東莞理工學院
DONGGUAN UNIVERSITY OF TECHNOLOGY

人工智能概论

第三章：线性模型

丁烨，计算机科学与技术学院
dingye@dgut.edu.cn

目录

线性模型

基本概念

- 给定一个样本 x，例如西瓜
- x 具备 d 个特征：$x = (x_1, x_2, \cdots, x_d)$
- 其中 x_i 是 x 在第 i 个特征上的取值
- 例如：西瓜 $= (x_{\text{色泽}} = \text{深绿}, x_{\text{根蒂}} = \text{蜷缩}, x_{\text{敲声}} = \text{清脆})$

线性模型

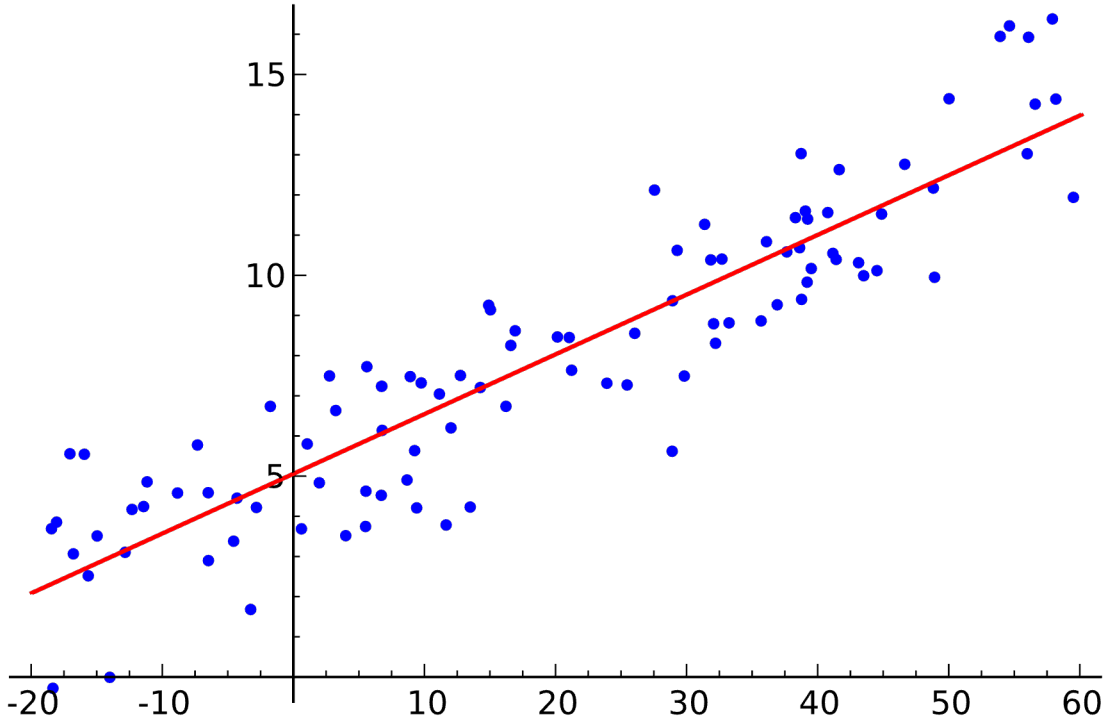
基本概念

- 线性模型试图学得一个函数（模型）
- 这个函数（模型）是特征的线性组合
- 例如：$f_{\text{好瓜}}(x) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.2 \cdot x_{\text{敲声}} + 0.1$
- 每个特征带有一个权重（Weight）：w
- 整个模型带有一个估计常数（Estimated Intercept）：b
- 函数的值 $f_{\text{好瓜}}(x)$ 对于所有的样本来说
- 尽可能的接近真相（Ground Truth）：$y \in \{\text{好瓜}, \text{坏瓜}\}$

线性模型

基本概念

- 如果 x 的维度很小
- 例如只有一维：
- $f_{\text{好瓜}}(x) = 0.5 \cdot x_{\text{敲声}} + 0.5$
- 你可以将 $f_{\text{好瓜}}(x)$ 和 (x, y) 可视化在一个二维坐标系里



