Introduction to Machine Learning Applications

Spring 2021

Lecture-15

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Agenda for today

- Unsupervised Learning example and Python exercises
- Text and NLP
- Class exercises

Announcements

- Homework-6 due tonight 11:59 pm ET
- Session-2 today will be focused on hw6

Natural Language Processing

- Language is complex and challenging

"The Scarecrow and the Tin Woodman stood up in a corner and kept quiet all night, although of course they could not sleep." - The Wonderful Wizard of Oz, L. Frank Baum

Language is challenging

- I looked straight at an eclipse of the sun.
- I looked after my little sister.
- I looked at the catalogue and saw the dress that I wanted to order.

Language is challenging

- River bank/ financial bank
- Mountain bike/ tandem bike
- Wine glass/ sea glass
- …

Subfields of Linguistics

Subfield	Description
Phonetics	The study of the sounds of human language
Phonology	The study of sound systems in human language
Morphology	The study of formation and internal structure of words
Syntax	The study of formation and internal structure of sentences
Semantics	The study of the meaning of sentences
Pragmatics	The study of the way sentences with their semantic meanings are used for particular communicative goals.

Table from: Linguistic Fundamentals for Natural Language Processing, June 2013, Emily M. Bender

Different Components

- Language models
- Part Of Speech tagging
- Parsing
- Named Entity Recognition
- Co-reference
- Automatic Speech Recognition
- Text To Speech
- Semantic roles

Important questions to ask

- What level of **representation**?
  - Words, sequences, trees
- What are the assumptions of a **model**?
  - Grammar, Hidden Markov model
- Efficient algorithms for **learning**?
  - Viterbi, Forward-Backward

Applications

- Text classification
  - Spam detection
- Question answering systems
  - Watson
- Machine translation
  - English to Korean
- Spoken dialogue systems
  - Siri, Alexa

In this class..

- Sentiment analysis
- Topic modeling
- Bag of words approach
- Word embeddings

Types of text

- Subjective text:
  - Blogs, Online product reviews, Movie reviews, Social media posts such as Tweets, etc.
- Objective text:
  - Wikipedia articles, NYT news articles, WSJ articles, etc.

Sentiment Analysis

- Computational inference of opinion, sentiment subjectivity in text

- Businesses are interested in customer opinions
- Individuals looking at the existing product reviews before purchasing a product
- Tweets during a presidential debate to predict who is the most favorable candidate

Sentiment Analysis

- There are practical issues
  - Cleaning the text
  - Tokenization
  - POS tagging
  - Semantics
  - Named entity recognition
  - Coreference resolution
  - …

Sentiment Analysis

Positive	Negative
My experience so far has been fantastic!	Your support team is useless
Package is nice.	Absolutely horrible!
This product is second to none.	This is better than nothing.
I will absolutely recommend it.	Yeah sure. So smooth!

Sentiment Analysis

- Unsupervised classification
  - Using POS tagging, bigrams and calculating Semantic Orientation (SO)
- Supervised classification
  - Naïve Bayes
  - Maximum entropy
  - Support vector machines
- Large area of research – Issues with fake text (such as fake reviews), spam filters,..

Exercise – 1 – Fake?

1. I was kinda doubtful about the "electroluminescent technology" of this USB cable. That it actually would work. But it actually did. Not only did the blue light function as they should, they were clear and bright, plus they turned off when the charge was finished. It's rad. They should make them in other colors too. Let's just say we're really impressed and are going to order a few more...
2. Such a cool product. I was so happy with how bright the lights on the cable are. It shipped super fast. The light shuts off when the charging is complete, so that’s super helpful. I don’t have to keep checking.

Topic Modeling

- Topic model is a type of statistical model to discover the abstract latent topics present in a given set of documents.
- Topic modeling allows us to discover the latent semantic structures in a text corpus through learning probabilistic distributions over words present in the document.
- It is a generative statistical model that allows different classes of observations to be explained by groups of unobserved data similar to clustering.
- **It assumes that documents are probability distributions over topics and topics are probability distributions over words.**

Latent Dirichlet Allocation

- Latent Dirichlet Allocation (LDA) was proposed by Blei et al. in 2003
- LDA assumes that the document is a mixture of topics where each topic is a mixture of words assigned to a topic where the topic distribution is assumed to have a dirichlet prior.

Exercise – 2

Doc1: The chicken curry is delicious.

Doc2: We learned that these chickens are raised cage-free.

Doc3: The delicious food was prepared by a famous chef.

Doc4: They add a lot of spices to prepare the food.

What are the topics?

TF-IDF

- tf-idf stands for Term frequency-inverse document frequency. The tf-idf weight is a weight often used in information retrieval and text mining. Variations of the tf-idf weighting scheme are often used by search engines in scoring and ranking a document's relevance given a query.
- This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus (data-set).

TF-IDF Computation

- **STEP-1: Normalized Term Frequency (tf)** -- tf(t, d) = N(t, d) / ||D|| wherein, ||D|| = Total number of term in the document
  - tf(t, d) = term frequency for a term t in document d.
  - N(t, d) = number of times a term t occurs in document d
- **STEP-2: Inverse Document Frequency (idf)** -- idf(t) = N/ df(t) = N/N(t)
  - idf(t) = log(N/ df(t))
  - idf(pizza) = log(Total Number Of Documents / Number Of Documents with term pizza in it)
- **STEP-3: tf-idf Scoring**
  - tf-idf(t, d) = tf(t, d)* idf(t, d)

TF-IDF Example

- Consider a document containing 100 words where:
  - the word “kitty” appears 3 times. The term frequency (i.e., tf) for kitty is then (3 / 100) = 0.03.
- Now, assume we have 10 million documents and the word kitty appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4.
- Thus, the Tf-idf weight for the word “kitty” is the product of these quantities: 0.03 * 4 = 0.12.

