The Elements of Statistical Learning notes 2.1-2.2 part

Augest 大魔王

2022 年 3 月 8 日

本文是关于《The Elements of Statistical Learning learning》的学习笔记, 主要涉及书的第一章到第二章的部分内容。该部分主要为入门介绍与基本概念的阐述。

1 介绍

输入变量 (inputs)，可以是测量得到或者预设的．这些变量对一个或多个**输出变量 (outputs)** 有影响．便是利用输入变量去预测输出的值．这样的过程称之为 ** 监督学习 (supervised learning)**.

在统计学中，**输入变量 (inputs)** 通常称**预测变量 (predictors)**, 也可以被叫做**自变量 (independent variables)**. 在模式识别中，被叫做**特征 (features)** 的说法。

输出变量 (outputs) 被称作**响应变量 (responses)**，也可以被叫做**因变量 (dependent variables)**。

2 变量类型与术语

输出变量的类型，根据度量可以分为**定量的 (quantitative)** 与**定性的 (qualitative)**。定性变量也被称为**类别型 (categories)** 或者**离散 (discrete) 型变量**，也被称作**因子 (factors)**。

对于两种类型的输出变量，考虑使用输入变量去预测输出变量是有意义的。当我们预测定量的输出时被称为**回归 (regression)**，当我们预测定

性的输出时被称为**分类 (classification)**. 我们将会看到这两个任务有很多的共同点，特别地，两者都可以看成是 ** 函数逼近 **.

输入变量也有各种类型，除了定性和定量以外，还有第三类**有序分类 (ordered categorical)**，如小 (small)、中 (medium) 和大 (large)，在这些值之间存在顺序，但是没有合适的度量概念（中与小之间的差异不必和大与中间的差异相等）.

定性的变量常用数字编码来表示. 最简单的情形是只有两个分类，比如说“成功”与“失败”, “生存”与“死亡”. 这些经常用一位二进制数来表示，比如 0 或 1, 或者用-1 和 1 来表示, 这些编码有时被称作**指标 (targets)**. 当存在超过两个的类别，存在其他可行的选择. 最有用并且最普遍使用的编码是**虚拟变量 (dummy variables)**. 这里有 K 个水平的定性变量被一个 K 位的二进制变量表示, 每次只有一个在开启状态, 尽管更简洁的编码模式也是可能的, 但虚拟变量在因子的层次中是对称的.

我们将经常把输入变量用符号 X 来表示. 如果 X 是一个向量, 则它的组成部分可以用下标 X_j 来取出.

定量的输出变量用 Y 来表示, 对于定性的输出变量采用 G 来表示.

我们] 使用大写字母 X, Y, G 来表示变量, 对变量的观测值我们用小写字母来表示; 因此 X 的第 i 个观测值记作 x_i

举个例子, N 个 p 维输入向量 $x_i, i = 1, \cdots, N$ 可以表示成 $N \times p$ 的矩阵 $\mathbf{X}$.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} \tag{1}$$

通常我们把维度用列表示，观测值用行表示。因此，根据上述矩阵，每一行都是一个输入向量，每一列则是一个维度。

我们简单定义统计学习，如下: 给定输入向量 X, 对输出 Y 做出一个不错估计, 记为 $\hat{Y}$. 并且如果 Y 取值为实数 $\mathbf{R}$, 则 $\hat{Y}$ 取值也是实数 $\mathbf{R}$; 同样地, 对于类别型输出, $\hat{G}$ 取值为对应 G 取值的集合 $\mathcal{G}$.