线性模型

基本概念

- 线性模型形式简单、易于建模
- 但却蕴涵着机器学习中一些重要的基本思想
- 许多功能更为强大的非线性模型（Non-linear Model）可在线性模型的基础上通过引入层级结构或高维映射而得
- 此外，由于 w 直观表达了各属性在预测中的重要性
- 因此线性模型有很好的可解释性（Comprehensibility）

线性模型

基本概念

- 例如若在西瓜问题中学得:
- $f_{\text{好瓜}}(x) = 0.2 \cdot x_{\text{色泽}} + 0.5 \cdot x_{\text{根蒂}} + 0.2 \cdot x_{\text{敲声}} + 0.1$
- 则意味着可通过综合考虑色泽、根蒂和敲声来判断瓜好不好
- 其中根蒂最重要
- 而色泽和敲声同等重要

目录

线性模型

线性回归

- 线性回归（Linear Regression）
- 找到一个线性方程尽可能的表达原始样本数据的分布
- 我们先考虑一个最简单的场景，样本的特征维度只有一维
- 寻找：$f(x) = wx + b$
- 使得：$f(x) \cong y$
- 找到这个线性方程之后
- 对每一个样本来说，给定其特征 x 就能预测对应的 $f(x)$

线性模型

线性回归

- 如何确定 w 和 b 呢?
- 显然，关键在于如何衡量 $f(x)$ 与 y 之间的差别
- 例如，你可以用均方误差（Mean Square Error，MSE）：

$$(w^*, b^*) = \underset{(w,b)}{\operatorname{argmin}} \sum_{j=1}^{m} (f(x_j) - y_j)^2$$

线性模型

线性回归

- 均方误差有非常好的几何意义
- 它对应了常用的“欧氏距离（Euclidean Distance）”
- 基于均方误差最小化来进行模型求解的方法称为：
- “最小二乘法（Least Square Method）”
- 在线性回归中，最小二乘法就是试图找到一条直线
- 使所有样本到直线上的欧氏距离之和最小

线性模型

线性回归

- 刚才我们考虑的场景比较简单，样本的特征维度只有一维
- 如果 x 具备 d 个特征：$x = (x_1, x_2, \cdots, x_d)$
- 那么我们就需要寻找：$f(x) = \mathrm{w}^T x + b$
- 使得：$f(x) \cong y$
- 这称为“多元线性回归（Multivariate Linear Regression）”

