第三讲 线性模型

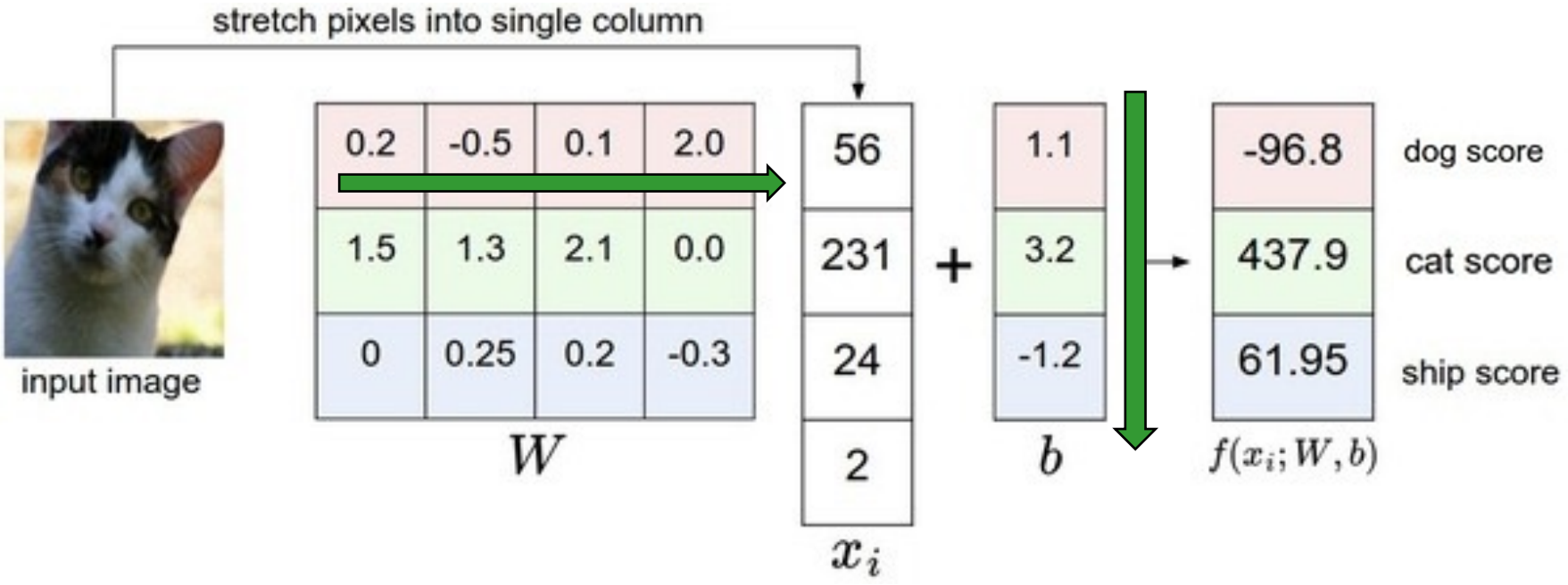
高级机器学习

提纲

- 引言
- 回归任务模型：最小二乘法
- 二分类任务模型：对数几率回归
- 多分类任务模型：一对一、一对其余、多对多

线性模型的基本思想

如何构建区分猫与狗的模型？对特征赋予不同的权值形成预测。



基本思想

- 线性模型一般形式

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

- 向量形式

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

$$\boldsymbol{w} = (w_1; w_2; \ldots; w_d) \quad \boldsymbol{x} = (x_1; x_2; \ldots; x_d)$$

问题：如何确定w和b的值，使得线性模型具有优良性能

线性回归（linear regression）

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

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

- 目的
  - 构建线性模型以其尽可能准确地输出实值标记

- 数据预处理：离散属性处理
  - 有“序”关系
    - 连续化为连续值，如学历（高-中-低）
  - 无“序”关系
    - 有k个属性值，则转换为k维向量，如颜色

