机器学习与数据挖掘

Machine Learning & Data Mining

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

Lecture 8: Linear Model II

8.1 Logistic Regression

二分类任务

- 预测值与输出标记

$$z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b \qquad\qquad y \in \{0, 1\}$$

- 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

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

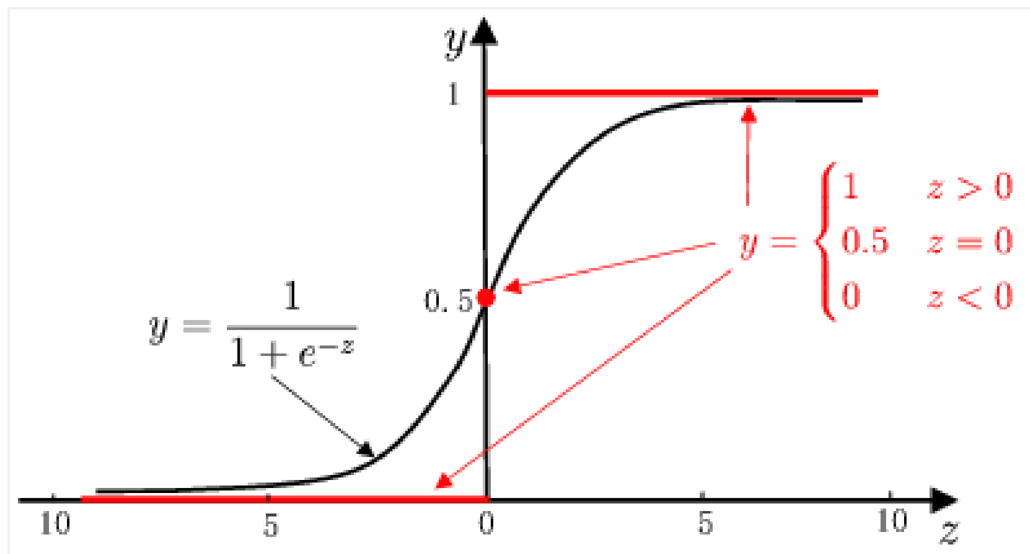
  – 预测值大于零就判为正例，小于零就判为反例，预测值为临界值零则可任意判别

二分类任务

- 单位阶跃函数缺点
  - 不连续
- 替代函数——逻辑函数（logistic function）
  - 单调可微、任意阶可导

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

单位阶跃函数与逻辑函数的比较



逻辑回归

- 运用逻辑函数

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

逻辑回归

- 对数几率（log odds）
  - 样本作为正例的相对可能性的对数

$$\ln\frac{y}{1-y}$$

- 逻辑回归优点
  - 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

逻辑回归 - 极大似然法

- 对数几率

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

逻辑回归是一种广义线性模型

显然有

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

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

逻辑回归 - 极大似然法

- 极大似然法（maximum likelihood）
  - 给定数据集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$
  - 最大化样本属于其真实标记的概率
    - 最大化对数似然函数

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

逻辑回归 - 极大似然法

- 转化为最小化负对数似然函数求解
  - 令 $\boldsymbol{\beta}=(\boldsymbol{w};b)$ $\hat{\boldsymbol{x}}=(\boldsymbol{x};1)$ 则 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ 可简写为 $\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$
  - 再令

$$p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)=p\left(y=1\mid\hat{\boldsymbol{x}};\boldsymbol{\beta}\right)$$

$$p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)=p\left(y=0\mid\hat{\boldsymbol{x}};\boldsymbol{\beta}\right)=1-p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)$$

则似然项可重写为

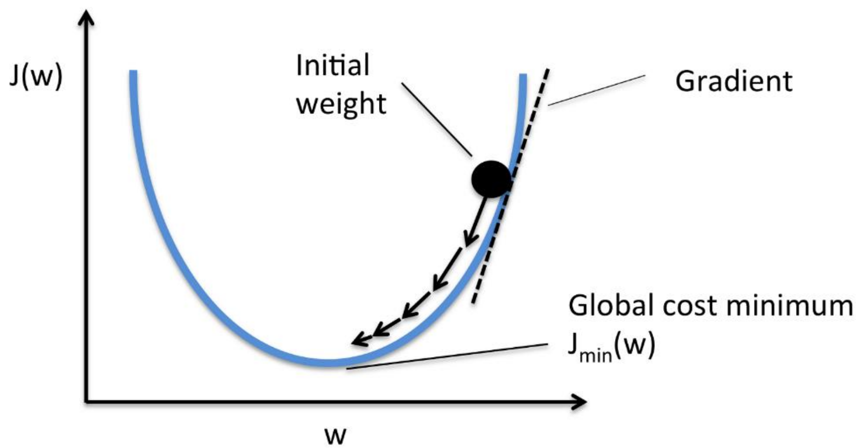
$$p\left(y_i\mid\boldsymbol{x}_i;\boldsymbol{w}_i,b\right)=y_ip_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)+\left(1-y_i\right)p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)$$