线性模型

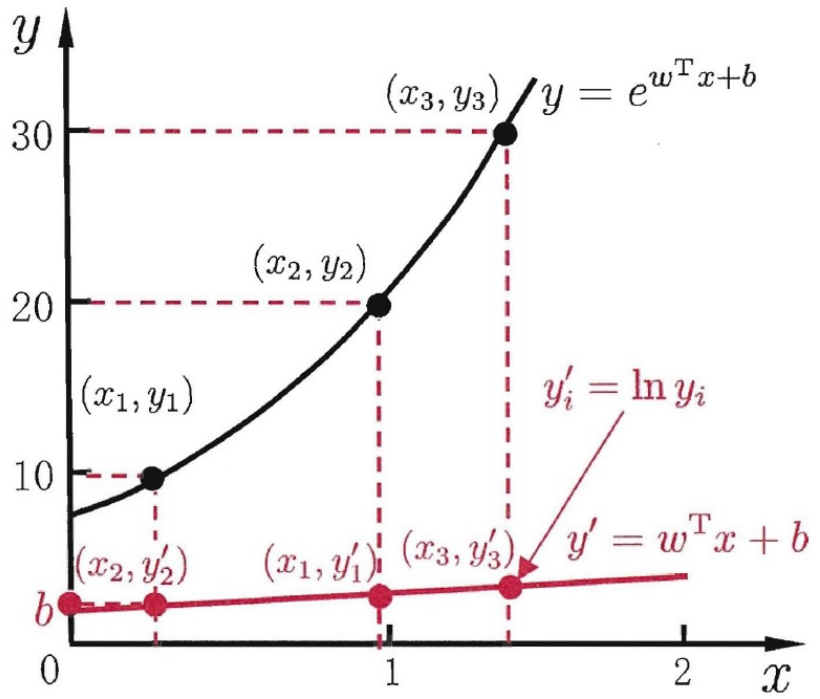
对数线性回归

- 线性模型虽简单，却有丰富的变化
- 标准的线性模型是一个线性函数
- 如果样本的分布倾向于指数尺度的话
- 我们也可以找到一个对应的对数函数来描述样本的分布

线性模型

对数线性回归

- 这个模型叫做：
- 对数线性回归（Log-linear Regression）
- $\ln f(x) = w^T x + b$



目录

线性模型

分类与回归

- 注意，刚才提到的特征值是连续的（Continuous）
- 例如，取值在 -3 到 3 之间的值（这个值也可以是小数）

线性模型

分类与回归

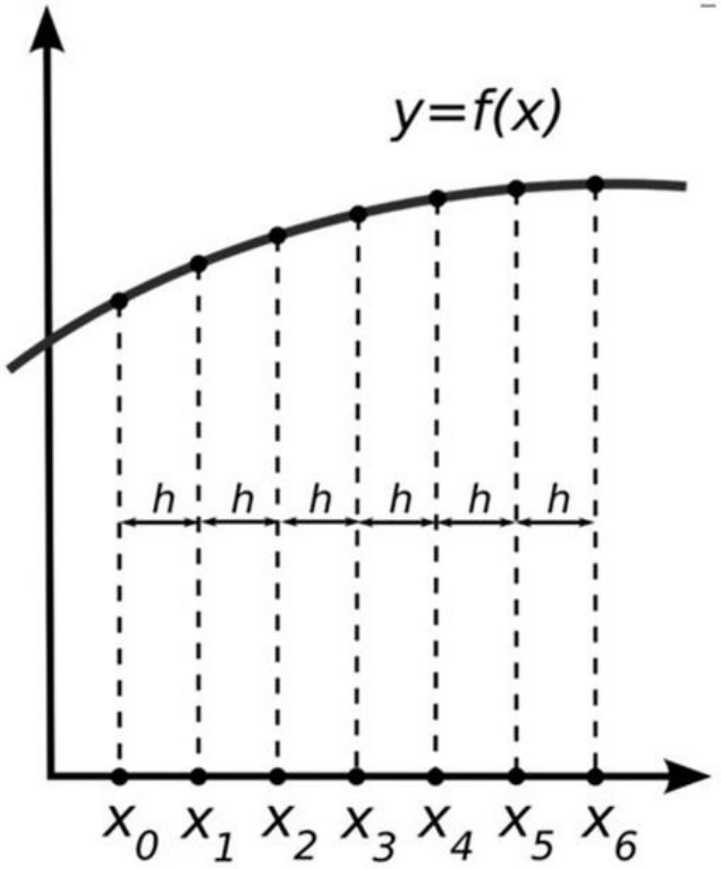
- 但实际中，你也会遇到特征值是离散的（Discrete）
- 例如，西瓜的颜色可能是：
- 白色、浅绿色、深绿色、黑色

线性模型

分类与回归

- 连续值和离散值是可以相互转换的

- 连续值转为离散值：
- 离散化（Discretization）
- 你需要指定一个离散函数，例如 👉



线性模型

分类与回归

- 离散值转为连续值：
- 如果是有序（Order）的离散值，例如：低、中、高
- 可以转换为：0、1、2
- 如果是无序的离散值，例如：白色、绿色、黑色
- 可以转换为多个二阶（Binary）特征，例如：
- 特征一（是否为白色）：0、1
- 特征二（是否为绿色）：0、1
- 特征三（是否为黑色）：0、1

线性模型

分类与回归

- 注意：
- 若将无序特征连续化，则会不恰当地引入顺序（Order）关系
- 对后续处理，如距离计算等，造成误导

线性模型

分类与回归

- 预测的真相 y 也有可能是连续值或者离散值
- 例如：
- 连续值：西瓜好的程度
- 离散值：西瓜是否是好瓜

线性模型

分类与回归

- 一般来说：
- 预测模型输出结果为连续的值：
- 回归（Regression）模型
- 预测模型输出结果为离散的值：
- 分类（Classification）模型

