STA 521: 项目2云数据

发布日期：**10月20日，星期四**

应付时间**12月6日（星期二）下午3点**

请仔细阅读!

- 在做项目之前，最好重温一下你的笔记、幻灯片和读物；并综合他们的主要观点。
- 我们对向Gradescope提交的错误/迟到的文件（请不要用电子邮件）采取零容忍政策。
- 这个项目的建议工作时间至少是24小时（至少12小时/人）。提前计划，尽早开始。
- 你需要提交两件事

  (a) 需要向Gradescope提交一份由Latex、Rnw或Word生成的主pdf报告（**字体大小至少为11pt，少于或等于12页**），即 "PROJ2写法"。
  - 提供一流的（研究论文级别的）写作，有用的、标记清楚的数字，并且在这个pdf中没有代码。紧凑地安排文本和数字（.Rnw可能对这一点不大有用）。
  - 你可以根据自己的喜好为报告选择一个标题和一个团队名称（*发挥创造力！*）。请在标题下面提供你的队友的姓名和电子邮件。
  - 你的报告应该以致谢部分结束，在这里你要简要讨论每个成员的贡献，**以及**你使用的再资源，记下你所得到的所有帮助，并简要概述你进行项目的方式。

  (b) 一个包含以下内容的压缩文件到Gradescope "PROJ2代码"（见第5节）。
  (i) 用于生成报告的原始Latex、Rnw、Qmd或Word。
  (ii) 你的R代码（在一个单独的R文件中使用CVmaster函数）。
  (iii) 一个README.md文件，详细描述了如何从头开始复制你的论文（假设研究人员可以访问图像）。

- **要直观*和*量化。**请记住，与家庭作业相比，项目的评分是不同的，只有一行答案而没有解释通常是不够的。做到

你的研究结果要简洁明了，并试图用数字和数据支持的良好论据来说服我们。设身处地为读者着想，大声朗读报告通常会有帮助。这一次的评分标准*非常高*。我们将对数字非常挑剔：缺乏适当的标题和轴标签将导致失去几分。对调整参数/超参数的选择缺乏解释也会导致失分。

项目概述

这个项目的目标是根据美国宇航局Terra卫星上的MISR传感器自动记录的辐射度，对极地地区的云层探测进行探索和建模。 你将尝试建立一个分类模型，利用现有的信号/特征区分图像中是否有云。你的数据集有 "专家标签"，可以用来训练你的模型。当你评估你的结果时，设想你的模型将被用来区分大量没有这些 "专家 "标签的图像上的云和非云。

在Sakai/Resources上，你会发现一个包含三个文件的压缩档案：**imagem1.txt**, **im- agem2.txt**, **imagem3.txt**。每个文件都包含一张来自卫星的图片。每个文件都包含若干行，每行有11列，如下表所述。所有五个辐射角都是原始特征，而NDAI、SD和CORR是基于主题知识计算的特征。关于这些特征的更多信息见**yu2008.pdf**一文。传感器的数据是多角度的，并记录在红色波段。关于MISR的更多信息，见**http://www-misr.jpl.nasa.gov/**。

01	y坐标
02	x坐标
03	专家标签（+1=云，-1=非云，0未标记）。
04	NDAI
05	ÄÄÄ
06	CORR
07	辐射角DF
08	辐射角CF
09	辐射角 BF
10	辐射角AF
11	辐射角AN

表1:云数据中的特征。

1 数据收集和探索（30分）

(a) **写一份半页的**论文**摘要**，至少包括研究的目的、数据、收集方法、其结论和潜在影响。

(b) **总结**数据，即不同类别的像素的百分比。用专家标签的x、y坐标和基于专家标签的区域颜色，**绘制出标记良好的美丽地图**。**你是否观察到一些趋势/模式？对于这个数据集来说，对样本的i.i.d.假设是合理的吗？**

(c) 对数据集进行**视觉和定量EDA**，例如，总结（i）特征本身之间的成对关系和（ii）专家标签与各个特征之间的关系。**你是否注意到**这些特征之间的**差异？**