Exercise –3

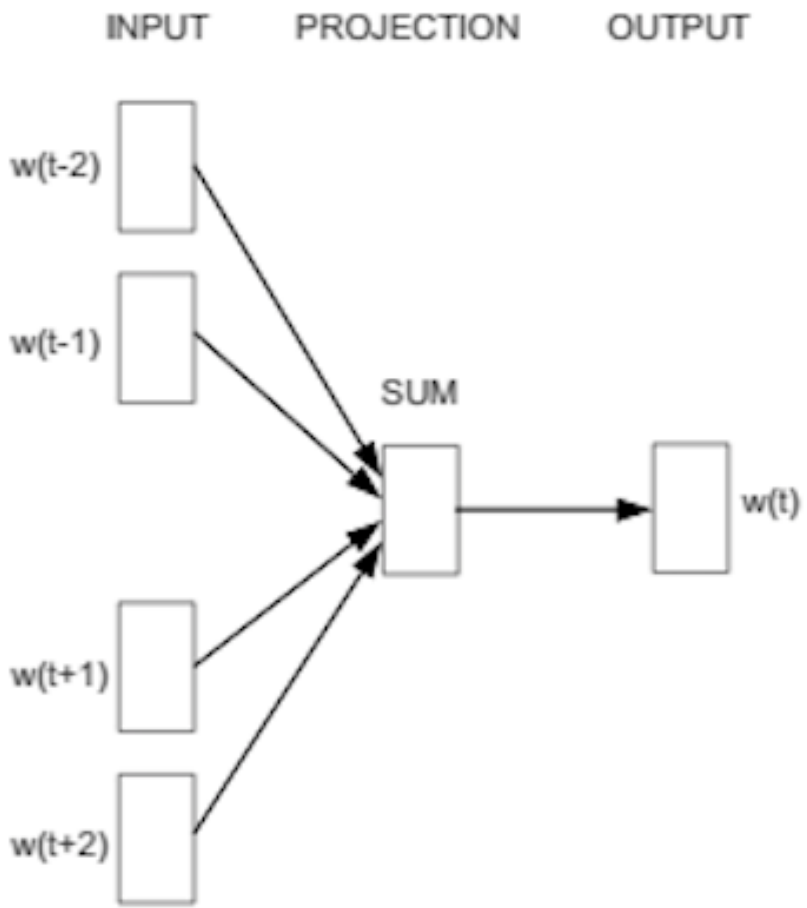
- Doc1: I love delicious pizza
- Doc2: Pizza is delicious
- Doc3: Cats love me

Compute the TF-IDF values for -- “love”, “Cats”, “delicious”

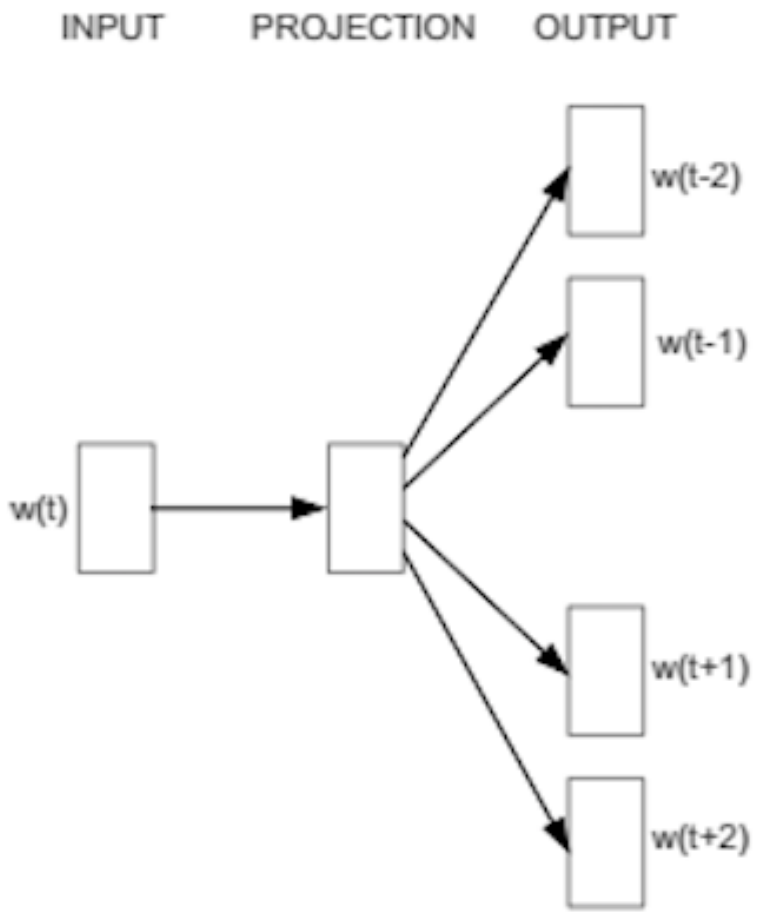
Word Embeddings

- These are learned word representations that allow words with similar meanings to have a similar representation.
- Each word in a given corpus is represented in terms of a vector in a predefined vector space.
- These vectors have often hundreds of dimensions, depending on the size of the corpus.
- These representations are learned based on how the words are used in the training corpus.
- Word2Vec is a statistical method or model that was built as a 2-layer neural network.

Word2Vec



CBOW



Skip-gram

Final Exercise

- We will work on the Twitter dataset. We will specifically do these tasks:

1. Preprocessing
2. N-grams
3. Topics
4. Sentiments