逻辑回归 - 极大似然法

- 故等价形式为要最小化

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

高阶可导连续凸函数，梯度下降法/牛顿法



逻辑回归

- 以牛顿法为例，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

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=-\sum_{i=1}^{m} \hat{\boldsymbol{x}}_{i}\left(y_{i}-p_{1}\left(\hat{\boldsymbol{x}}_{i} ; \boldsymbol{\beta}\right)\right)$$

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

逻辑回归

- 牛顿法求解最值问题:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

- 梯度下降法:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n f'(x_n)$$

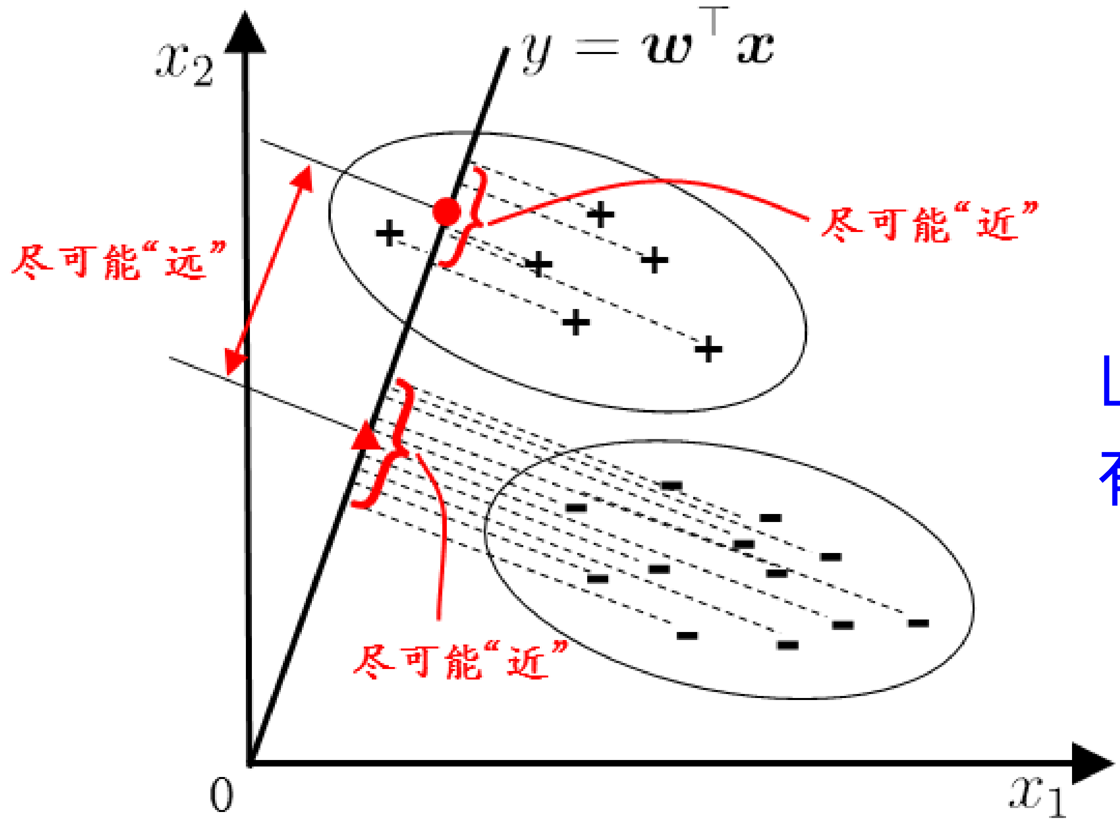
牛顿法：https://zh.wikipedia.org/wiki/%E7%89%9B%E9%A1%BF%E6%B3%95

二分类任务-线性判别分析

- 线性判别分析（Linear Discriminant Analysis）
  - LDA的思想：给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近，不同类的投影点尽可能远离；
  - 在对新样本进行分类时，将其投影到这条直线上，根据投影点的位置确定新样本的类别；

二分类任务-线性判别分析

- 线性判别分析（Linear Discriminant Analysis）



LDA也可被视为一种有监督降维技术

二分类任务– 线性判别分析

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 $\boldsymbol{\Sigma}_i$
- 两类样本的中心在直线上的投影：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_0$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_1$
- 两类样本的协方差：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_0\boldsymbol{w}$ 和 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{w}$

二分类任务– 线性判别分析

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

二分类任务－线性判别分析

- 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

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

二分类任务– 线性判别分析

- **结果**

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

- **求解**
  - 奇异值分解 $\mathbf{S}_w = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$