线性模型

对数几率回归

- 上一节讨论了如何使用线性模型进行回归学习
- 但若要做的是分类任务该怎么办？
- 实际上，只需要使用一个单调可微函数
- 将分类任务的真相与线性回归模型的预测值联系起来

线性模型

对数几率回归

- 例如，你可以使用单位阶跃函数（Unit-step Function）

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

- 即：若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

线性模型

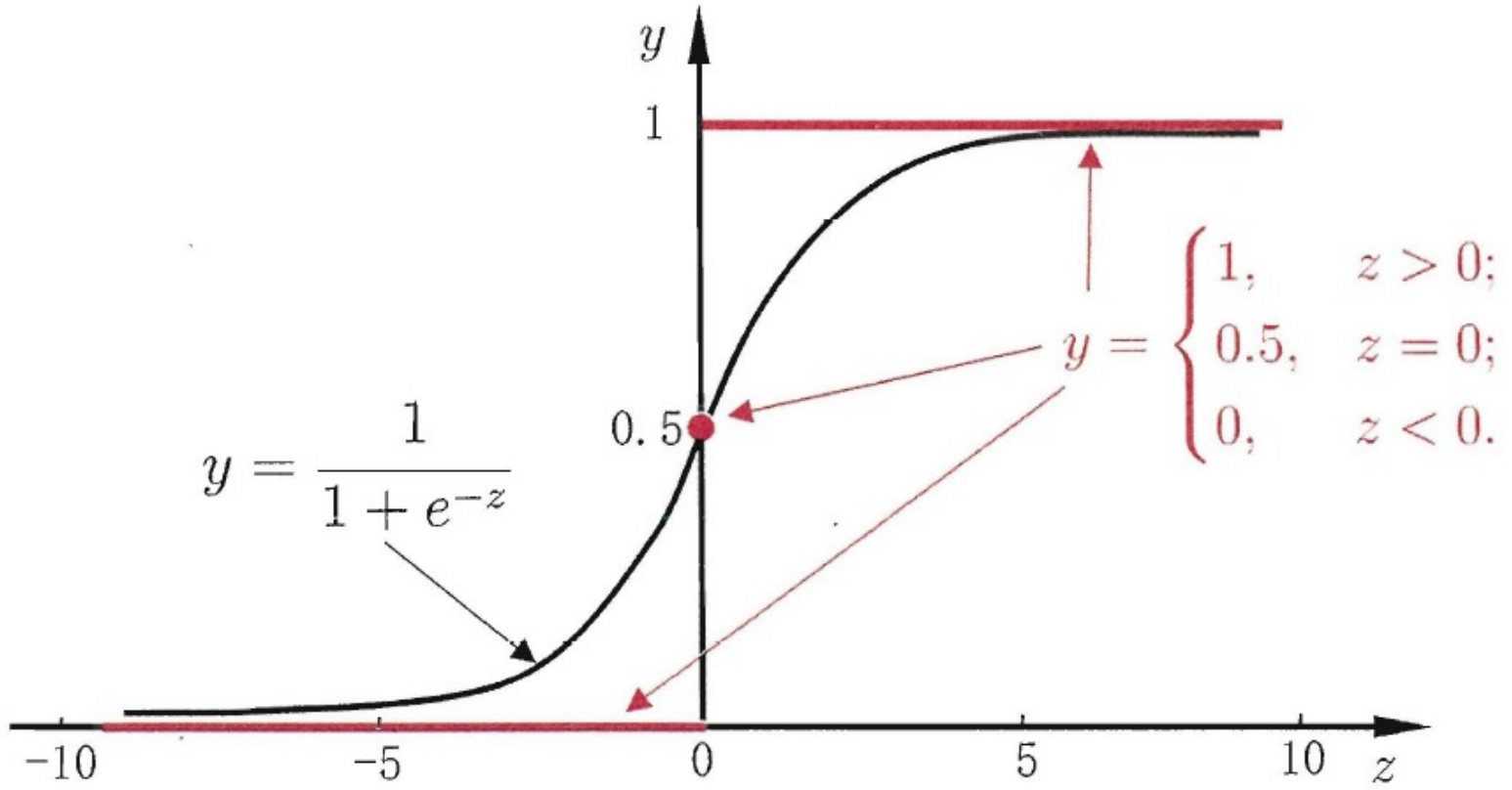
对数几率回归

- 单位阶跃函数不连续，因此并不完全单调可微
- 我们希望找到能在一定程度上近似单位阶跃函数的“替代函数”
- 并希望它单调可微
- 对数几率函数（Logistic Function）正是这样一个常用的替代函数：

