Statistical Learning for Data Science DS 64640

UNIVERSITY OF NOTRE DAME | DATA SCIENCE Naughton, Fall 2019

Statistical Learning for Data Science
Fall 2019

Final Project Presentation (Group) 11/04/2019
FInal Project Written Report and Reflection (Individual) 11/06/2019

Description

The goal of this project is to identify and explore interesting aspects of data in a real-world context, provide extensive explanations about each step, and report on your analysis of the results. The final project will involve communicating results of your analysis and technical results in a group presentation that effectively translates your work. You will work in groups of 3-4 students for the group presentation component and turn in an individual write-up component and a short reflection assignment.

The project will consist of the following tasks:

- Analyzing data and selecting the model best suited to solve the problem.
- Implementing and optimizing algorithms discussed in this course.
- Using Python to perform computations and analyze the results.
- Completing a slide deck for the group presentation.
- Effectively communicating methodology and results during the group presentation.

Project Problem - Classification

Given numerous images, create an image recognition algorithm to accurately identify images and explore how these images are related. Identify and discuss interesting aspects of the data. Optimize an algorithm to correctly identify images with the lowest possible error rate.

You will use the CIFAR-10 image dataset (https://www.cs.toronto.edu/~kriz/cifar.html), a widely used dataset for machine learning and image recognition. The data consists of 10 classes of images: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Each of the images are 32x32 color pixels with 6,000 images per class.

The original problem included 50,000 images for training and 10,000 for testing. You do not need to use the entire dataset. For this project, you should formulate a binary classification that you want to solve with this dataset. You will pick **two** classes of images for this problem. For example, you may choose to build a model that classifies images as either cats or dogs, or you might relabel the dataset to classify the images as objects or animals.

Although the goal is to build a good classifier, you are not expected to create the most accurate classifier possible. Instead you will be evaluated on your methodology: how you defined the problem, the process by which you solved the problem (exploratory data analysis, preprocessing, modeling and optimization) and how well you clearly communicate your analysis (group presentation and individual write-up).

As a data scientist you should be able to turn an ambiguous problem into a practical one problem that you can solve. For this reason, you are not required to use all of the data as it is. Feel free to use a subset of the original data, or use different sized training and test sets, but be able to defend the choices you make in defining the problem. It is better to have an effective solution to your problem, even if it isn't the optimal one.

Deliverables

A. Individual Write-up
B. Group Presentation
C. Project Reflection

A. Individual Write-up

You are to work individually on and turn in a written report summarizing the results of your project and your code. The actual work can and should be done as a group. You can work together to use the same code, analyses, results and plots. But this write-up describing the work you did should be in your own words.

The report should be:

- 3–10 pages in length (not including graphs)
- In 12-point font, double spaced, with 1-inch margins
- Saved as PDF

Your **individual write-up** should consist of the following sections:

I. **Define the Questions**

- A. Formulate a binary classification problem using the CIFAR-10 dataset.

II. **Get the Data**

- A. Describe the steps you used to prepare the data to answer the questions for this classification problem.
- B. Explain how you split the data for analysis and evaluation, and why you did it that way.
- C. Did you make any changes to the data or labels?

III. **Explore the Data**

- A. Perform exploratory data analysis (EDA) to first understand the dataset and hypothesis what features or methods might be useful (e.g. what features of the images are most similar or different?)
- B. Summarize your findings from the EDA.

IV. **Preprocessing**

- A. Defend any preprocessing and explain why you chose those particular method(s).
- B. How did you handle the color images?
- C. If you used dimension reduction methods, how/why did you optimize it the way you did (i.e. number of components used, include scree plot where possible)?

V. **Initial Model**

- A. Choose an unoptimized models discussed in this course to build a classifier, and explain why you chose them. This should be a flexible model that you can optimize for better performance. This can be a Random Forest, Gradient Boosted Decision Trees, or a Neural Network.
- B. Evaluate the performance of the initial model using the appropriate metrics discussed in this course, and explain why you chose these.
- C. Justify what this model was selected. What are the advantages and disadvantages for this model compared to others?

VI. **Model Optimization**

- A. Optimize the initial model to get better performance metrics.
- B. Discuss the methods used to optimize your model. How did you choose the tuning parameters?

VII. **Figures**

- A. Include figures to compare the performance of the optimized model and base model.

B. Group Presentation

Group Presentation via Zoom Live Session

- 10 minutes via Zoom
- 2–3 mins of Q&A

Your group presentation should deliver important conclusions that you can draw based on the questions you defined, your exploratory analysis phase and the predictive model you build.

