DATA IS POTENTIAL

Importance of Publicly Available Datasets for Developing Machine Learning Solutions

NIST Technical Language Processing Community of Interest

Dr. Nicholas Propes, Seagate, Senior Staff Data Scientist

April 14, 2021

SEAGATE

Seagate Overview

- Manufacturer of data storage solutions (hard drives, etc.)

- Machine Learning used for:
  - Machinery health monitoring
  - Product quality monitoring
  - Marketing
  - Hard Drive Design
  - etc.

- Type of Data:
  - Time Series / Events
  - Images
  - Point Cloud
  - Logs
  - Textual
  - etc.

- Common Challenges
  - Getting the data
  - Understanding the data
  - Labeling the data

What are Public Data Sets and Examples?

- A public data set are datasets made available to the general public

- Where can you find public datasets:
  - Dataset Search: datasetsearch.research.google.com
  - Kaggle
  - UCI Machine Learning Repository
  - TensorFlow / Keras
  - etc.

- Examples (from ubuntupit.com)
  - MNIST
  - ImageNet
  - Twitter Sentiment Analysis
  - Amazon Reviews Dataset
  - Spam SMS Classifier Dataset
  - YouTube Dataset
  - Chars74K Dataset
  - Facial Image Dataset
  - … etc.

- Datasets associated with a standard:
  - Face Image ISO Compliance Verification

What is an Ideal Data Set?

- Free
- Data explained well (system description, mode changes, sensors, etc.)
- Labelled accurately
- Training, Testing, and Validation data sets are static (for comparing results), covers behaviors of interest, and has sufficient variation (unbiased)
- Limited pre-preprocessing
- Reusable for different applications
- Not too large (or streaming)
- Associated with a standard

How do we use Public Datasets?

- Learning / Test new ideas
- Compare approaches
- Debugging implementations
- Transfer Learning
- Test approach on multiple data sets (solutions not biased toward a single data set)
- etc.

Creating a Public Dataset

- Define Problem (e.g. classification, regression, performance measures, narrow scope)
- Get the data and set up for ease of access
- If required, label the data (labor intensive)
- Document the data (e.g. descriptions of data columns, system, sensors, etc.)
- Test out your own solutions to find issues

Public Data Sets Yield Faster Results

- Example Use Case: Testing out new ML ideas

- Prerequisites:
  - Understand the data and see if it is appropriate
  - Get access to data
  - Clean data

- If Public Data Set Available and understood well from multiple use:
  - ~~Understand the data and see if it is appropriate~~
  - ~~Get access to data~~
  - Clean data (?)

Fast answers to questions such as:

- Does your approach make sense?
- Is your approach implemented correctly?
- etc.

Main Takaways

- Can find public data sets through: Google Dataset Search and many others
- Ideal Data Sets: sufficient variation, coverage, reusable, static, standards
- Primary Uses: Learning and Testing and Comparing Approaches
- Benefits: After familiarizing with some public data sets, those can be reused quickly