HOMEWORK 3

1.

(1) Write Newton-Raphson algorithm to estimate logistic regression.

Reminder: you need to derive the equation

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^\top} = -\sum_i x_i x_i^\top p(x_i;\beta)\big\{1 - p(x_i;\beta)\big\}. \tag{0.1}$$

Generate $X = (1, X_1, X_2)$, where $X_j \sim N(0, I_N)$.

Set true parameter $\beta = (0.5, 1.2, -1)^\top$.

Set $N = 200, 500, 800, 1000$.

Estimate β using NR algorithm for $R = 200$ rounds of simulation. For each round of simulation, terminate the iteration when $\max_j |\widehat{\beta}_j^{old} - \widehat{\beta}_j^{new}| < 10^{-5}$. Denote $\widehat{\beta}_j^{(r)}$ as the estimation of β_j in the rth round of simulation. Then please: for each j, draw $(\widehat{\beta}_j^{(r)} - \beta_j)$ in boxplot for $N = 200, 500, 800, 1000$.

(2) 假设有 m^+ 个正例和 m^- 个负例，令 D^+ 与 D^- 分别表示正例、负例集合。定义排序“损失”如下：

$$\ell_{rank} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \Big(I(f(x^+) < f(x^-)) + \frac{1}{2} I(f(x^+) = f(x^-)) \Big) \tag{0.2}$$

理解：若正例的预测值小于负例，则记一个“罚分”，若相等，则记 0.5 个罚分。定义 AUC：

$$AUC = 1 - \ell_{rank}. \tag{0.3}$$

考虑一种简单的情况，即当数据中不存在 $f(x^+) = f(x^-)$ 时，定义排序“损失”如下：

$$\ell_{rank} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \Big(I(f(x^+) < f(x^-)) \Big) \tag{0.4}$$

试证明以上定义的 AUC 即有限样本下 ROC 曲线下方的面积。

2. 客户流失预警数据分析及算法实现，编程语言可以使用 R/python。

针对附录中描述的客户流失预警数据，完成以下任务：

(1) 读入训练数据；

(2) 绘制因变量和各个自变量的箱线图（提示：可以对右偏分布的数据取对数）；

(3) 以是否流失为因变量，对自变量进行标准化（使其均值为 0，方差为 1，提示：在 R 中可使用 scale 函数），建立逻辑回归模型，给出系数估计结果，并对结果进行解读（提示：使用 glm() 函数建立逻辑回归模型）；

(4) 使用建立好的逻辑回归模型，分别对训练集和测试集进行预测，得到每个用户的预测流失概率值（提示：使用 predict() 函数进行预测）；

(5) 借助问题 4 中预测的结果，分别绘制训练集和测试集上预测结果的 ROC 曲线，计算相应的 AUC 值，并根据 ROC 曲线和 AUC 值对模型进行评价（提示：使用 R 包 pROC 中的 plot.roc() 函数绘制 ROC 曲线）。

提交时间：10 月 28 日，18:30 之前。请预留一定的时间，迟交作业扣 3 分，作业抄袭 0 分。

附：客户流失预警数据集介绍

- 训练数据集：sampledata.csv
- 测试数据集：predata.csv

数据文件来自国内某运营商，数据已经进行了清理，数据集中的变量包括：是否流失（churn）、在网时长（tenure）、当月花费（expense）、个体的度（degree）、联系强度（tightness）、个体信息熵（entropy）、个体度的变化（chgdegree）、花费的变化（chgexpense）共 8 个变量。具体的变量说明表如下所示：

	变量名		详细说明	备注
因变量（下月）	churn	是否流失	1=流失 0=不流失	流失率 1.25%
自变量（当月）	tenure	在网时长	连续变量 单位：天	客户从入网到截止数据提取日期时在网时间
	expense	当月花费	连续变量 单位：元	客户在提取月份时的花费总额
	degree	个体的度	连续变量 单位：人数	和客户通话的总人数，去重之后的呼入与呼出加总
	tightness	联系强度	连续变量 分钟/人	通话总时间除以总人数
	entropy	个体信息熵	连续变量	$E_i = -\sum_{a_{ij}=1} p_{ij} * \log(p_{ij})$，其中$E_i$为个体 i 的信息熵，$a_{ij} = 1$代表个体 i 和 j 通过电话，$p_{ij}$代表 j 和 i 通话的分钟数据占 i 总通话分钟的比例
	chgdegree	个体度的变化	连续变量 单位：%	（本月个体的度-上月个体的度）/上月个体的度
	chgexpense	花费的变化	连续变量 单位：%	（本月花费-上月花费）/上月花费