Importance of Data

On today’s agenda:

1. Acquiring data
    - Public data (easy access)
    - Public data (difficult)
    - Collect your own
2. Preparing data
    - Data exploration
    - Data cleaning
    - Data preprocessing
3. Tutorial

Introduction

The success of Deep Learning is tied to the amount and quality of the training data. The noisier the data the harder it is to extract the important features and the easier it is to overfit to the noise.

Garbage in garbage out!

Acquiring data

Thankfully there is a lot of structured and unstructured data freely available on the internet.

- Easily accessible public data
- ‘Blobs’ of data that you have to sift through
- You can always collect your own data!

Acquiring data: Easy access

- Government organizations
- Non-profit organizations
- Social media

Acquiring data: Easy access

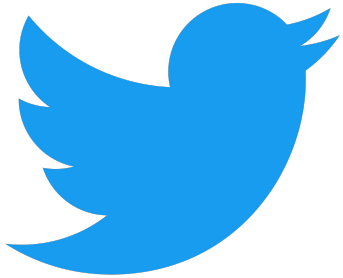
- Government organizations
  - Data.gov
  - Census data
  - Energy Information Administration
  - ...

Acquiring data: Easy access

- Non-profit organizations
  - Wikipedia
    - Millions of articles in hundreds of languages
    - Highly structured and categorized
  - OpenStreetMap
  - ...

Acquiring data: Easy access

- Social media
  - Twitter
    - Good source for text data
    - Popular for sentiment analysis
  - Flickr
    - High quality image data
    - Convenient licencing



flickr

Acquiring data: Crawling

Sometimes there is no API to easily query large quantities of data. The data might not even be structured.

In such cases you have to put in some effort and write code to access and parse the data.

- wget and bash for loops
- Use tools like Scrapy

Acquiring data: Collecting your own

Companies usually have access to proprietary data and have their own collection methods. If, however, you need to collect new data there are a few things to pay attention to:

- Pay attention to similar data you have access to*
- Use multiple sources, you can even use multiple recording devices
- Pay heed to the balance of the data

* You might even be able to utilize transfer learning!

Preparing data

There are three steps to take after acquiring a dataset

- Data exploration
- Data cleaning
- Preprocessing

Preparing data: Exploration

- What are the recorded features?
- What range of values do they take?
- Is the format consistent across all samples?
- Are there redundancies or obvious correlations?
- Balance and completeness

Preparing data: Cleaning

- Missing values and duplicates
- Irrelevant features
- Irrelevant samples
- Outliers

Preparing data: Cleaning

- Missing values and duplicates

Samples with missing values can still carry relevant information. Some feature analysis can guide you when deciding whether to discard a sample/feature or to incorporate it into the data.

  - You can classify missing data as its own class
  - You can use interpolation to fill in the missing value

Preparing data: Cleaning

- Irrelevant features

Some features might not be relevant to your ML problem. Discarding such features would allow you to build simpler models. This might however hinder future expansion of the model.

Preparing data: Cleaning

- Irrelevant samples

Some samples may be rendered invalid as they do meet your criteria such as assumptions made about the data.

Example: When analysing user interactions with a website, bots will inevitably make your life difficult!

Preparing data: Cleaning

- Outliers

Some samples, although valid, are not representative of the real world data and only serve to skew the data rendering it unbalanced.

Example: In a balanced dataset of 10,000 samples you cannot have multiple 1 in a million events!

Preparing data: Preprocessing

- Feature engineering
- Vectorization
- Normalization and batch-normalization
- Data augmentation

Preparing data: Preprocessing of Text

- Tokenization
  - By character
  - By word (with or without stemming)

- Vectorization
  - One-hot encoding
  - Binning

Tutorial

Practice makes perfect!