最简单的线性回归

- 单一属性的线性回归

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

- 参数估计方法：最小二乘法（least square method）

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

最小二乘法

- 最小化均方误差

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

- 分别对 w 和 b 求导，可得

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

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

最小二乘法

- 得到闭式（closed-form）解

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

$$b = \frac{1}{m}\sum_{i=1}^{m} (y_i - w x_i)$$

其中

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

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

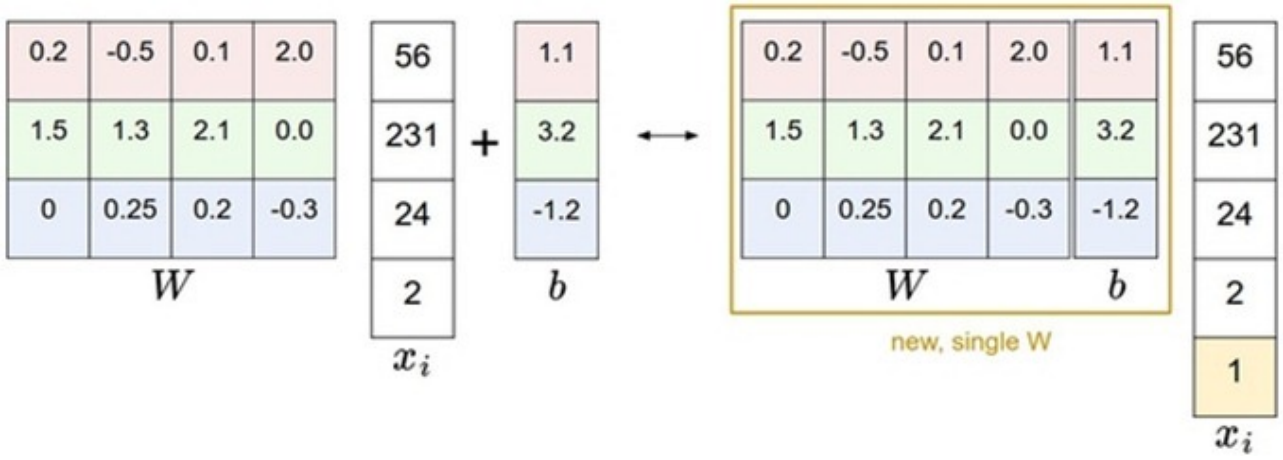
$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b \text{ 使得} f(\boldsymbol{x}_i) \simeq y_i$$

- 三步：齐次表达、最小二乘法、满秩讨论

多元线性回归 – 齐次表达

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

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



多元线性回归 – 最小二乘法

最小二乘法

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

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

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

令上式为零可得 $\hat{\boldsymbol{w}}$ 最优解的闭式解

多元线性回归 - 满秩讨论

- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 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
  - 引入正则化（参见6.4节，11.4节）

广义线性回归模型

- 一般形式

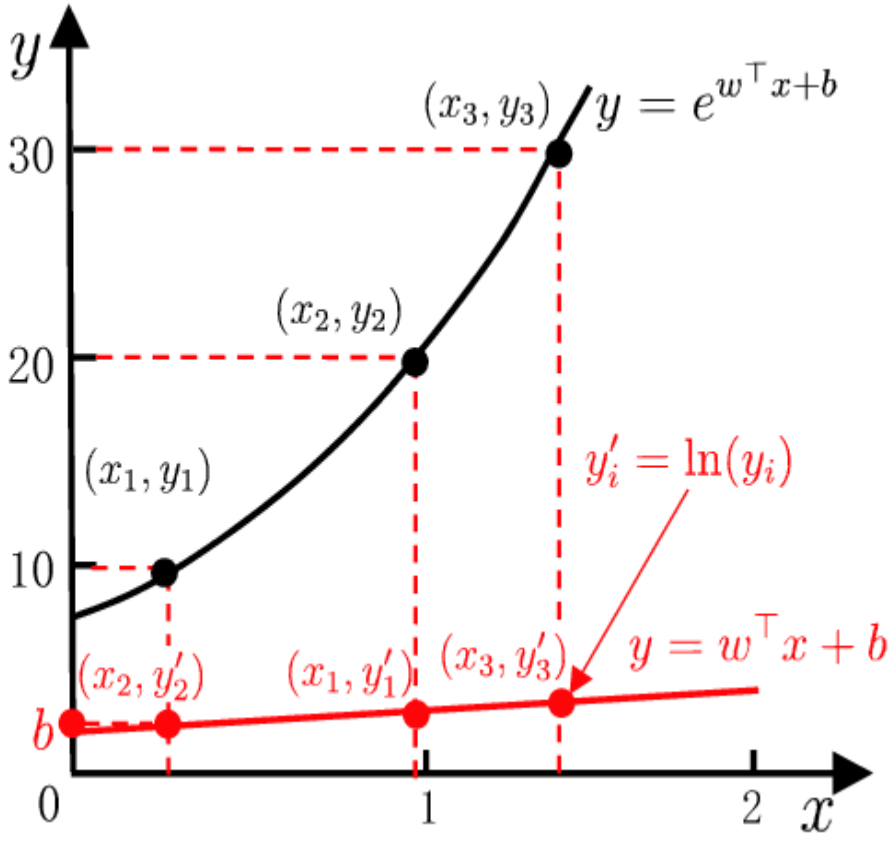
$$y = g^{-1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b\right)$$

