第三章 线性模型

王照国

2019 年 3 月 23 日

1 线性回归

$\simeq$ 表示同伦符号，两个拓扑空间如果可以通过一系列连续的形变从一个变到另一个，那么就称这两个拓扑空间同伦。

1.1 对数线性回归

$$\ln y = w^T x + b$$

对数线性回归实质上是试图让$e^{w^T x+b}$逼近y，对应的是在求取输入空间到输出空间的非线性映射

1.2 广义线性模型

更加一般的情况，对于单调可微函数g(.)

$$y = g^{-1}(w^T x + b)$$

这样的函数称为“广义线性模型”

2 对数几率回归(logistic function)

$$y = \frac{1}{1 + e^{-(w^T x+b)}}$$

3 线性判别分析LDA

思想：给定训练样本集，设法将样本投射到一条直线上，使得同类样本点的投影点尽可能近，异类样例的样本点尽可能远；在对新的样本点做预测的时候，先将样本点投射到这条直线上面，然后对其进行预测

4 多分类学习

考虑N个类别，多分类学习的基本思路是“拆解法”，将多分类问题拆为若干个二分类问题求解，策略分为三种：“一对一”，“一对多”，“多对多”

5 类别不平衡问题

5.1 解决类别不平衡问题的三个方法:

1.对数据较多的类别进行欠采样

2.对数据较少的类别进行过采样，重复使用一部分数据

3.分类的阈值调整：$0.5- > \frac{small}{small+large}$

5.2 各自存在的缺点：

1,2 都改变了数据的原始分布

1 浪费了数据

2 造成了过拟合

3 “训练集是真实样本总体的无偏采样”的假设往往在实际中不成立