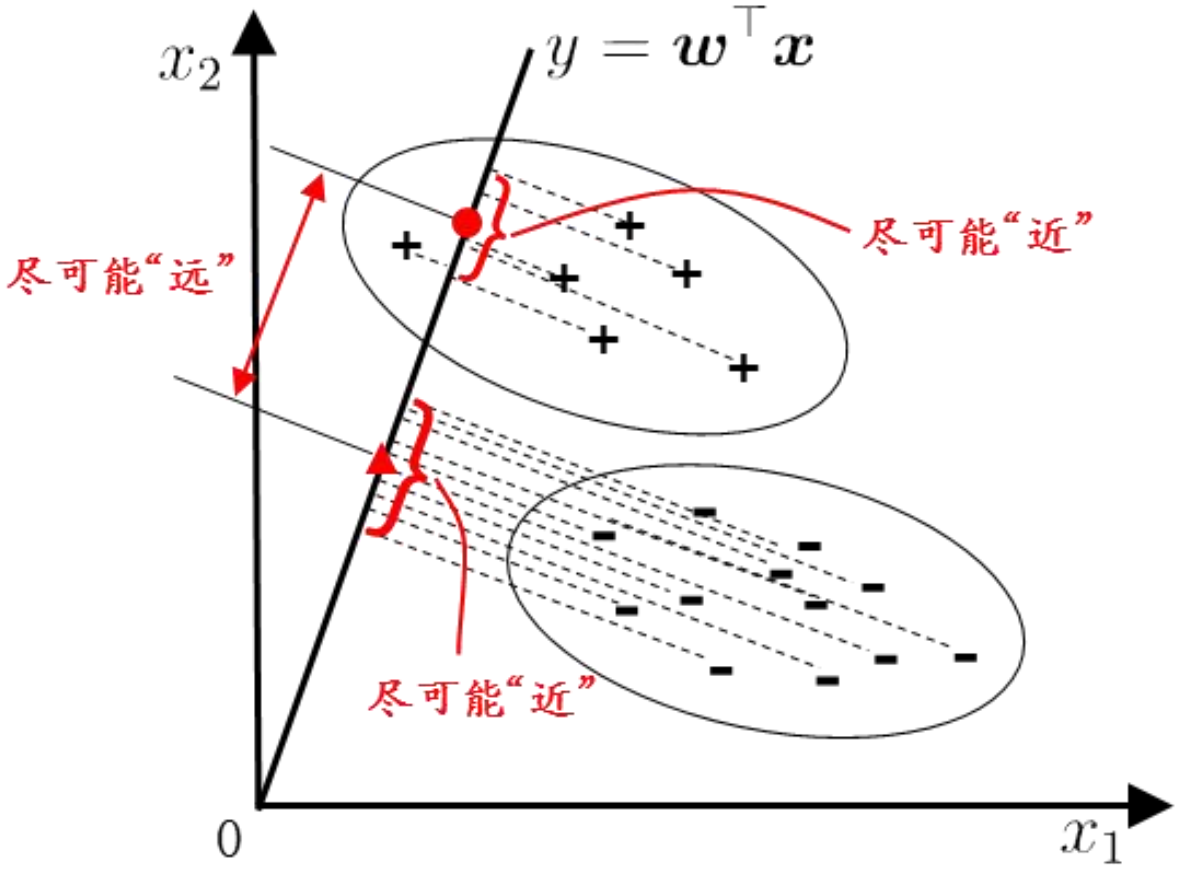
$$w_k = w_k + \alpha \frac{\partial \ln L(w)}{\partial w_k}$$

则：

$$w_k = w_k + \alpha \sum_{i=1}^{m} x_{ik}\left[y_i - \pi(x_i)\right]$$

●线性判别分析（Linear Discriminant Analysis）

[Fisher, 1936]



LDA也可被视为一种监督降维技术

- LDA的思想
  - 欲使同类样例的投影点尽可能接近，可以让同类样例投影点的协方差尽可能小
  - 欲使异类样例的投影点尽可能远离，可以让不同类中心之间的距离尽可能大
- 一些变量
  - 第i类示例的集合 X_i
  - 第i类示例的均值向量 $\boldsymbol{\mu}_i$
  - 第i类示例的协方差矩阵 $\boldsymbol{\Sigma}_i = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$
  - 两类样本的中心在直线上的投影： $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$
  - 两类样本投影点的协方差： $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$

- 最大化目标

$$J = \frac{\left\| \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_0 - \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\mu}_1 \right\|_2^2}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Sigma}_0 \boldsymbol{w} + \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\Sigma}_1 \boldsymbol{w}}$$
$$= \frac{\boldsymbol{w}^{\mathrm{T}} \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^{\mathrm{T}} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 \right) \boldsymbol{w}}$$

- 类内散度矩阵

$$\mathbf{S}_w = \boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1$$
$$= \sum_{\boldsymbol{x} \in X_0} \left(\boldsymbol{x} - \boldsymbol{\mu}_0 \right) \left(\boldsymbol{x} - \boldsymbol{\mu}_0 \right)^{\mathrm{T}} + \sum_{\boldsymbol{x} \in X_1} \left(\boldsymbol{x} - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{x} - \boldsymbol{\mu}_1 \right)^{\mathrm{T}}$$