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

线性模型

对数几率回归


y
1
0.5
$y = \frac{1}{1+e^{-z}}$
$y = \begin{cases} 1, & z > 0; \\ 0.5, & z = 0; \\ 0, & z < 0. \end{cases}$
−10
−5
0
5
10
z

线性模型

对数几率回归

- 对数几率函数是一种“Sigmoid 函数”，即形似 S 的函数
- 它将 z 值转化为一个接近 0 或 1 的 y 值
- 并且其输出值在 $z = 0$ 附近变化很陡

线性模型

对数几率回归

- 对数几率函数实际上是在：
- 用线性回归模型的预测结果去逼近真相的对数几率
- 因此，其对应的模型称为：
- 对数几率回归（Logistic / Logit Regression）
- 虽然名字是“回归”，但对数几率回归实际是一种分类学习算法

线性模型

对数几率回归

- 对数几率回归有很多优点：

1. 可以直接对分类的可能性进行建模，无需事先假设数据分布，这样就避免了假设分布不准确所带来的问题
2. 它不是仅预测出“类别”，而是可得到近似概率预测，这对许多需利用概率辅助决策的任务很有用
3. 此外，对数几率函数是任意阶可导的凸函数，有很好的数学性质，现有的许多数值优化算法都可直接用于求取最优解

目录

线性模型

线性判别分析

- 线性判别分析（Linear Discriminant Analysis，LDA）
- 一种经典的线性学习方法
- 在二分类问题上因为最早由 Fisher 于 1936 年提出
- 亦称“Fisher 判别分析”

