EP2420 Project 1 - Advanced Project

André Silva

November 19, 2020

You can use the tex file of this document as a style file for your report. The sections with heading in small font, e.g. General Rules, are for explanation and are not part of your report. The sections with heading in large font, e.g. Project Overview, belong into your report.

General Rules

The project report should be brief and concise. The language should be precise. Avoid long sentences. Use notations and terms in a consistent way.

Once you are satisfied with a draft, check it for grammar and spelling. Have a colleague proofread the text before you send it in.

Students whose first language is not English often have problems with the proper use of articles. [**?**] contains a short text that explains articles.

Project Overview

Describe in your own words the goal of the project and the main results you achieved in one or two paragraphs.

Background

Introduce important concepts that you apply in the project in short form. Example of such concepts include Principal Component Analysis (PCA), Hyper-parameter search, or random forest regression. Half a page for a concept is sufficient. Provide references where you draw information from for the description of the concept. Name the libraries or software packages you use in the project.

Data Sets and Data Pre-processing

Characterize the dataset(s) your are using in the project. The description can include following items:

1. give the number of samples and features; for selected features, compute statistics of feature values like mean, std, min, max, 25 and 95 quantiles;
2. provide time series plot and a density plot of target;
3. add a heatmap to visualize the correlation among features and between features and target.

Describe the procedures you apply for data cleaning, pre-processing of the data, and outlier removal.

Task I

This section describes the objective, the work, and the results of Task I.

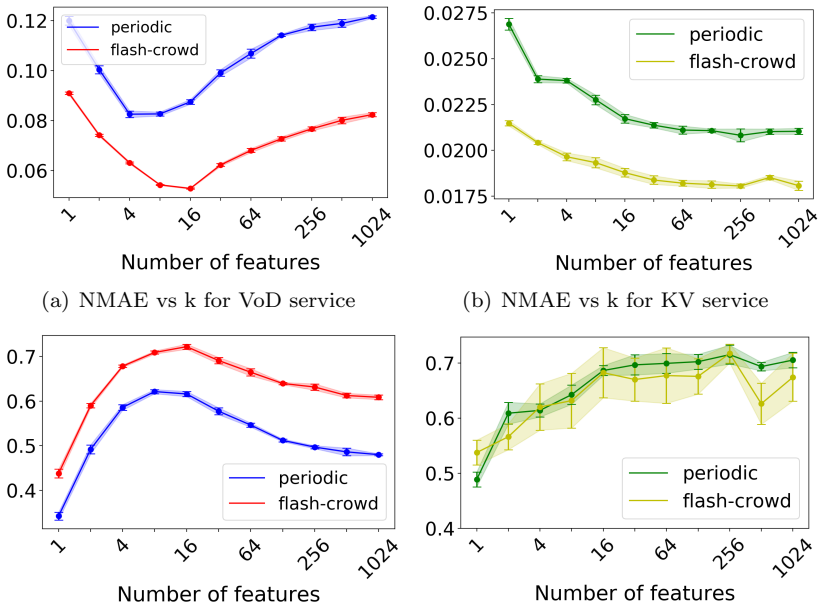
.....

Task V

This section describes the objective, the work, and the results of Task V.

How to report the results of a data analysis

1. Describe precisely the data you use and the methods you employ to obtain the results. The reader should be able to reproduce the results.
2. Include tables and/or graphs in the description.
3. When you give evaluation results with numbers, 3 significant digits are generally sufficient. Providing more digits is often misleading due to uncertainties in measurements, statistics, etc.
4. Make the caption of a table or graph expressive, allowing the reader to grasp the main findings without reading all the surrounding text. See the caption of Figure 1 as an example.
5. For a graph, provide in sequence:
   - Explain the axes, the points on the graph, the bars around the points, etc.
   - Describe the properties of a graph. For example, is it monotonic, does it exhibit non-continuities? etc.
   - Highlight surprising, non-intuitive behavior of the graph with respect to the evaluation.
   - Explain or suggest explanation for the observed surprising behavior.



(a) NMAE vs k for VoD service

(b) NMAE vs k for KV service

(c) R^2 vs k for VoD service

(d) R^2 vs k for KV service

Figure 1: Frame rate and response time prediction on the subspace spanned by the top k features. The vertical axis shows prediction accuracy, the horizontal axis shows the dimensionality of the subspace.

Example of describing the results

Figure 1 shows the evaluation results for all four scenarios. We use blue curves for the scenario VoD periodic load, red for VoD flash-crowd load, green for KV periodic load, and yellow for KV flash-crowd load. The prediction accuracy is measured in $NMAE$ and R^2. One point on a curve represents the average value from ten-fold cross validation. The vertical bar represents the standard deviation. In most cases, the standard deviation is small and barely visible.

The curves in Figure 1(b) follow our expectation: when the number of dimensions increases, the prediction error falls monotonically (at least up to k=256). In contrast, Figure 1(a) shows that, for the VoD service, the prediction error can be significantly lower for small values of k than for the full feature space X where $k = 1409$. This is surprising and suggests that many features do not contribute to an accurate prediction of the frame rate and actually reduce accuracy. For VoD, the optimal value of k with the smallest prediction error is between 8 and 16, while for KV the optimal k is around 256.

Discussion

List and discuss the key findings of each task.
Discuss the difficulties you encountered in the project and how you dealt with them.
Optionally, suggest further investigations one can conduct to achieve the project objectives.

References

[1] U. of Adelaide, "Articles in english grammar," 2014. [Online]. Available: https://www.adelaide.edu.au/writingcentre/sites/default/files/docs/learningguide-articlesinenglishgrammar.pdf