$g(\cdot)$ 称为联系函数（link function），单调可微函数

$g(\cdot) = \ln(\cdot)$ 是广义线性模型的特例

广义线性回归模型

- 输出标记的对数为线性模型逼近的目标



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

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

提纲

- 引言
- 回归任务模型：最小二乘法
- 二分类任务模型：对数几率回归
- 多分类任务模型：一对一、一对其余、多对多

线性分类

- 预测值与输出标记 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ $y \in \{0, 1\}$
- 寻找函数将分类标记与线性模型输出联系起来
- 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

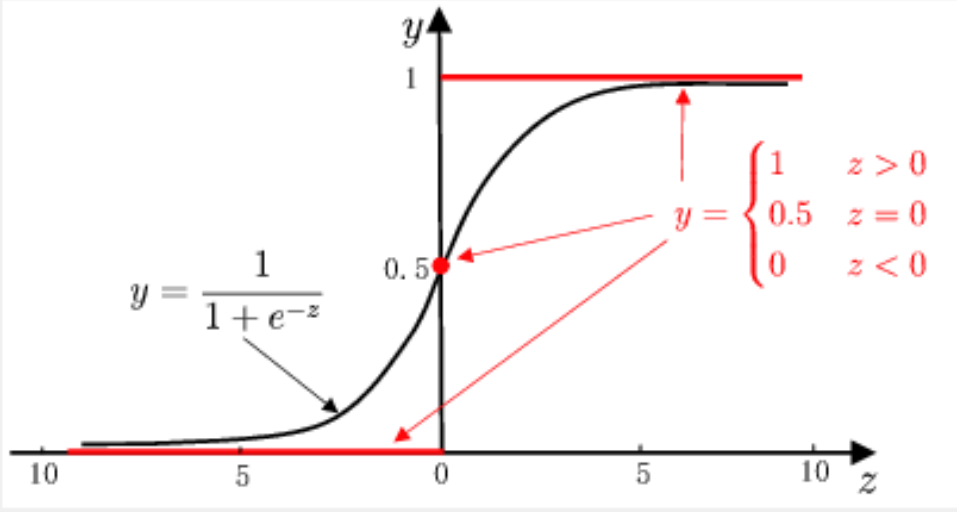
线性分类

- 缺点：单位阶跃函数不连续，不可导

- 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



对数几率回归（logistic regression）

以对率函数为联系函数：

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

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

几率(odds)，反映了 $\boldsymbol{x}$ 作为正例的相对可能性

“对数几率”
(log odds, 亦称 logit)

“对数几率回归”（logistic regression）
简称“对率回归”

- 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

注意：它是
分类学习算法！

对数几率回归 – 极大似然法

- 对数几率

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

显然有

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

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

对数几率回归 – 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
    - 对数似然函数

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

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

对数几率回归 – 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 $p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=p(y=1\mid\hat{\boldsymbol{x}};\boldsymbol{\beta})$

