Homework 1 Tutorial

Artificial Intelligence II - Fall 21-22
Dept. of Informatics & Telecommunications
National & Kapodistrian University of Athens

George Katsogiannis (katso@athenarc.gr)

Calculating the Gradient

$$MSE(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}\left(h_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2$$

Keep in mind:

- The Chain Rule
- What are the variables in the equation? Matrices? Vectors?
- How can I rewrite a summation of vectors as a matrix multiplication?

Jupyter Notebooks

- An interactive computational environment for creating notebooks that contain code, text, multimedia, etc.
- Very helpful for data science projects, experimenting and visualising data and results
- Google Colab
  - Allows us to run our notebooks online
  - Offers access to GPUs, TPUs

Structuring a NLP Project

1. Loading and exploring the dataset
2. Data pre-processing
3. Training a model
4. Making predictions
5. Evaluating the model/predictions
6. Experimenting with different choices
7. Presenting the results

Loading and Exploring the Dataset

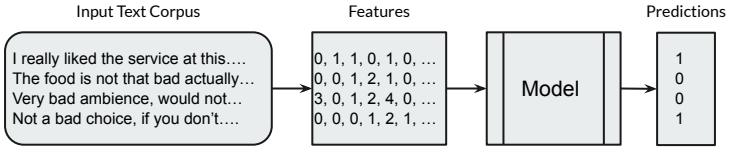
- Load the data and examine its structure
- Visualise some examples to gain insights
- Split dataset into train/validation/test

Useful library: ***pandas***

- *"pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool"*
- Functions to load multiple types of files (csv, tsv, etc.)
- Helps display and get quick information on large amounts of data
- Easy to handle and manipulate tabular data

Data Pre-processing

- Data cleaning
  - Is some part of the data unnecessary or can confuse the model?
  - What is the best way to handle it?
- Create features for the model
  - ML/DL models work with numbers
  - How can we represent texts with vectors?

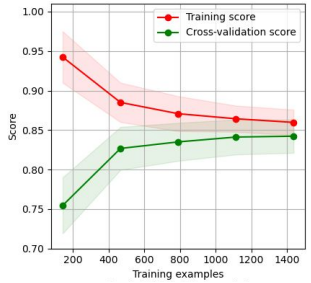


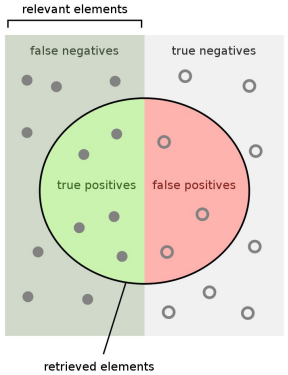
- NLTK *(not necessary)*
  - Tokenization
  - Lemmatization
  - Stemming
- Scikit Learn
  - Feature Extraction
  - Count Vectorizer, Hashing Vectorizer, TF-IDF Vectorizer
- Word Embeddings *(next HWs)*
  - GloVe, Word2Vec, FastText, etc.

Evaluating the model/predictions

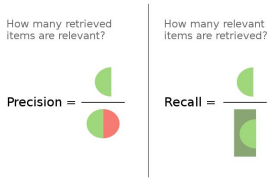
- Use a performance measure
  - How good is my model?
  - How does it compare to other models?
- Examine some correct/wrong predictions
  - What is my model good/bad at?
  - What could I do to improve these mistakes?
- Plot learning curves
  - Is my model overfitting/underfitting?
  - How fast does it learn?

- Scikit Learn
  - Precision
  - Recall
  - F1 Score
  - All-in-one metrics
  - Confusion Matrix
  - Learning Curves





	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP



Experimenting and Presenting Results

- No model/pre-processing/representation that is perfect for every task
- Our experience can provide intuitions on what works better for certain problems
- But the ultimate goal is to give an interesting overview of the problem and an insightful comparison of different approaches

Nice things to include in your report:

- Some examples from the dataset and any observations you might make
- An explanation of each step/approach used in your project
- An evaluation of your final model
- Performance comparisons with other models you tried