Methods of Statistical Learning: 36-462/36-662
Final Project, Spring 2024

Grain Image Classification

Deliverables and deadlines

Here are some key deliverables and deadlines upfront:

- This project is due on April 22 at 11:59pm, submitted on Canvas.
- You can team up with (at most) one classmate to work on the project. Each team only has to submit one of everything. Hence you can read "you" throughout as "your team".

Disclaimer

This project is an open-ended work-in-progress. If you find any issues with the data or have difficulty interpreting something in the following description please post on Piazza and we will update this document as necessary.

Introduction

This project concerns predicting the grain type (binary) using 16 geometric features extracted from the images of grains. All features are computed from pixel-based digial images using computer vision techniques. You can view this project as a simple and primitive version of the algorithms used in Applications such as "Seek" and "Merlin Bird ID" that identifies plant and animal species by their images or sound recordings.

Your project has a few parts:

1. Data exploration and pre-processing.
2. Building and validation of predictive algorithms.
3. Actual submission to a prediction contest.
4. Some follow-up analysis of your results.

More details

Download the file `train_data.csv` and load it into your R session. This is the training data that you will use for model building. Your prediction target is the `Y` variable.

Data pre-processing and exploration: You will need to do some pre-processing or exploration of your data. Particularly, many of the features are highly correlated, which may cause ordinary linear models to be unstable. Consider appropriately drop/combine some of the features, or use regularization. You may also realize that some features are nonlinear transforms of others.

General advice for training and tuning your predictors: Split your data into a train and a validation set. Use the train set to fit and the validation set to evaluate and select a good model. Use small subsets of the data initially until you get a feeling for what works and what does not.

I have also provided you with a test set containing only the features. Using an external source to obtain the labels for the test set and using this to tune your model is considered cheating. **Do not do this**.

Making predictions

How can you make your predictions? You can use any of the techniques we have discussed in class. You can use any of the variables in the data set, and you can also consider constructing new variables by combining or transforming the variables that are present in the data set. Given the high correlation between the features, combining them seems to make sense. Some of the features are transformations of others and you can discover this from your exploratory analysis. You should not use external information sources for this project.

Submitting predictions

You will submit a single RData file with your predictions. This file should contain the following variables:

1. `y.guesses`: A single vector of the predicted grain type on the test data. This vector should have the same length as the number of test cases in `test_data_x.csv` (**and in the same order**).

2. `test.acc`: A single number indicating your best guess at the 0/1 error rate of your classifier on the test set.

3. `team.name`: A string with your team's name. These may be revealed in class, so make it anonymous if you wish. Your report will link the team name to individual names for grading purposes.

To make this file, if you have the appropriate variables in your workspace, you can type

```
save(list=c("y.guesses","test.acc","team.name"),file="stat462project.RData")
```

This will create `stat462project.RData` file, which you can upload on canvas. (Please rename the file to include your team name before uploading.)

Write-up

Along with your contest entries, you will submit a write-up of your work. This write-up should be a polished report, with figures and snippets of R code as you deem helpful. You don't need to submit your R code in its entirety. Your report should have the following sections (you can of course add subsections if you want), and should be no more than 8 pages.

Introduction: Describe your data set. What is the problem you are trying to solve? This can be quite brief.

Exploration: Exploration of your data. You don't need to do the typical "exploratory data analysis" that you might do in 36-401, but you should provide proper motivation for your work and explain any insights and exploration that led to your features and models. This can include unsupervised approaches that we learned in the last part of the term if you detect any interesting structure with them. (If you don't find anything interesting, then just describe what you tried. You don't need to artificially manipulate the data to find something that's not there.)

Supervised analysis: How did you make your predictions? Describe this process in detail. Again, you can use any of the classification/regression techniques that we learned in the first half of the course, or any other techniques as long as they are adequately described. What predictor variables did you include? How did you engineer features from the data? What technique did you use for prediction, and why did you choose it? If there were tuning parameters, how did you pick their values? Can you explain anything about the nature of the relationship between the predictors in your model and the predictions themselves?

Analysis of results: Once you have predictors you happy with, you should think about their performance a little more. At the least, you should try to address the following questions:

- What kinds of sample points do you do well or poorly on?
- Suppose you had more time, what would your next steps be, i.e. what would you like to try next to improve your predictive performance?

Any other analysis of your results would be welcome here.

Evaluation

Your predictions will be evaluated against the true values. The overall performance will be measured based on your mis-classfication error for the classification task, and the agreement between the actual accuracy and your guessed accuracy. We will try our best to provide the results and feedbacks to you. The total score of the project will be 60% from the test error rate and your guess of test accuracy, and 40% from the report.

Cheating

Don't cheat. We know that there are ways to cheat on this final project. If we suspect you of cheating (e.g., if you have a remarkably low misclassification rate/low mean squared error, but your method is not really statistically motivated), then we reserve the right to give you a 0.