

作业1：线性回归

本次作业deadline为2019.3.12，任何编程问题请提供源代码，作业有任何问题请及时联系助教。

- 1. 名词解释
- 2. 证明
- 3. 过拟合问题
- 4. 前列腺特异抗原水平预测

1. 名词解释

请对下列名词给出你的理解。

- 人工智能（Artificial Intelligence）
- 模式识别（Pattern recognition）
- 机器学习（Machine Learning）
- 深度学习（Deep Learning）
- 统计学习（Statistical Learning）

2. 证明

试证明线性回归中的 R^2 与Pearson相关系数 r 的关系：

$$R^2 = r^2$$

其中

$$R^2 = \frac{\sum_1^n (\bar{y} - \hat{y}_i)^2}{\sum_1^n (\bar{y} - y_i)^2}, r = \frac{cov(x, y)}{\rho_x \rho_y}$$

x, y 均为一维向量。

3. 过拟合问题

利用模型 $y = \theta_1 \times x + \theta_0 + \epsilon$ 生成一组仿真数据 (x, y) ，其中 x 服从 $N(0, 1)$ 的正态分布。 $\theta_1 = 3$ ， $\theta_0 = 6$ 。残差项 ϵ 服从正态分布 $N(0, \sigma^2)$ ，分别考虑 $\sigma = 0.5$ 和 2 的情况，回答以下问题。

- (1) 随机生成10个训练样本数据，分别用线性模型，一元二次和一元三次模型对改组数据进行回归，得到回归模型的参数，绘制散点图和回归曲线，计算RSS并比较大小。
- (2) 再随机生成100个测试样本，用(1)中的模型预测y值，并比较三种模型的预测效果。
- (3) 将(1)中的“随机生成10个训练样本数据”改为“随机生成100个训练样本数据”，重复步骤(1)(2)。
- (4) 请多次重复(1)-(3)，对 σ 的取值、模型复杂程度、训练样本量和模型效果之间的关系进行总结。

4. 前列腺特异抗原水平预测

附件提供了一些前列腺癌患者临床指标的数据。请使用前四个临床数据（即lcavol, lweight, lbph, svi）对前列腺特异抗原水平（lpsa）进行预测。在给出的prostate_train.txt文件和prostate_test.txt文件中，前4列每一列代表一个临床数据（即特征），最后一列是测量的前列腺特异抗原水平（即预测目标的真实值）；每一行代表一个样本。

- (1) 在不考虑交叉项的情况下，利用Linear Regression对prostate_train.txt的数据进行回归，给出回归结果，并对prostate_test.txt文件中的患者进行预测，给出结果评价。
- (2) 如果考虑交叉项，是否会有更好的预测结果？请给出你的理由。

数据名词解释：

lcavol: log cancer volume

lweight: log prostate weight

lbph: log of the amount of benign prostatic hyperplasia

svi: seminal vesicle invasion

lpsa: level of prostate-specific antigen