Technical Report: Digit Recognition

Group 6: Yilin Li, Hien Nguyen, Lyn Peterson

12/6/2019

Your technical report should be an .Rmd file that contains the following sections. So as not to make the compilation (knitting) of the document not take too long, consider setting cache = TRUE in the curly braces of any R chunk with substantial computing. Please knit both to pdf and github document (.md).

Abstract

A brief overview of the area that you'll be investigating, the research question(s) of interest, your approach to analysis, and the general conclusions.

Overview:

Research question: Can we build classifiers to recognize what the digit (0-9) is in a given image based on 60,000 training images?

Approach to analysis:

General conclusions:

Introduction

Overview of the setting of the data, existing theories/models (particularly if you are working in a descriptive/inferential setting), and your research questions.

The Data

Where does the data come from? How many observations? How many variables? What does each observation refer to (what is the observational unit)? What sorts of data processing was necessary to get the data in shape for analysis?

The data comes from the MNIST database of handwritten digits. The digits have been size-normalized and centered in a fixed-size image.

The data is split into a training set of 60k images and a test set of 10k images.

Data processing: Since the pre-downloaded data is already well-formatted, minimal effort was spent on data processing. We performed two additional steps of data processing:

(1) Each 28x28 image was flattened into a single row of 784 pixels.

(2) All pixels were rescaled from 0-255 to 0-1.

Exploratory Data Analysis

Explore the structure of the data through graphics. Here you can utilize both traditional plots as well as methods from unsupervised learning. Understanding the distribution of your response is particular important, but also investigate bivariate and higher-order relationships that you expect to be particular interesting.

- The digit classes are well-separated.
- Handwriting patterns and variations.
- PCA & early predictions.

Modeling

Construct (descriptive and/or predictive) (classification and/or regression) models that address your research questions. You are encouraged to fit many different classes of models and see how they compare in terms the bias/variance tradeoff (do you have a Rashomon effect going on?). Also be sure to guard against overfitting through cross-validation or shrinkage/penalization (don't forget about ridge regression and the lasso).

This will be the most extensive section and will include your results as well.

- Classification Tree (with pruning)
- Random Forest (with bagging)
- Boosted Tree
- Logistic Regression for Multinomial Distribution (with Ridge and Lasso)
- KNN
- SVM

All models are cross-valiated.

Discussion

Review the results generated above and sythensize them in the context from which the data originated. What do the results tell your about your original research question? Are there any weaknesses that you see in your analysis? What additional questions would you explore next?

- 4-9 is the most difficult pair to predict across models
- No overfitting with PCA, but some with raw data
- Best model
- Compare with published models

References

At minimum, this will contain the full citation for your data set. If you reference existing analyses, they should be cited here as well.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998. Online Version.