- 类间散度矩阵

$$\mathbf{S}_b = \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right) \left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right)^{\mathrm{T}}$$

- 广义瑞利商（generalized Rayleigh quotient）

$$J = \frac{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}}$$

- 令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} = 1$，最大化广义瑞利商等价形式为

$$\min_{\boldsymbol{w}} \ -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}$$
$$\text{s.t. } \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} = 1$$

- 运用拉格朗日乘子法

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w}$$

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w}$$

- 同向向量

$$\mathbf{S}_b\boldsymbol{w} = \lambda\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)$$

同向向量

- 结果

$$\boldsymbol{w} = \mathbf{S}_w^{-1}\left(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\right)$$

- 求解
  - 奇异值分解 $\mathbf{S}_w = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$
- LDA的贝叶斯决策论解释
  - 两类数据同先验、满足高斯分布且协方差相等时，LDA达到最优分类

- 全局散度矩阵

$$\begin{aligned}\mathbf{S}_t &= \mathbf{S}_b + \mathbf{S}_w \\ &= \sum_{i=1}^{m} (\boldsymbol{x}_i - \boldsymbol{\mu}) (\boldsymbol{x}_i - \boldsymbol{\mu})^T\end{aligned}$$

- 类内散度矩阵

$$\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}$$

其中 $\mathbf{S}_{w_i} = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i) (\boldsymbol{x} - \boldsymbol{\mu}_i)^T$

- 求解得

$$\begin{aligned}\mathbf{S}_b &= \mathbf{S}_t - \mathbf{S}_w \\ &= \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}) (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T\end{aligned}$$

- 类内散度矩阵

$$\mathbf{S}_w = \sum_{i=1}^{N} \mathbf{S}_{w_i}$$

其中 $\mathbf{S}_{w_i} = \sum_{\boldsymbol{x} \in X_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$

- 类间散度矩阵

$$\mathbf{S}_b = \sum_{i=1}^{N} m_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

- p个投影方向 $\boldsymbol{W} = [w_1, w_2, \ldots, w_p]$
  - 最小化类内投影协方差之和

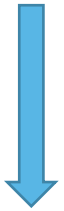
$$\sum_{j \in p} w_j^T \mathbf{S}_w w_j = tr(\boldsymbol{W}^T \mathbf{S}_w \boldsymbol{W})$$

  - 最大化类间投影距离之和

$$\sum_{j \in p} w_j^T \mathbf{S}_b w_j = tr(\boldsymbol{W}^T \mathbf{S}_b \boldsymbol{W})$$

- 优化目标 $\max_{\mathbf{W}} \frac{\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}\right)}{\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}\right)}$

其中 $\mathbf{W} \in \mathbb{R}^{d\times(N-1)}$



$$\mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W}$$

$\mathbf{W}$ 的闭式解则是 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的前d' $(d' \leq N-1)$个最大非零广义特征值所对应的特征向量组成的矩阵

- 多分类LDA将样本投影到d'维空间， d'通常远小于数据原有的属性数d，因此LDA也被视为一种监督降维技术

- 多分类学习方法
  - 二分类学习方法推广到多类
  - 利用二分类学习器解决多分类问题（常用）
    - 对问题进行拆分，为拆出的每个二分类任务训练一个分类器
    - 对于每个分类器的预测结果进行集成以获得最终的多分类结果
- 拆分策略
  - 一对一（One vs. One, OvO）
  - 一对其余（One vs. Rest, OvR）
  - 多对多（Many vs. Many, MvM）

- 一对一

- 训练N(N-1)/2个分类器，存储开销和测试时间大
- 训练只用两个类的样例，训练时间短

一对其余

- 训练N个分类器，存储开销和测试时间小
- 训练用到全部训练样例，训练时间长

多对多（Many vs Many, MvM）

若干类作为正类，若干类作为反类

- 类别不平衡（class imbalance）
  - 不同类别训练样例数相差很大情况（假设正类为小类）

    类别平衡正例预测 $\frac{y}{1-y} > 1$ ⇒ $\frac{y}{1-y} > \frac{m^+}{m^-}$ 正负类比例

- 再缩放
  - 欠采样（undersampling）
    - 去除一些反例使正反例数目接近（EasyEnsemble [Liu et al.,2009]）
  - 过采样（oversampling）
    - 增加一些正例使正反例数目接近（SMOTE [Chawla et al.2002]）
  - 阈值移动（threshold-moving）

作业三

1. 写出最大熵模型的DFP算法

2. 线性判别分析仅在线性可分数据集上能获得理想结果，设计一个改进方法，使其能较好地用于非线性可分数据

3. 二分类问题下的LDA投影方向求解

给定两类数据：类别0：样本均值 $\mu 0=[1,2]^T$　　类别1：样本均值 $\mu 1=[4,6]^T$

假设协方差矩阵相等，类内散度矩阵 Sw=[2　1；1　3]

感 谢 观 看

统 计 机 器 学 习

主讲人：彭振华

数学与计算机学院

2026年