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

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

则似然项可重写为

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

对数几率回归

- 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

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

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

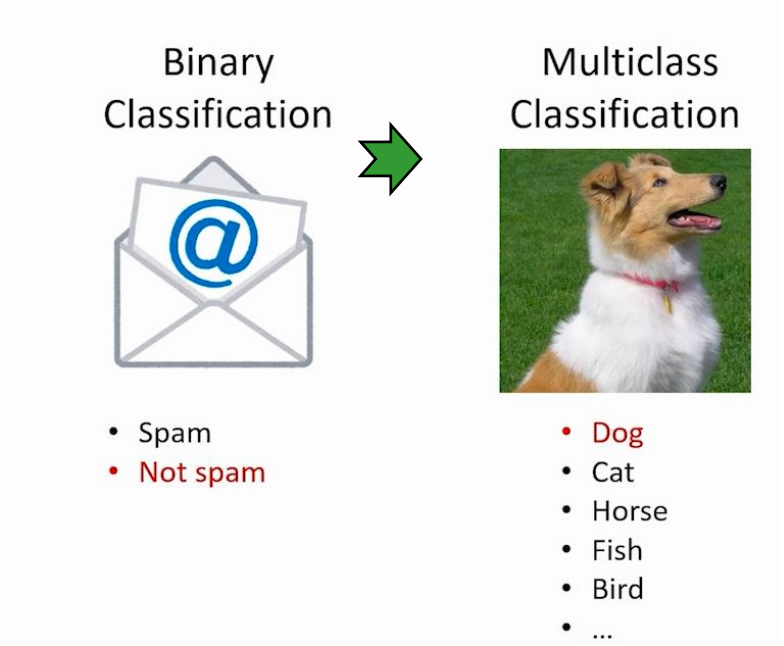
高阶可导连续凸函数，梯度下降法/牛顿法 [Boyd and Vandenberghe, 2004]

提纲

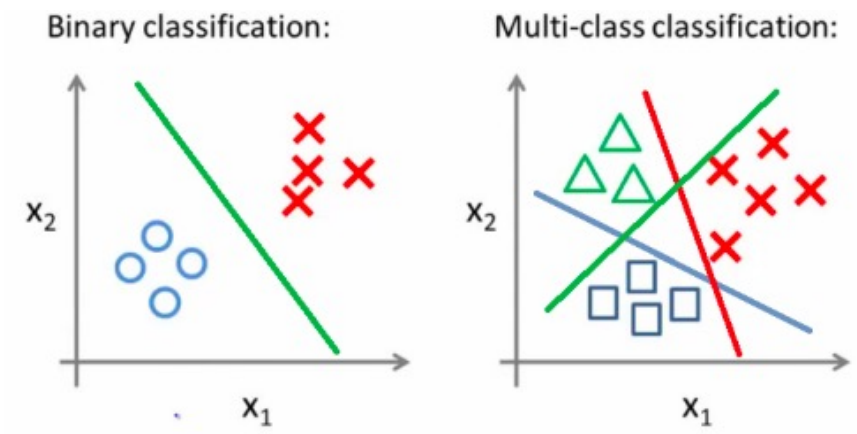
- 引言
- 回归任务模型：最小二乘法
- 二分类任务模型：对数几率回归
- 多分类任务模型：一对一、一对其余、多对多

多分类任务

- 多分类任务



- 基本思想



多分类任务

- 多分类学习方法
  - 技巧：利用二分类学习器解决多分类问题
    - 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）

