Table of Contents

第四周作业

Part1

对于数据集auto，进行以下任务：

- (c) 建立一个二元变量，每加仑里程量($mpg01$)，1表示每加仑汽油该型号撒所拍里程数(mpg)在$\frac{3}{4}$分位数以上，0表示每加仑汽油该型号车所跑里程数(mpg)在$\frac{3}{4}$分位数以下。将数据集按照留出法随机分为训练集(比例0.8)和测试集（比例0.2）,对于函数$P(mpg01=1)=\frac{exp\{3.85-0.01*weight\}}{1+exp\{3.85-0.01*weight\}}$编写程序估计0-1损失下的泛化误差；

- (d) 将数据集按照分层等比例分按照训练集(比例0.8)和测试集(比例0.2),对于函数$P(mpg01=1)=\frac{exp\{3.85-0.01*weight\}}{1+exp\{3.85-0.01*weight\}}$编写程序估计0-1损失下的泛化误差

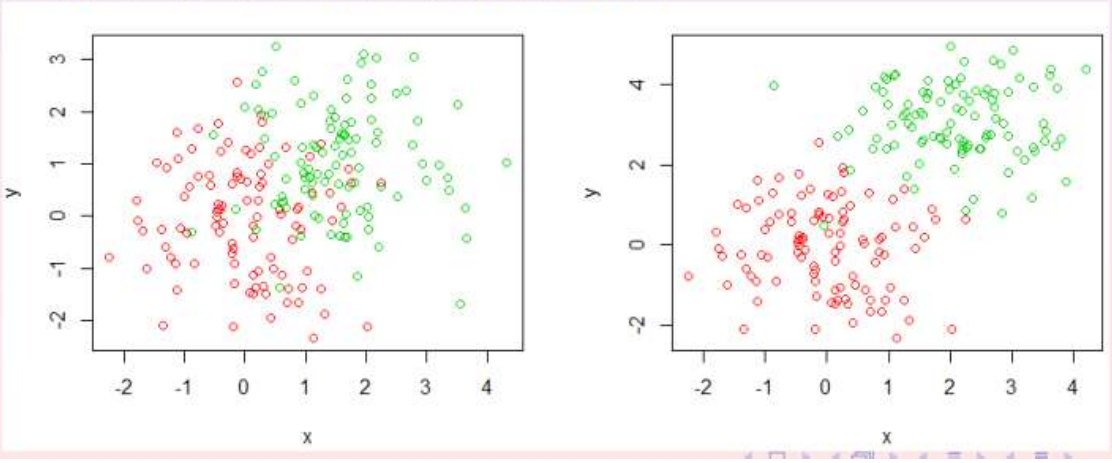
- (e) 根据以上实验，结合教材，分析分层抽样和不同的抽样方式对泛化误差的影响

- (f) 数据集包含1000个样本，其中500个正例，500个反例，将其划分为包含70%样本的训练集和30%样本的测试集用于留出法评估，试估算共有多少种划分方式

part2

PPT内思考题

- 对于图中的两个图，你认为如果用线性模型来划分两组数据，他们的P-R曲线会有怎样的不同？



part3

如何绘制ROC与AUC

- 假设已经得出一系列样本被划分为正类的概率，然后按照大小排序，下表是一个示例，表中共有20个测试样本，“Class”一栏表示每个测试样本真正的标签（*p*表示正样本，*n*表示负样本），“Score”表示每个测试样本属于正样本的概率。

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

- (a) 参照例题数据，作ROC曲线和P-R曲线（要求自写代码绘制，函数为佳）

- (b) 参照例题数据绘制均衡代价曲线（注：$cost_{01} = 1, cost_{10} = 1$）。如果该数据是一个银行追查网络钓鱼攻击的一个学习器的训练数据，漏网的代价(FN)是误报代价(FP)的五倍（$cost_{01} = 5, cost_{10} = 1$），那么请绘制代价曲线。如果上述代价是3倍，代价曲线会怎么变化，请比较三种不同代价下的代价曲线有什么变化?

- (c) 如下数据是16次实验中的误分类结果，请比较测试误差率是否小于30%($\epsilon \leq 30$)，实验中每次测试集$n = 10$的误分数量(mc)如下，凭借这些数据使用t检验等于该算法的泛化误差给出判断。请给出检验问题的表述，计算统计量及其$p - value$,给出结论。

3	4	1	2	0	1	2	3
1	2	4	0	0	3	4	5

ps: 这周作业由于发布时间较晚，延长至下周三（3月23日）22：00，另外同学们写作业的时候记得写一会快捷键保存一下，免得没保存，这是一个好习惯^-^