HOMEWORK 4

1. 证明题

(1) Suppose for the ith subject we observe x_i and y_i. Let $p(x_i;\beta) = P(Y=1|X=x_i)$. Maximum likelihood estimation:

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{N}\Big\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\Big\} \\ &= \sum_i \Big\{y_i x_i^\top \beta - \log\left(1+\exp(x_i^\top\beta)\right)\Big\}\end{aligned}$$

Please derive the blue part.

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

Reminder: you need to derive the equation

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

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 $R = 200$ times. For each j, draw $(\widehat{\beta}_j^{(r)} - \beta_j)$ in boxplot for $N = 200, 500, 800, 1000$. ($\widehat{\beta}_j^{(r)}$ is the estimation for β_j after r iterations.)

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

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

理解：若正例的预测值小于反例，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，推荐使用 R 语言，提交 rmarkdown 输出的报告。具体任务见 word 文档。

最后以 HTML/PDF 的形式提交报告。报告中需包括题目内容中涉及的代码和相关文字解释、结果分析。

提交时间：11 月 9 日，晚 20:00 之前。请预留一定的时间，迟交作业扣 3 分，作业抄袭 0 分。