How to go about a machine learning project

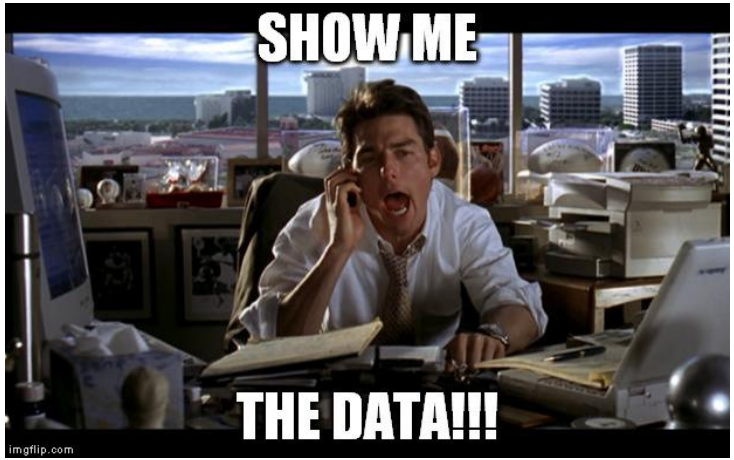
2018 GEN 511 / Machine Learning

Defining the problem

- The first step in any project is defining your problem. You can use the most powerful and shiniest algorithms available, but the results will be meaningless if you are solving the wrong problem.
- Steps to define a problem
  - **Step 1**: What is the problem?
  - **Step 2**: Why does the problem need to be solved?
  - **Step 3**: How would I solve the problem? (Domain Knowledge)

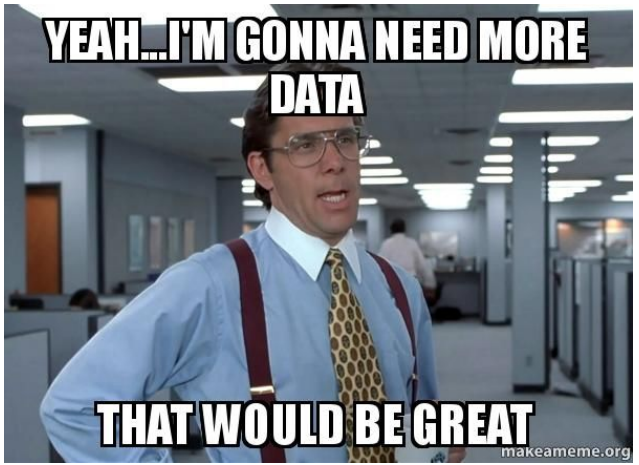
Finding appropriate dataset

- Hardest part of a machine learning project
- Common source
  - Kaggle
  - UCI
  - Data.gov
- Other option
  - Collect your own data
  - Scrape the web



What to do if the data is not sufficient?

- Plentiful high-quality data is the key to great machine learning models. But good data doesn’t grow on trees, and that scarcity can impede the development of a model
- **Data augmentation**
  - Rotating images
  - Adding noise to data
- By augmenting your dataset, you can get excellent results with small data.



Preparing Data

Data might not always be in the form that you can directly feed to your learning algorithm

- Formatting data
- Data annotation
- Cleaning - Removing missing values
- Normalization

Selecting the write hypothesis class

- Visualize the data
- Will depend on the problem statement
- Also depend on the data you have
- Will require domain knowledge

Improving the results of your model

- Get more data
- Feature engineering
- Feature selection
- Ensembling models

Presenting the result

- Mean accuracy score
- Confusion matrix
- True positive rate, True negative rate
- Area under ROC

DEMO

Our Problem Statement

- Predict the effectiveness of malaria vaccine using gene expression data.
- What - evident from the problem statement itself.
- Why - will help in reducing the trial phase of a vaccine.
- How - Read, Read and then read some more

Data source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18323

Data preparation

Time to move to the terminal