Project Audience: Your presentation should answer the questions that are most important to the project audience (the client and stakeholders). The audience is a stakeholder who does not have statistical knowledge however requires information from your presentation and report in order to make decisions for the company.

Group presentations should consist of the following components:

1. Describe the problem you chose to solve.
2. Present a brief version of the EDA (most interesting findings).
3. Explain preprocessing/transformations done to the data and defend your reasoning.
4. Describe the base model that you chose, and discuss the advantages or disadvantages for that model as it relates to the project. Also explain the metrics you chose to use.
5. Explain what you did to optimize the base model that you chose.
6. Compare the optimized model's metrics to the base version and briefly explain differences. Did optimizing help? Did optimizing produce better training results than the base model yet worse test results?

C. Project Reflection

Self-reflection and evaluation are important to your professional growth. You will submit a short report (**approximately 1 page**), reflecting on your own work for the assignment, as well as work by your group members, and the presentations given by your classmates. This can relate to any aspect of the project -- the project chosen, the methods used, the presentation, or your group's workflow. Specifically, you should include:

- A reflection on what you did (or your group did) for your project. Include at least **one thing** that went well, and at least **one thing** that didn't go well. Be specific in your answers.
- Your candid feedback regarding your group members contributions to the project. Did anyone not contribute equally or in a timely manner?

- Include **at least one highlight** that another group (or groups) presented, that you wish you had considered. Be specific in your answers: what did they do, and which group(s) did it?
- A suggestion for your classmates on what didn't work well. Include **at least one opportunity** for improvement. Be specific in your answers: what could have been better, and which group(s) does it apply to?

Your feedback will not be shared with the other groups or group members, but you should write it as if this is a peer evaluation in which you are highlighting what they did well, and suggesting areas for improvement.

Project Submission Instructions

What to submit:

- Group slide presentation (Google slides, PowerPoint or PDF) from **one person** in the group.
- Individual 3–10 page written report (PDF, no code) from **each person** in the group.
- Code/script/Jupyter notebook (PDF) from **each person** in the group.

Project Grading Rubric

	Excellent	**Good**	**Satisfactory**	**Below Satisfactory**
Presentation	Effective storytelling techniques are employed. There is a clear and comprehensive understanding of the work that was carried out. Presentation demonstrates interesting and nuanced insights into the dataset.	Some storytelling techniques are employed. There is a mostly clear and comprehensive understanding of the work that was carried out. Presentation demonstrates some insights into the dataset.	Little effective storytelling techniques are employed. There is some understanding of the work that was carried out. Presentation demonstrates only a few insights into the dataset.	Little to no evidence of storytelling techniques employed in the presentation. No clear rationale as to why or how work was carried out. There are no insights into the datasets. There is no slidedeck.

	Slidedeck is very well-organized and very effectively conveys information and looks professional.	Slidedeck is well-organized and effectively conveys information.	There is a slidedeck but it is not well-organized and/or does not effectively convey information.	
Project Reflection	Completed all required elements of the reflection assignment.	Completed most of the required elements of the reflection assignment (missing 1 part).	Did not complete all required elements of the reflection assignment (missing half).	The reflection assignment was not completed, or only 1 part was complete..
Individual Write-Up	Provides a complete and clear description of the analysis process. Contains good insights into the analysis including what worked and what didn't. Overall, is an accurate understanding of the analysis. Appropriate tone for intended audience. Very few errors in spelling or grammar.	Provides an almost complete and clear description of the analysis process. Contains some insights into the analysis including what worked and didn't work. Overall, is a mostly accurate understanding of the analysis. Appropriate tone for intended audience, but uses some jargon. Few errors in spelling or grammar.	Parts of the description of the analysis process are missing and/or are not clear. Lacking in number and quality of insights into the analysis. Overall, contains many inaccuracies. Written in a tone that is appropriate for the audience. Uses too much jargon. Many grammatical and spelling errors throughout.	Little to no description of the analysis process, or is very unclear.. Has very few insights, or are very low quality. Overall, is an inaccurate understanding of the analysis. Inappropriate tone for intended audience. Would even be difficult for a fellow data scientist to understand. Extensive grammatical and spelling errors throughout.
Code/Script Quality and Correctness	Excellent quality code (is clear and easy to understand). Good usage of comments or documentation. Can be rerun and produce the same results reported.	Good quality code. Adequate usages of comments. Can be rerun and produce the same results reported, but with some errors or inconsistencies.	Poor quality code. Little to no usage of comments. Can be rerun, but does not produce the same results reported.	No code or code that is not readable. No comments. The code does not run due to errors.

Final Project Grade Distribution	
10	Presentation
15	Individual write-up
5	Project reflection
5	Code/script quality and correctness
35	**Total (toward final grade)**