多分类任务 – 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

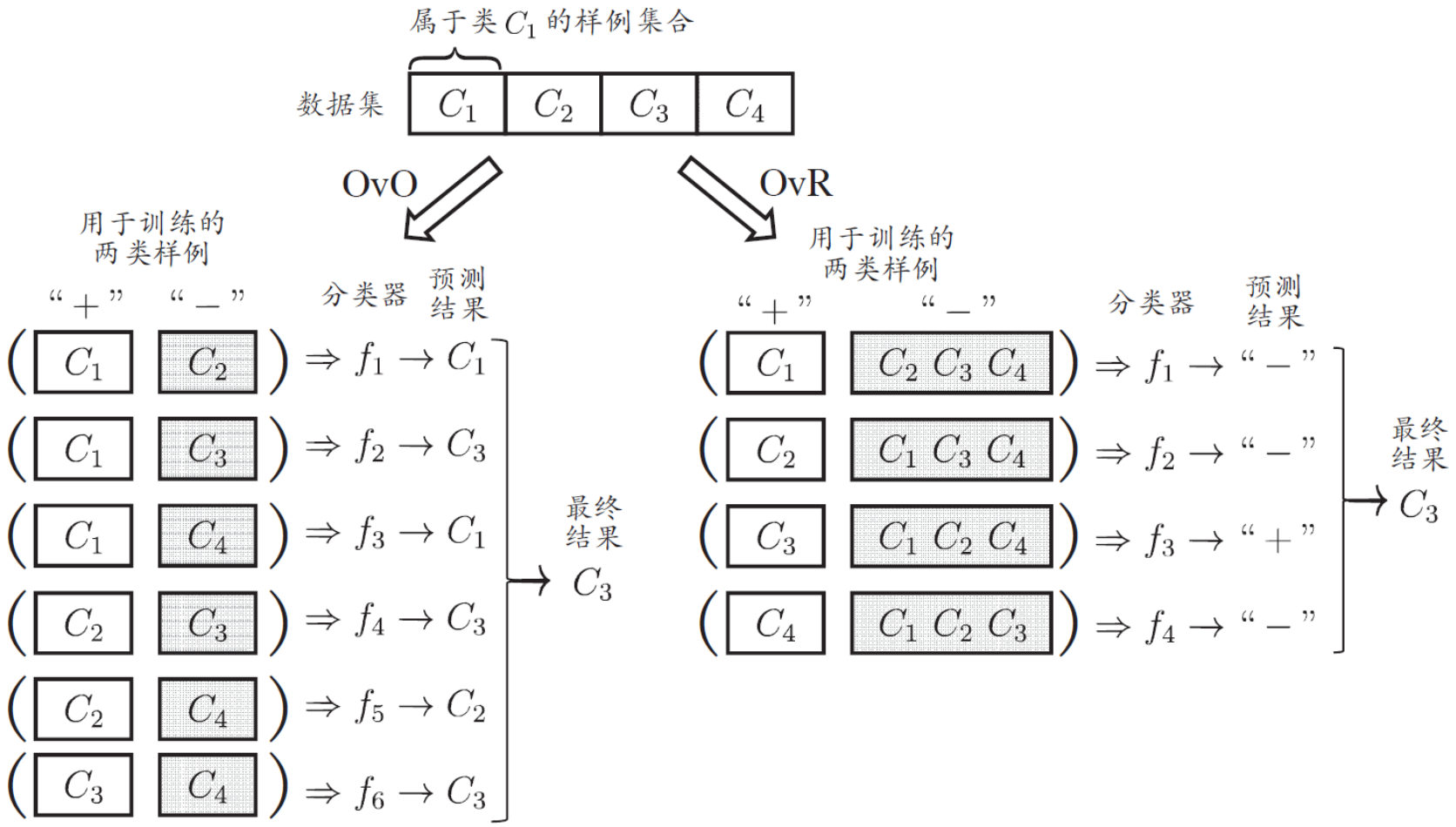
- 测试阶段
  - 新样本提交给所有分类器预测
    - N(N-1)/2 个分类结果
  - 投票产生最终分类结果
    - 被预测最多的类别为最终类别

