统计学习引论笔记

Kevin

目录

1 基本概念

假定我们观测到一个定量应变量 Y，p 个不同的自变量 $X = (X_1, \cdots, X_p)$，它们之间存在关系：

$$Y = f(X) + \epsilon.$$

其中，f 是某个固定但未知的函数，ϵ 是随机误差项（error term），ϵ 与 X 独立且均值为 0，f 表示 X 为 Y 提供的系统性信息。简单来说，统计学习就是一系列估计 f 的方法。

为什么要估计 f？主要有两种目的：预测和推断。从预测来说，我们获得 f 的估计 $\hat{f}$，然后从 X 可以估计 Y：$\hat{Y} = \hat{f}(X)$。$\hat{f}$ 一般被当成一个黑箱（black box），只要能通过它得到准确的估计，我们不太关心它的具体形式。$\hat{Y}$ 的准确性取决于两个误差：可消除误差和不可消除误差。统计学习方法的目标就是最小化可消除误差。从推断来说，我们需要考察自变量如何影响应变量，此时 $\hat{f}$ 不能看作黑箱了，我们需要知道它的准确形式。

怎么估计 $\hat{f}$？我们需要一系列训练数据：

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}, \quad \text{where} \quad x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T.$$

基于这些训练数据，我们需要寻找某个函数 $\hat{f}$，满足：

$$Y \approx \hat{f}(X), \forall (X, Y).$$

大致上，统计学习方法分为参数方法和非参数方法。

1.1 评估模型准确性

对于回归问题，最常用的测度是均方误（mean squared error，MSE）：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

2 线性回归

面对一个数据集，比如 Advertising，包括四个变量，sales，TV，radio，newspaper。我们要问几个问题：

1. 广告预算与销售额之间有关系吗?
2. 如果有上述关系，这种关系有多强?
3. 哪种广告媒介影响销售额?
4. 广告媒介对销售额的影响能准确估计吗?
5. 能准确预测未来的销售额吗?
6. 广告支出与销售额之间的关系是线性的吗?
7. 广告媒介之间存在协同效应吗?

2.1 简单线性回归

假设 X 和 Y 之间近似存在线性关系：

$$Y \approx \beta_0 + \beta_1 X.$$

一旦获得了参数的估计值 $\hat{\beta}_0, \hat{\beta}_1$，我们就可以根据某个特定的 X，预测 Y：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

其中，$\hat{y}$ 表示在 $X = x$ 时 Y 的一个预测。

我们要找到最能够拟合数据集的参数估计，使得组成的直线尽可能靠近数据点。有多种方式来衡量 “近”，最常用的方法之一是最小二乘法（OLS）。简单线性回归的 OLS 估计为：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

其中，$\bar{y}, \bar{x}$ 为样本均值。

2.2 多元线性回归

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

作多元回归，一般要回答四个问题：

1. 响应变量和自变量之间相关吗?
2. 是否所有的自变量都能解释 Y，抑或只有一部分有用?
3. 模型对数据的拟合程度如何?
4. 给定自变量值，如何预测响应变量的值? 预测的准确性如何?

2.2.1 问题 1：响应变量与自变量之间有关系吗？

对于这个问题，可以用 F 检验。原假设为所有的参数均为 0：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0; \quad H_a : \exists j, \beta_j \neq 0.$$

构造 F 统计量：

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)},$$

可以证明，当原假设为真时，上述统计量中分子和分母的期望值均为 σ^2（误差方差），因此，如果自变量和响应变量不存在关系，F 统计量应该接近 1。到底 F 多大时，可以拒绝原假设? 这要看 n 和 p 的大小。当 n 较大时，即使 F 统计量只比 1 稍大也可以拒绝原假设。反之，当 n 比较小时，需要更大的 F 值。当 H_0 为真且误差项服从正态分布时，F 统计量服从 F 分布。由此，可以从相应的 p 值来判断是否可以拒绝原假设。

（TODO：F 检验与 t 检验的关系，为何需要 F 检验?）

2.2.2 问题 2：哪些变量是重要的？

多元回归分析的第一步，是计算 F 统计量并检查相应的 p 值。如果这一步得出至少有一个自变量与响应变量有关，那么很自然地，我们要找出哪些是重要的（应该进入模型），这项任务称为 “变量选择”。我们可以查看每一个变量的 p 值（t 检验），但是如果自变量数量很多（p 很大），这种办法会导致错误。

通过选取不同的变量，可以生成 2^p 个模型。当 p 较大时，尝试所有的模型显然是不切实际的。我们需要一种自动且有效的方法来选取一部分模