线性模型

线性判别分析

- LDA 的思想非常朴素：
- 设法将样本投影到一条直线上
- 使得：
- 同类样本的投影点尽可能接近
- 异类样本的投影点尽可能远离

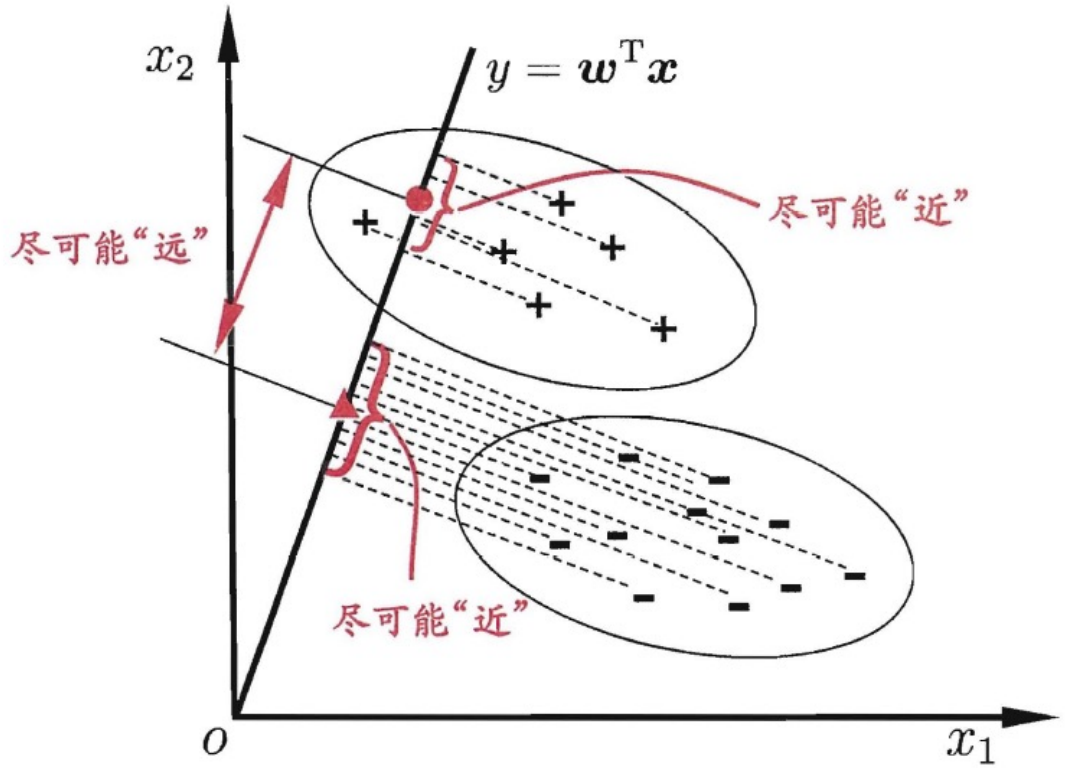


图 3.3 LDA 的二维示意图. “+”、“−”分别代表正例和反例, 椭圆表示数据簇的外轮廓, 虚线表示投影, 红色实心圆和实心三角形分别表示两类样本投影后的中心点.

线性模型

线性判别分析

- LDA 的距离定义主要基于协方差（Covariance）
- 同类样本的投影点尽可能接近：
- 采用：类内散度矩阵（Scatter Matrix）
- 异类样本的投影点尽可能远离：
- 采用：类间散度矩阵

