ADM first challenge

Jaime Davila

2/23/2022

Due date:

Wednesday March 16th-6pm.

Introduction

In class we have gone through the process of building for distinguishing a `2` from `7` by using some predefined features. In this challenge you will use the `mnist` dataset from `dslabs` to create an end-to-end classifier that distinguishes between two digits that will be given to each group.

Instructions

The end product of this challenge will be a knitted Rmd document which should contain the following sections:

Dataset creation

- Your dataset should have in total 1000 randomly selected digits (feel free to use a `set.seed` command so that your results are reproducible).
- Your training dataset should have 800 observations and your testing should have 200 observations.

Feature definition

- You are allowed to use only *2 features*. Notice that you need to calculate those features directly from dataset. Make sure to describe what those features represent and why you chose them. Are those features capturing any intuition that you have about distinguishing those two digits?

Model creation, optimization and selection

- Create at least two different models for this classification and make sure to optimize the parameters those models have.
- Calculate the missclassification rates for both models and select the model with the lowest error rate.

Visualization

- Plot the probabilities across a grid and the decision boundary for your selected model

Changing things up

- Create a new dataset that includes your two chosen digits and the digit `5`. Create training and testing datasets that include `5` and your two given digits.
- Calculate the same 2 features for this new testing and training dataset.
- Calculate the missclassification rate on this new dataset. Create also the confusion matrix and comment on what digits seem to get confused more and why.
- Plot the probabilities across a grid and the decision boundary for your model.