多分类任务 – 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

多分类任务 - 两种策略比较



多分类任务 - 两种策略比较

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

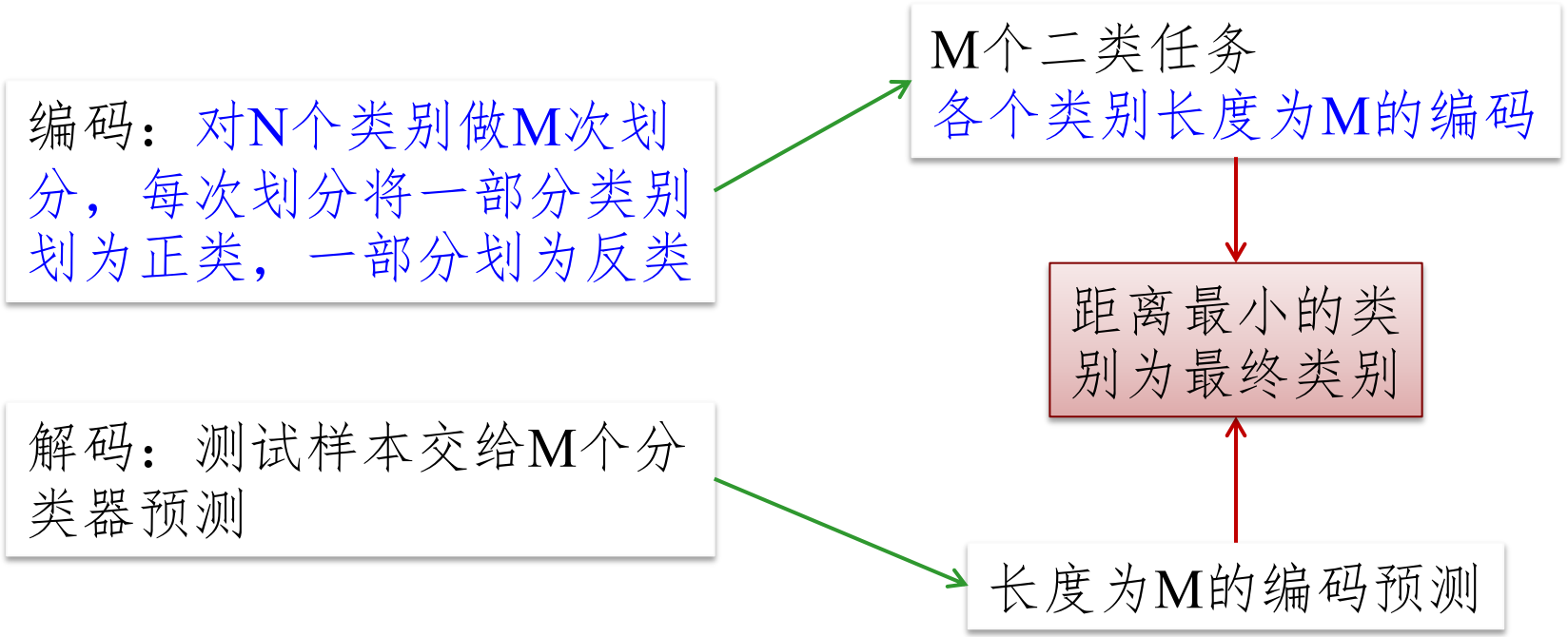
预测性能跟具体数据分布有关；一般来说，不同类别的数据分布有比较明显的差别时，两者性能相当，都可以取得较好的效果。对于类别数据分布有较多交叠的情况，当类别比例比较均衡且每个类的数据量不太少时，“一对一”可能更有优势；当数据量比较少，且类别分布不够均衡时，可能采用“一对多”辅助于类别不平衡数据方法会更合理。

另有一种能够缓解类别的不平衡，又可以缓解数据量不足的方法就是采用“多对多”的策略。

多分类任务 – 多对多

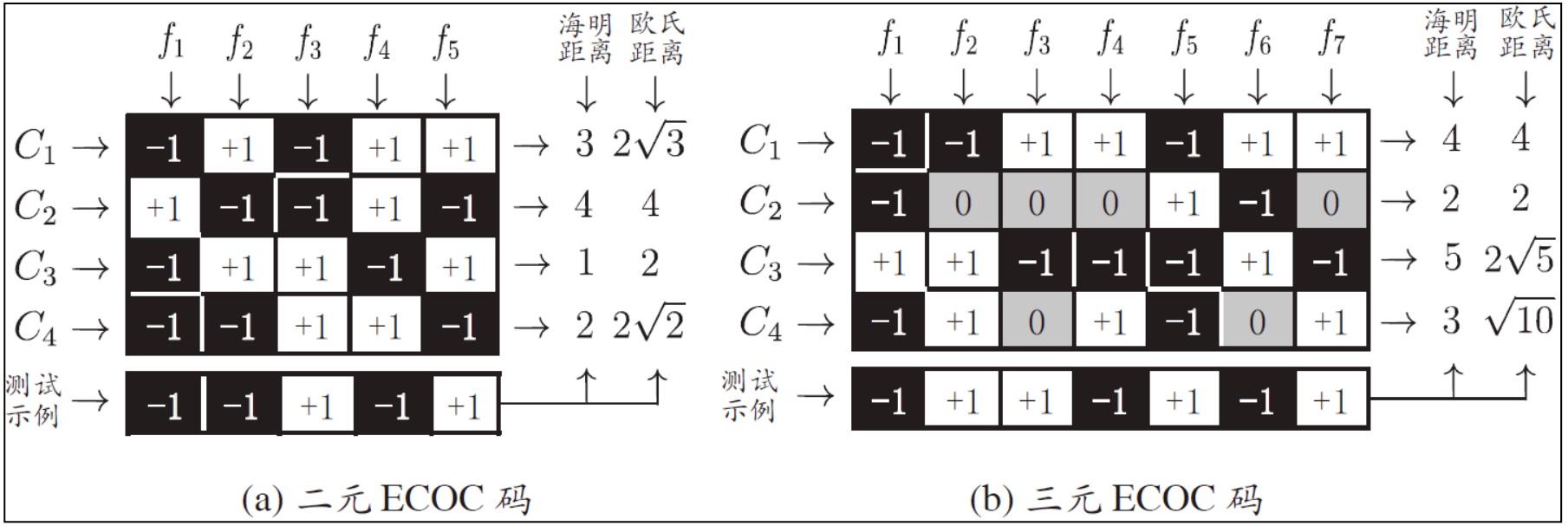
- 多对多（Many vs Many, MvM）
  - 若干类作为正类，若干类作为反类

