Tuesday, October 26, 2021

First Session 10.00 - 13.00

---

0 Introduction to Unstructured Data Analysis

0.1 Why?
0.1.1 The importance of Unstructured Data
0.2 What?
0.2.1 Information Extraction
0.2.2 Natural Language Processing
0.2.3 Machine Learning
0.3 How?
0.3.1 Learning Tools
0.3.2 Google Colab
0.3.2.1 How Upload a File on Google Colab
0.3.2.2 How Download a File on Google Colab
0.4 End Notes

1 Introduction to Machine Learning

1.1 What is Machine Learning?
1.1.1 Type of Machine Learning Models
1.1.2 Jargon
1.1.3 Type of Data
1.2 Feature Normalization
1.2.1 The maximum absolute scaling
1.2.2 The min-max feature scaling
1.2.3 Z-Score
1.3 Cost Functions
1.3.1 Linear Cost Function
1.4 Gradient Descent
1.5 Validation and Testing
1.6 Bias and Variance
1.7 Regularization
1.7.1 Ridge Regression
1.7.2 Lasso Regression
1.7.3 Elastic Net Regression
1.8 References

2 Supervised and Unsupervised Models

2.1 What is Supervised Learning
2.1.1 Linear Regression
2.1.2 Example 1 - Predicting Iowa House Prices (from Kaggle)
2.1.2.1 The Problem
2.1.2.2 Categorical Features
2.1.2.3 Loading data (J. C. Hull, 2019, Chapter 3)
2.1.2.4 Linear Regression with sklearn
2.2 What is Unsupervised Learning
2.2.1 k-Means Clustering
2.2.1.1 How the K-means algorithm works
2.2.1.2 A Distance Measure
2.2.1.3 K-means algorithm example problem
2.2.2 Example 2 - Country Risk
2.2.2.1 The Country Risk Dataset (J. C. Hull, 2019, Chapter 2)
2.2.3 Simple exploratory analysis

2.2.4 K means cluster
2.2.5 Perform elbow method
2.2.6 List the result
2.2.7 Silhouette Analysis
2.3 References

Second Session 14.00 - 17.00
---

3 Introduction to Deep Learning

3.1 What is a Neural Network
3.2 Mc Culloch and Pitts Artificial Neuron
3.3 Feedforward Neural Networks
3.3.1 One Hidden Layer NN
3.3.2 The Loss Function
3.3.3 Formula of Batch Training
3.3.4 Generate Sample Dataset
3.3.5 Weights Initialization
3.3.6 Forward Propagation
3.3.7 Loss Function
3.3.8 Delta Rule
3.3.8.1 Gradient Calculation
3.3.8.2 Weights Update
3.3.9 Batch Loader
3.4 Keras: the Python Deep Learning API
3.5 Other types of Neural Networks
3.6 Appendix - Calculation of Activation Functions Derivatives
3.6.1 Derivative of the Hyperbolic Tangent
3.6.2 Derivative of the Sigmoid
3.7 References and Credits

4 Basic Text Analysis

4.1 What is Text Mining and Why it's so Important
4.2 Package Required
4.3 What is a Corpus
4.4 Processing Raw Text
4.4.1 Processing HTML Files
4.4.2 Urllib
4.4.3 BeautifulSoup
4.5 Regular Expressions
4.5.1 Why we need Regular Expression
4.5.2 The Re Module
4.5.3 Python Regex Metacharacters
4.5.4 Further Examples
4.6 References and Credits

==========
Wednesday, October 27, 2021

First Session 10.00 - 13.00
---

5 Natural Language Processing: Tools and Ideas

5.1 Introduction
5.2 The NLTK package
5.2.1 Using NLTK Corpus
5.2.2 Loading Your Own Corpus
5.3 Text Preprocessing
5.3.1 Tokenization
5.4 Lemmatisation and Stemming
5.4.1 NLTK Stemming
5.4.2 NLTK Lemmatization
5.4.3 Stemming or lemmatization?
5.5 Part-of-Speech Tagging
5.5.1 What is POS Tagging?
5.5.2 Example 1 - Reading the Newspaper
5.5.3 Where we can use POST
5.5.4 Build your own POS Tagger
5.5.4.1 Exploiting Context
5.6 spaCy
5.6.1 What is spaCy?
5.7 References and Credits

6 Text Vectorization

6.1 What are Word Vectorization and Why It is so Important
6.2 Bag-of-Words (BOW)
6.2.1 Sample corpus of text documents
6.2.2 Simple text pre-processing
6.2.3 Frequency Vectors
6.2.4 One-Hot Encoding
6.2.5 Term Frequency-Inverse Document Frequency
6.2.6 Extracting Features for New Documents
6.3 Document Similarity
6.4 Word Embedding
6.5 Word2Vec
6.5.1 CBOW Architecture
6.5.2 Skip-gram model
6.6 Implementing a word2vec model using a CBOW NN architecture
6.6.1 Load up sample corpus
6.6.2 Build Vocabulary
6.6.3 Build (context_words, target_word) pair generator
6.6.4 Build CBOW Deep Network Model
6.6.5 Train model for 5 epochs
6.6.6 Get word embeddings
6.6.7 Build a distance matrix to view the most similar words
(contextually)
6.7 GloVe
6.7.0.1 Co-Occurence Matrix
6.8 Pre-Trained Embedding Models
7 References and Credits

Second Session 14.00 - 17.00

---

7 Classification for Text Analysis