线性模型

线性判别分析

- 在对新样本进行分类时
- 将其投影到同样的直线上
- 再根据投影点的位置来确定新样本的类别

目录

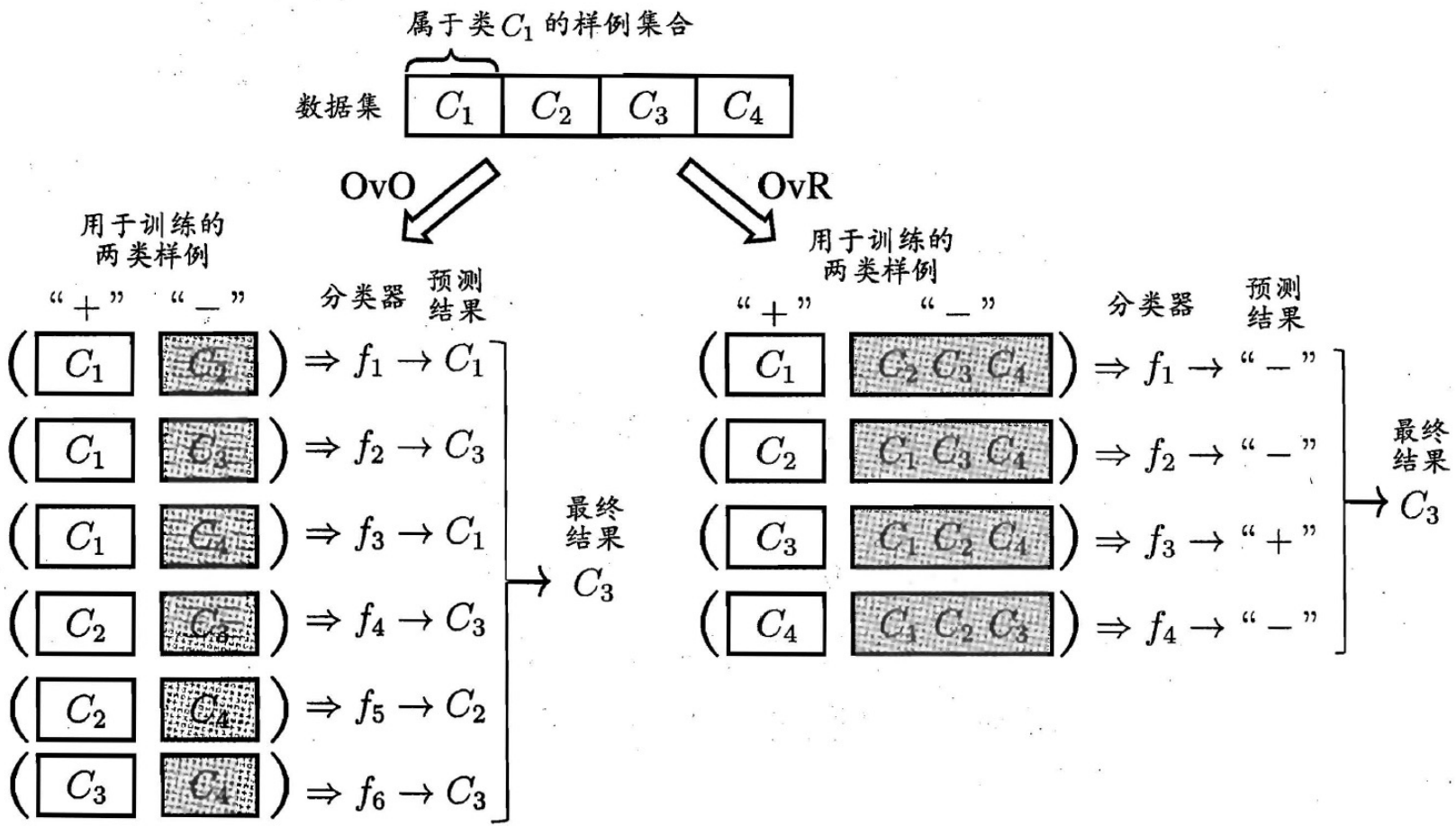
线性模型

多分类学习

- 多分类学习本质上是二分类（Binary Classification）学习的推广
- 因此，大多数情况下我们可以利用二分类学习器来解决多分类问题
- 最经典的拆分策略有三种：
- 一对一（One vs. One，OvO）
- 一对其余（One vs. Rest，OvR）
- 多对多（Many vs. Many，MvM）

线性模型

多分类学习


属于类C_1的样例集合
数据集
C_1 C_2 C_3 C_4
OvO
OvR
用于训练的两类样例
“+” “−”
分类器
预测结果
C_1 C_2 ⇒ $f_1 \rightarrow C_1$
C_1 C_3 ⇒ $f_2 \rightarrow C_3$
C_1 C_4 ⇒ $f_3 \rightarrow C_1$
C_2 C_3 ⇒ $f_4 \rightarrow C_3$
C_2 C_4 ⇒ $f_5 \rightarrow C_2$
C_3 C_4 ⇒ $f_6 \rightarrow C_3$
最终结果
C_3
用于训练的两类样例
“+” “−”
分类器
预测结果
C_1 C_2 C_3 C_4 ⇒ $f_1 \rightarrow$ “−”
C_2 C_1 C_3 C_4 ⇒ $f_2 \rightarrow$ “−”
C_3 C_1 C_2 C_4 ⇒ $f_3 \rightarrow$ “+”
C_4 C_1 C_2 C_3 ⇒ $f_4 \rightarrow$ “−”
最终结果
C_3

目录

- 基本概念
- 线性回归
- 对数几率回归
- 线性判别分析
- 多分类学习
- 类别不平衡问题

线性模型

类别不平衡问题

- 前面介绍的分类学习方法都有一个共同的基本假设
- 即不同类别的训练样本数目相当
- 如果不同类别的训练样本数目稍有差别，通常影响不大
- 但若差别很大，则会对学习过程造成困扰

线性模型

类别不平衡问题

- 例如，有 998 个反例，但正例只有 2 个
- 那么算法只需返回一个永远将新样本预测为反例的模型
- 就能达到 99.8% 的准确度（Precision）
- 然而这样的模型往往没有价值，因为它不能预测出任何正例
- 性能度量（例如 F1）也会很低