设计框架：纠错输出码（Error Correcting Output Code, ECOC）



多分类任务 – 多对多

- 纠错输出码 (Error Correcting Output Code, ECOC)



- ECOC编码对分类器错误的容忍和修正能力：编码越长、纠错能力越强
- 同等长度的编码，理论上，两个类别间的编码距离越远，纠错能力越强

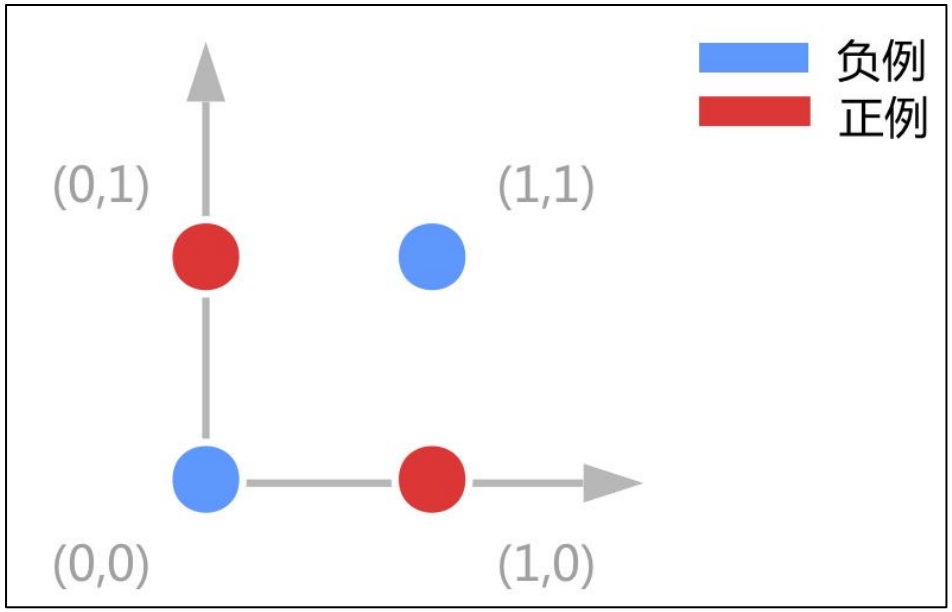
线性模型的优点

- 形式简单、易于建模
- 具有一定的可解释性

- 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，表明根蒂对判别好坏最重要；
  - 而敲声的系数比色泽大，说明敲声比色泽更重要

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

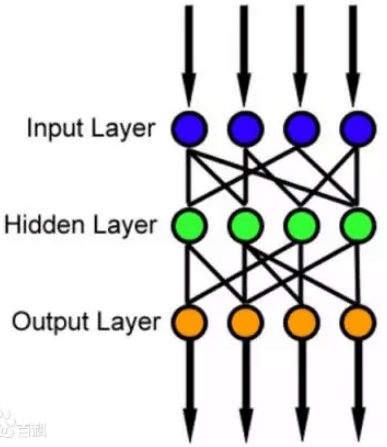
线性模型的缺陷



- 难以处理非线性问题

✓ 非线性模型的基础

  - 如，引入层级结构（如神经网络）或高维映射（如，支持向量机）



小结

- 回归任务
  - 掌握最小二乘法原理和推导
- 二分类任务:
  - 熟悉对数几率回归
- 多分类任务
  - 熟悉一对一、一对其余、多对多的原理
- 线性模型的优缺点