STA101: 魅力统计 Fascinating Statistics

作业 3

注: 使用 RStudio 作答的题目, 请在答案最后附上 R 语言代码.

1. (10 分) 在线性回归模型$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$下, 假设$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$独立并都服从$N(0, \sigma^2)$, 请使用极大似然法估计$\beta_0$和$\beta_1$, 得到$\hat{\beta}_0$和$\hat{\beta}_1$的形式, 并与最小二乘法的估计量比较, 它们是否有区别?

2. (10 分) 为什么在逻辑回归模型的参数估计中, 不是去最小化$\sum_{i=1}^{n}(y_i - \hat{p}_i)^2$? 其中$y_i$表示观测的类别, 取值为 0 或 1, $\hat{p}_i$表示逻辑回归模型得到的对$P(Y_i = 1)$的估计。

3. (15 分) 我们在讲逻辑回归模型的参数估计时提到, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$没有显式表达式, 需要借助梯度下降法、牛顿法等优化算法。请你自行学习这两个优化算法分别是什么, 有什么区别和联系。如果你是老师, 现在想要向学生讲述这两个优化算法, 你会怎么讲? 请把你的讲述内容用 PPT 呈现出来(提供 PPT 截图)。

4. (20 分) 在课程中讲到的 titanic 数据中, 定义一个名为 FamilySize 的新变量, 它等于 SibSp+Parch+1, 即乘客随行的兄弟姐妹、配偶、父母、子女的数量加上自己。然后再基于 FamilySize 定义一个家庭大小的类别型变量 FamilySizeCat: 如果 FamilySize = 1, 则 FamilySizeCat = ‘Single’; 如果 FamilySize 在 2 到 4 之间, 则 FamilySizeCat = ‘Small’; 如果 FamilySize 大于等于 5, 则 FamilySizeCat = ‘Large’. 注: 本题请使用 RStudio 作答.

   (1) 使用 ggplot()函数, 按家庭大小类别画出乘客的幸存情况的柱状图。(5 分)

   (2) 把 FamilySizeCat 加入课程中拟合的逻辑回归模型, 得到 FamilySizeCat 的两个回归系数(请提供模型拟合结果系数的截图), 并对这两个回归系数作出具体解释。(10 分)

   (3) 画出(2)中所拟合模型的 ROC 曲线, 与课程中所拟合模型的 ROC 曲线画在同一张图上(用不同的颜色, 请提供你所画图形的截图), 比较这两个模型的优劣。(5 分)

5. (15 分) k 近邻法虽然是一个非常直观和简单的方法, 但其在处理大规模数据集时的计算效率并不高, 在对新实例进行分类预测时, 需要遍历整个训练数据集以找到最近的 k 个邻居. 为提高 k 近邻法的计算效率, 有人提出可以利用 KD 树(k-dimensional tree)这一数据结构, 在近邻搜索中快速排除与目标距离较远的区域, 提高搜索效率. 请点击这里阅读百度百科关于 KD 树的介绍, 特别是结构实例、构建算法、查找算法部分.

(1) 请仿照百度百科介绍中 KD 树的构建实例，给定 7 个二维数据点{(1, 6), (2, 7), (3, 2), (4, 9), (5, 5), (7, 8), (8, 4)}，构建一棵 KD 树. 给出空间划分的详细步骤，并绘制最后生成的 KD 树. (10 分)

(2) 请仿照百度百科介绍中最近邻的查找算法，基于(1)中生成的 KD 树，以欧氏距离为距离度量，搜索实例(8.5, 5.2)的最近邻点. 给出搜索的详细步骤. (5 分)

6. (10 分) 在课程 6.4 节中，我们使用 titanic 数据演示了朴素贝叶斯法的求解过程，其中仅考虑了性别、乘客等级、年龄段 3 个特征. 利用上面第 4 题定义的乘客随行家庭大小的类别型变量 FamilySizeCat，使用朴素贝叶斯法对不同乘客幸存的概率进行估计，将结果填写至如下表格中: (答案保留 4 位小数)

Sex	Pclass	AgeCat	FamilySizeCat	幸存概率估计
Male	1	Child	Single	
Male	1	Child	Small	
Male	1	Child	Large	
Male	1	Senior	Single	
Male	1	Senior	Small	
Male	1	Senior	Large	
Male	3	Child	Single	
Male	3	Child	Small	
Male	3	Child	Large	
Female	3	Child	Single	
Female	3	Child	Small	
Female	3	Child	Large	

7. (10 分) 假设有观测数据$\{(y_i, x_{i1}, x_{i2}),\ i = 1, 2, \ldots, 5\} = \{(1, 1, 2), (1, 2, 3), (1, 3, 3), (-1, 2, 1), (-1, 3, 2)\}$. 试求出最大间隔分离超平面的表达式，并手动画一个二维散点图，在图上画出该分离超平面及支持向量.

8. (10 分) 请你出一道与第六章内容有关的作业题，并给出解答。题目和解答各占 5 分。评分会考虑题目的趣味性与难度(难度适中，不能太简单)。