线性模型

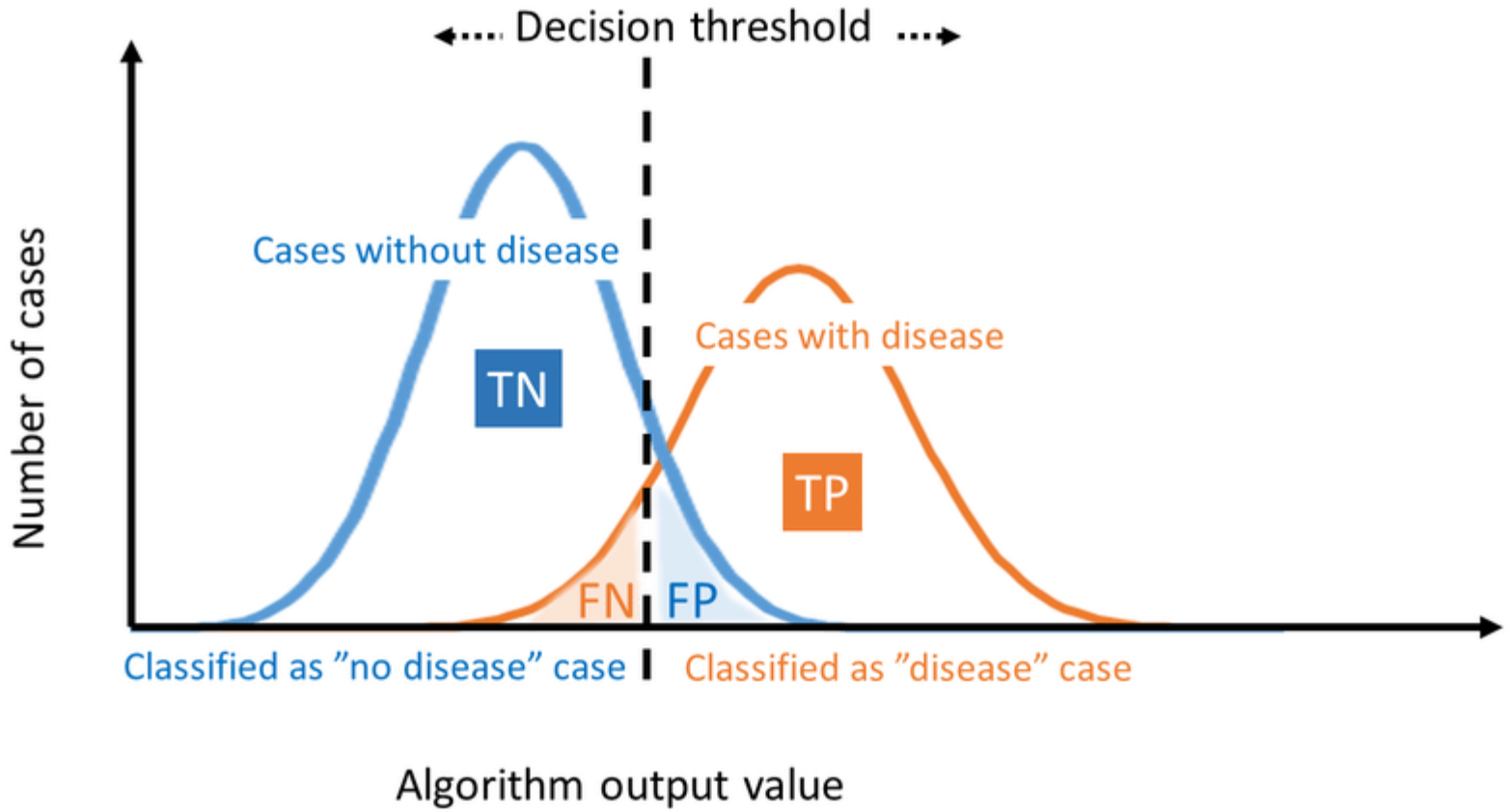
类别不平衡问题

- 处理类别不平衡（Class Imbalance）问题的基本策略：
- 再缩放（Rescaling）
- 欠采样：去除一些样本使得正反例数量接近
- 过采样：增加一些样本使得正反例数量接近（例如：SMOTE）

线性模型

类别不平衡问题

- 阈值移动（Threshold Moving）



总结

- 「西瓜书」周志华《机器学习》，清华大学出版社
- https://item.jd.com/12762673.html

- 「南瓜书」谢文睿、秦州《机器学习公式详解》，人民邮电出版社
- https://github.com/datawhalechina/pumpkin-book/

总结

- Scikit-Learn
- https://scikit-learn.org

- TensorFlow
- https://www.tensorflow.org

TensorFlow

東莞理工學院
DONGGUAN UNIVERSITY OF TECHNOLOGY

Thank You!

丁烨，计算机科学与技术学院
dingye@dgut.edu.cn