多分类学习

- 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）

多分类学习– 一对一

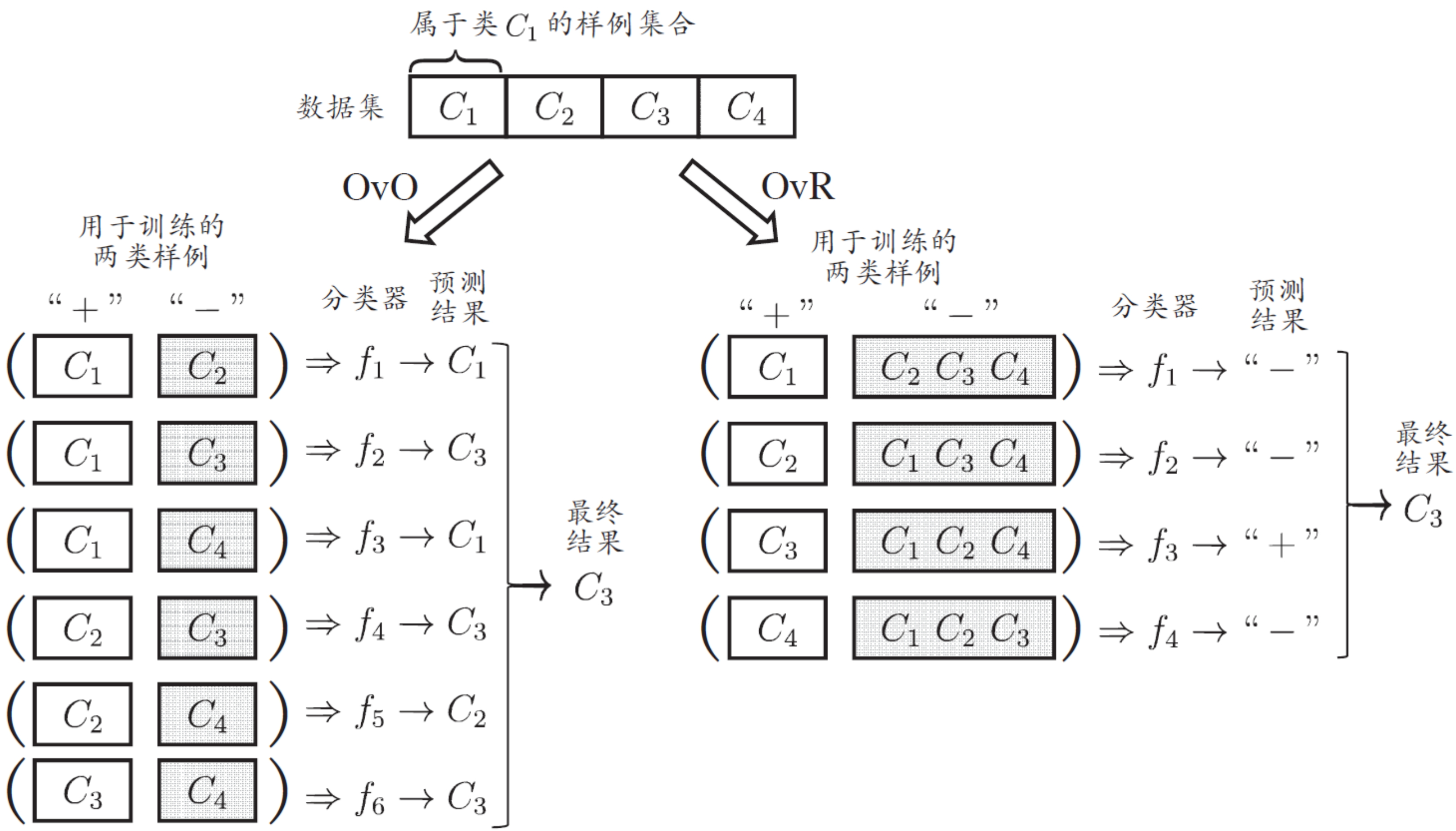
- 拆分阶段
  - N个类别两两配对
    - N(N-1)/2 个二类任务
  - 各个二类任务学习分类器
    - N(N-1)/2 个二类分类器

- 测试阶段
  - 新样本提交给所有分类器预测
    - N(N-1)/2 个分类结果
  - 投票产生最终分类结果
    - 被预测最多的类别为最终类别

多分类学习– 一对其余

- 任务拆分
  - 某一类作为正例，其他反例
    - N 个二类任务
  - 各个二类任务学习分类器
    - N 个二类分类器
- 测试阶段
  - 新样本提交给所有分类器预测
    - N 个分类结果
  - 比较各分类器预测置信度
    - 置信度最大类别作为最终类别

多分类学习- 两种策略比较



类别不平衡问题

- 类别不平衡（class imbalance）
  - 不同类别训练样例数相差很大情况（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）

优化提要

- 各任务下（回归、分类）各个模型优化的目标
  - 最小二乘法：最小化均方误差
  - 逻辑回归：最大化样本分布似然
  - 线性判别分析：投影空间内最小（大）化类内（间）散度
- 参数的优化方法
  - 最小二乘法：线性代数
  - 逻辑回归：凸优化梯度下降、牛顿法
  - 线性判别分析：矩阵论、广义瑞利商

Thank you!

权小军 中山大学数据科学与计算机学院