基于辐射度或其他特征（CORR、NDAI、SD）的两类（云、无云）？

2 准备(40分)

现在我们已经对数据进行了EDA，我们现在准备训练我们的模型。

(a) (数据分割）**将整个数据**（imagem1.txt, imagem2.txt, imagem3.txt）分成三组：训练、验证和测试。仔细思考如何分割数据。**建议至少有两种不同的方法**来分割数据，并考虑到数据的非即期性。

(b) (基准）**报告一个琐碎的分类器的准确性**，该**分类器**在验证集和测试集上将所有标签设置为-1（无云）。在什么情况下，这样的分类器会有很高的平均准确率？*提示：这样的步骤提供了一个基线，以确保手头的分类问题不是微不足道的。*

(c) (第一顺序重要性) 假设专家的标签是真理，在不使用花哨的分类方法的情况下，建议三个 "最佳 "特征，**使用定量和可视化的理由**。清楚地定义你的 "最佳 "特征标准。只有相关的图谱是必要的。一定要仔细考虑这个问题，因为它涉及到后续问题。

(d) 在R语言中编写一个通用的交叉验证（CV）函数**CVmaster**，它将一个通用的分类器、训练特征、训练标签、折叠次数***K***和损失函数（至少分类精度应该有）作为输入，并在训练集上输出***K*折叠**CV损失。请记得把它放在第五节的github文件夹里。

3 建模（40分）

我们现在尝试拟合不同的分类模型，并使用不同的标准评估拟合的模型。在接下来的三部分中，我们希望你能尝试*逻辑回归和至少三种其他方法*。

(a) **尝试几种分类方法，用交叉验证法（CV）评估它们的适合性。对你所尝试的方法的假设进行评论，并说明这些假设在本案例中是否得到满足。**由于CV没有验证集，你可以合并你的训练集和验证集来适应你的CV模型。**报告**不同褶皱的准确性（而不仅仅是不同褶皱的平均值）和测试准确性。应该报告创建折叠的两种方式的CV结果（如第2(a)部分中的回答）。对结果进行简要评论。确保你诚实地提及你所尝试的所有分类方法。

(b) **使用ROC曲线来比较不同的方法。**选择一个临界值，并在ROC曲线上突出显示。解释你对截止值的选择。

(c) (奖励)使用其他相关的指标来评估配合度。使用量化的措施，并展示干净的、可解释的数字!

4 诊断学 (50分)

*免责声明：*本节中的问题是开放式的。要有视觉效果和定量!黄金标准的论据将能够说服美国国家航空航天局（NASA）使用你的分类方法--在这种情况下，将获得奖励分。

(a) 通过展示一些诊断图或与收敛或参数估计有关的信息，对你选择的一个好的分类模型做一个深入的分析。

(b) 对于你的最佳分类模型，你是否注意到错误分类误差中的任何模式？同样，使用定量和可视化的分析方法。你是否注意到特定区域的问题，或特定范围的特征值？

(c) 基于4(a)和4(b)部分，你能想到一个更好的分类器吗？你认为你的模型在未来没有专家标签的数据上的效果如何？

(d) 当你修改分割数据的方式时，你在第4(a)和4(b)部分的结果是否有变化？

(e) 为你的结论写一个段落。

5 可重复性（10分）

除了上述结果的写法外，请提交一个压缩文件，其中包含重现你的写法所需的一切，以Gradescope "PROJ2代码"。具体来说，想象一下，在某一时刻，在三个图像文件中发现了一个错误，未来的研究者想用新的、经过修正的图像文件检查你的结果是否成立。这位研究人员应该能够轻松地重新运行你的所有代码，并生成你的所有数字和表格。这个压缩文件应该包含。

(i) 用于生成报告的原始Latex、Rnw、Qmd或Word。

(ii) 你的R代码（在一个单独的R文件中使用CVmaster函数）。

(iii) 一个README.md文件，详细描述了如何从头开始复制你的论文（假设研究人员可以访问图像）。