AM-217 Regression Analysis May 13, 2021

Homework 5

Instructor: Cheng Hua 花成 *Due Date: May 26, 2021, 10:00pm*

1 Review Data

In this problem, we will examine a review dataset. There are, as usual, two parts to this dataset:
在这个问题中，我们将检查一个评论数据集。这个数据集有两个部分:

- yelp_train.csv: for training your models, and model selection
- yelp_test.csv: for assessing prediction accuracy

Each row corresponds to a single review from a user rating a particular business. Each review consists of a 1 to 5 star rating as well as a text commentary review. The variables included are as follows:
每一行都对应着一个用户对某一特定商家的单一评论。每条评论由 1 到 5 颗星的评价以及文字评论组成。包括的变量如下。

- date: the date the review was posted
- words and characters: the number of characters or words in the text portion of the review
- stars: the number of stars given by the user in this review
- funny, useful, cool: number of ‘funny’, ‘useful’, and ‘cool’ votes this review received
- user_funny, user_useful, user_cool: number of ‘funny’, ‘useful’, and ‘cool’ votes ever received by the user writing the review
- user_average_stars: the average number of stars provided by the user across all his/her reviews
- user_review_count: the number of reviews this user has previously written
- biz_review_count: the number of reviews this business has ever received
- biz_open: whether the business is open (as opposed to closed permanently)
- biz_stars: the average number of stars across all reviews for this business (rounded)

Our goal is to use methods learned in this course to predict whether or not a review is a positive review (which we'll define as one that is a 4 or 5-star review).
我们的目标是使用在本课程中学到的方法来预测一条评论是否是正面评论（我们将其定义为 4 或 5 星的评论）。

Part 1

Create a column called positive which is equal to 1 if the review is positive and 0 otherwise. Make any other transformations to the dataset you determine to be helpful or necessary.
创建一个名为 positive 的列，如果评论是正面的，它就等于 1，否则就是 0。对数据集进行任何其他你认为有帮助或必要的转换。

Part 2

Try any appropriate methods you feel appropriate to predict positive.
尝试一些你觉得合适的方法来预测 positive。

Part 3

Summarize your findings and show all steps in model selection as well as the tuning of any parameters. You should provide some interpretation of the coefficients.
总结你的发现，并展示模型选择的步骤以及任何参数的调整。提供一些对系数的解释。

2 Bone Data

For this problem, we will model changes in bone mineral density as a function of age using various non-linear regression methods.
对于这个问题，我们将使用各种非线性回归方法，将骨矿物质密度的变化作为年龄的函数来建模。

Here is a description of the variables in dataset bone.csv:

- idnum - identification number of the individual
- age
- gender
- spnbmd - relative spinal bone mineral density measurement

Each value of spnbmd is the difference in spinal bone mineral density measurements taken on two consecutive visits, divided by the average. Each value of age is the average age over the two visits. Even though we have repeated-measures data, we will simply ignore this fact (effectively ignoring the first column in the dataset). spnbmd is our response variable and age and gender will be our predictors. Using non-linear regression methods, you could examine differences in the timing of female vs. male adolescent growth spurts.
spnbmd 的每个值是连续两次就诊时测量的脊柱骨矿物质密度的差异，再除以平均值。年龄的每个值是两次访问的平均年龄。尽管我们有重复测量数据，但我们将简单地忽略这一情况（实际上，我们忽略了数据集中的第一列）。spnbmd 是我们的响应变量，年龄和性别将是我们的预测变量。使用非线性回归方法，你可以研究女性与男性青春期生长突增的时间差异。

Part 1

Begin with a plot of spnbmd against age, colorcoded by gender. Does it appear that there are differences in bone mineral density (BMD) trajectories between males and females?
画 spnbmd 与年龄的关系图，并按性别进行颜色编码。男性和女性之间的骨矿物质密度（BMD）轨迹是否存在差异？

Part 2

Now, you will fit a series of models to spnbmd using a polynomial of age for each gender separately. Use knots at c_1 and c_2, where c_1 and c_2 are the 33rd and 67th percentiles of the data, respectively. That means you should have 3 bins on age: ‘age $\leq c_1$’, ‘$c_1 <$ age $\leq c_2$’, and ‘age $> c_2$’.

For each of the following models, you should (1) construct the model, (2) state the number of coefficients fitted, including the intercept, and (3) show a plot like the one in Part 1 with your fitted models.
现在，你将对 spnbmd 进行一系列的模型拟合，对每个性别分别使用年龄的多项式。在 c_1 和 c_2 处使用结点，其中 c_1 和 c_2 分别是数据的第 33 和第 67 个百分位。这意味着你应该有 3 个关于年龄的分类："age$\leq c_1$"，"$c_1 <$age$\leq c_2$"，以及"age$> c_2$"。

对于以下每个模型，你应该 (1) 构建模型，(2) 说明拟合的系数数量，包括截距项，(3) 用你的拟合模型展示一个类似于第 1 部分的图。

The models to fit are:
需要拟合的模型是:

a) step functions

b) cubic spline

c) natural spline

d) Piecewise quadratic

e) Continuous piecewise quadratic

In this part, *d*) and *e*) are a bit harder. Please follow the slides and see how you would construct the models in python.
在这一部分，*d*) 和 *e*) 有点难度。请按照课件的内容，看看你将如何在 python 中构建这些模型。

Part 3

Describe which model in Part 2 appears most suited to the data, based on visual inspection. Are there aspects of this fitted model that seem concerning?

For this best approach, implement 5-fold cross-validation to select between using 2, 3, 4, 5, 6, or 7 knots, spaced out at uniform percentiles of the data. For example, with 2 knots, we would want the knots to be placed at the 33rd and 67th percentiles, as we did above. What choice of knots works best? Refit a model on all of the data with this number of knots and plot the model.
检查各个图像，描述第二部分中的哪个模型最适合于数据。这个适合的模型是否有需要注意的方面?

对于这种最佳方法，实施 5 倍交叉验证，在使用 2、3、4、5、6 或 7 个结点之间进行选择，在数据的百分位上进行等比例间隔。例如，如果使用 2 个结点，我们希望结点放在第 33 和第 67 个百分位数，就像我们上面做的那样。选择几个节点最有效? 在所有的数据上用这个节点数量重新建立一个模型并画图。

What to Submit

A compiled .ipynb file with codes and results in PDF format documenting your work.

There are many ways to do it:

- In Jupyter notebook, click File → Download as → PDF via LaTeX(.pdf)
- Print the page and save it as PDF
- Use markdown and convert it to PDF