AM-217 Regression Analysis April 23, 2021

Homework 4

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

Background

In this homework assignment, we will consider the classification problem of identifying handwritten digits. The dataset consists of pixel data for 16 by 16 images, each of which contains a single handwritten digit from 0 to 9. Each pixel is either on (1) or off (0).

There are 2 datasets:

- digits_train.csv - a training set
- digits_test.csv - a test set

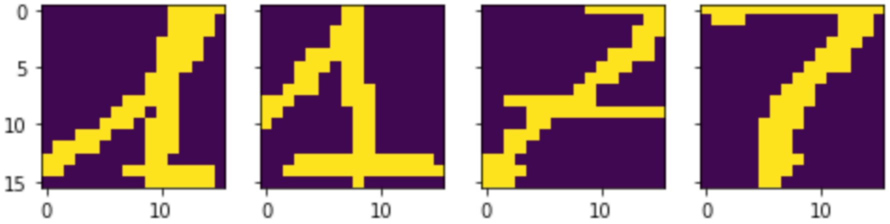
在这个作业中，我们将考虑识别手写数字的分类问题。该数据集由 16 乘 16 的图像的像素数据组成，每张图像包含一个从 0 到 9 的手写数字。每个像素要么开（1），要么关（0）。一共有训练与测试两个数据集。

The training and the test set each have 257 columns. The first 256 (our predictors) are binary values (0 or 1) indicating whether the pixel is activated. The last column is a ‘digit’ column with the actual digit that was meant to be represented (the true class label). You can plot the digits using the following code.

训练集和测试集各有 257 列。前 256 列（我们的预测因子）是二进制值（0 或 1），表示该像素是否被激活。最后一列是一个”数字”列，为实际的数字。你可以用如下的程序来显示数字。

```
def plotDigit(k, dat):
    f = np.reshape(np.array(dat.loc[k][0:256]), [-1, 16])
    print("row: " + str(k) + "| digit: " + str(dat.loc[k][256]))
    return f

fig, (ax1, ax2, ax3, ax4) = plt.subplots(1, 4, sharey=True, figsize=(8, 2))
ax1.imshow(plotDigit(7, train))
ax2.imshow(plotDigit(53, train))
ax3.imshow(plotDigit(151, train))
ax4.imshow(plotDigit(622, train))
fig.show()
```



Part 1: Logistic Regression

(1) Which digits do you think will be most difficult to distinguish between or classify? Show a few images (no more than 6) to justify your answer.

你认为哪些数字将最难区分或分类？请展示一些图片（不超过 6 张）。

(2) Based on your answer to (1), pick any pair of two numbers that are difficult to distinguish. Run a logistic regression. Make a confusion matrix on both the training set and the test set.

根据你对（1）的回答，选择任何一对难以区分的两个数字。运行一个逻辑回归。在训练集和测试集上分别做一个混淆矩阵。

(3) Pick any pair of two numbers that you feel are easy to distinguish. Run a logistic regression. Make a confusion matrix on both the training set and the test set.

选择任何一对你觉得容易区分的两个数字。运行一个逻辑回归。在训练集和测试集上分别做一个混淆矩阵。

(4) Briefly describe your observations from (2) and (3).

简要描述你在（2）和（3）中的观察结果。

(5) Show 1 or 2 plots if there are any wrong predictions on the test set.

如果在测试集上有任何错误的预测，显示 1 或 2 个你预测错误的图。

Part 2: Multinomial Logistic Regression

(1) Run a multinomial logistic regression of all classes. Examine a confusion matrix on training set. What do you observe?

对所有的数字进行一个多元逻辑回归。检查训练集的混淆矩阵。你能观察到什么?

(2) Examine a confusion matrix on test set. Which digit(s) is(are) relatively difficult to classify? How does this compare to your initial guesses from Part 1?

检查测试集的混淆矩阵。哪一个 (些) 数字是比较难分类的? 这与你在第一部分中的最初猜测相比有什么不同?

(3) What is the prediction accuracy for the test set?

测试集的预测精度是多少?

(4) Show 3 or 4 plots if there are any wrong predictions on the test set.

如果在测试集上有任何错误的预测，显示 3 或 4 个你预测错误的图。

Part 3: With Ridge Penalty and Cross-Validation

(1) Apply the ridge penalty to fit a multinomial logistic regression model. We will use the following built-in l2 penalty for ridge regression, and C is the hyperparameter for you to choose. We use solver='lbfgs'. It is the best solver choice for medium-sized dataset.

应用岭回归来拟合一个多元逻辑回归模型。我们将使用以下内置的 l_2 惩罚来进行岭回归，C 你需要选择的超参数。我们使用 solver='lbfgs'。它是中小型数据集的最佳 solver 选择。

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(penalty = ’l2’, solver=’lbfgs’, multi_class=’multinomial’,
     C = for_you_to_choose)
```

Specify your own selection range and use 10-fold cross-validation on the training set to select your C. Show a plot that displays the accuracy error on the y-axis and values of $\log(C)$ on the x-axis. (Note: You will need to use the GridSearchCV function, and the scoring function is 'accuracy' in classification problems.)

指定你自己的 C 的选择范围，在训练集上使用 10 倍交叉验证来选择你的 C。显示一个图，在 Y 轴上显示交叉验证的预测精度，在 X 轴上显示 $log(C)$ 的值。(注意：你可以使用 GridSearchCV 函数，在分类问题中，评分函数是”accuracy”。)

(2) Report C_{min} (the C value that achieves the best cross-validation accuracy). Report the resulting prediction accuracy on both training and test sets.

找到 C_{min}（达到最佳交叉验证精度的 C 值）。分别报告训练集和测试集上的预测精度。

(3) For $C = C_{min}$, how many zero coefficients are estimated in total for this value of C? Does this meet your expectation?

对于 $C = C_{min}$ 的值，总共有多少系数被估计为 0？这符合你的期望吗？

(4) Do you expect lasso regression to give a better or worse result? Please briefly answer. You don't need to implement it.

你认为 lasso 回归的结果会比 risge 是好还是坏？请简要地回答。你不需要实现它。

Part 4: Dimension Reduction

(1) Transform the original X with 30 principal components from PCA. (Since the X only consists of 0s and 1s, you don't need to standardize it.) Then, redo Part 3 (1) and (2) with these 30 principal components. What do you observe?

用 PCA 对原始 X 进行转换, 我们先使用前 30 个主成分。(由于 X 只由 0 和 1 组成，你不需要对它进行标准化。) 然后，用这 30 个主成分重新做第 3 部分的（1）和（2）。你观察到了什么？

(2) Now use 50 principal components. What do you observe?

现在使用前 50 个主成分，你发现了什么？

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