Data 4380

Data Problems
Spring 2022

Amir Farbin

Project Title

Objective

- Source: Kaggle, Keras Example, …
- Data:
  - What is a data point? e.g. “image of a flower”, patient, …
  - What is the types? e.g. images, table of features, time series features, …
  - How many instances? How large? How is it divided?
- Task/Metrics
- Resources

Crypto Forecasting

Objective: Regress Target Metric

https://www.kaggle.com/c/g-research-crypto-forecasting/overview/evaluation

- Describe:
- Source: Kaggle
- Data:
  - What is a data point? Row represents a time period of a specific crypto currency's exchanges. Holds set of features, for example high, low price.
  - What is the types? time series features in csv
  - How many instances? ~ 25 Million data points. Divided into training and example sets
    - How large? 2.9 G How is it divided?
- Task: Regression: predict returns in the near future for prices → Defined a "target" metric
  - Metric: Compute over various crypto and weighted.
- Resources: Good example leading through data visualization and simple linear regression example.

Project Proposal

- ***Abstract***: Short summary of everything below. A few sentences.
- ***Introduction***: (1-2 slides)
  - Create the context. Explain the domain.
  - Plan for rest of the presentation.
- ***Motivation*** (2-3 slides)
  - Why is this topic interesting? Why now?
  - Define the specific problem you will attack. Why do you think it can be done?
  - Previous work. List references.
- ***Problem Formulation*** (5-10 slides)
  - Is this a new problem? Previous work you are reproducing or extending? Is there existing code? What language/libraries have been used?
  - Datasets: sources of data. How it was/will be collected/compiled.
    - Type(s), statistics, size, quality, …
      - What resource will you need?
    - Does it need processing, book-keeping, etc? If so, what needs to be done? What are some tools you can use to do it?
  - What is the goal/task? What type of ML algorithm does it map to?
    - What libraries/tool do you expect to use?
  - What are the metrics used to train the algorithm and to assess it's performance?
    - Estimate expected performance based on previous work or educated guess. What is the state of the art?
  - How will you train, test, and validate?
  - What is the goal of the package you will provide? What could it look like.
- ***Workplan*** (2-3 slides)
  - What do you expect to do for each stage/presentation?