线性模型

廖星宇

2017 年 6 月 12 日

1 多分类学习

二分类问题可推广到多分类，主要思路是” 拆解法”，考虑有 N 个类别，经典的拆分策略有三种:

（1）” 一对一” One vs. One (OvO)

$C_N^2 = \frac{N(N-1)}{2}$ 个二分类器，每个分类器只针对两类样本，预测最多的结果作为最终结果。

OvO 每次并不需要训练全部的样本，只需要训练对应选定的两种样本，训练时间短，但是分类器的数目过多，存储开销和测试时间大。

（2）” 一对其余” One vs. Rest (OvR)

构造 N 个二分类器，每种分类器都是选择一个类别为正类别，剩下的为负类别，如果其中一个分类器预测为正，则为最终结果，否则考虑每个分类器的置信值。

每次训练需要全部的样本，训练时间更长，但是存储开销和测试时间更小。

（3）” 多对多” Many vs. Many（MvM

每次将若干类分成正类，若干类分成负类，正、反类的构造需要特殊的设计，一种常见的 MvM 技术：” 纠错输出码”（Error Correcting Output Codes)

编码：对 N 个类别做 M 次划分，每次划分将一部分划为正类，一部分划为反类，形成一个二分类训练集，训练一个分类器，一共有 M 个分类器，

每个类别也有一个 M 维的编码。

解码：用 M 个分类器对样本进行预测，M 个预测结果，与每个类别的编码进行比较，选择最小的距离。

通过这种方式对分类器的错误具有一定的容忍和修正能力，比如 M 个分类器中其中一个分错了，但是对整体的影响特别小，编码越长，纠错能力越强。

2 类别不平衡问题

类别不平衡 (class-imbalance) 是指分类任务中不同类别的训练样例数目差别很大。

$y = w^T x + b$ 对新样本 x 进行分类，得到 y 与阈值比较得到预测结果，几率 $\frac{y}{1-y}$ 反映正例可能性与反例可能性的比值，如果 $\frac{y}{1-y} = 1$，则表示正反样例的可能性相同。

如果正、反例数目不同时，m^+ 表示正例数目，m^- 表示反例数目，则观测几率就是 $\frac{m^+}{m^-}$，假设训练集是真实样本总体的无偏采集，观测几率就代表真实几率，则 $\frac{y}{1-y} > \frac{m^+}{m^-}$ 表示预测为正