Midterm Review

Topics to Go Over

- TF-IDF
- Logistic Regression
- Feature Engineering
- Word2Vec
- Recurrent Neural Network

TF-IDF

- Helps us with **ranked retrieval**
  - User’s query + document corpus and compute score for every document compared to query, and how relevant they are
- General idea
  - Vectors that encode both the query and the document
  - Take similarity of vectors as a proxy for relevance!

TF-IDF

- If a word appears a lot in the document, it's probably **relevant** to that document (i.e if I have a document discussing pasta, and I see the word pasta 50 times, it's definitely relevant!)
- Not all words are equally useful (*the*, *of*, *a)*
- **TF: Term Frequency**
  - How often does an individual word appear in the document?
- **IDF: Inverse Document Frequency**
  - How many documents does a word appear in?
- If a word appears a lot in a given document, it's probably important.
  - BUT if a word appears in many documents, probably not as important

TF-IDF

$$\mathbf{w_{i,\,j} = f_{i,j} \log (D / d_i)}$$

- Weight of word i in document j
- $f_{i,\,j}$ = frequency of word i in document j
  - Divide number of times word appears in a document by the total number of words in the document
- D = total number of documents in collection
- d_i = number of times word appears in any document in corpus
- Vector representation of both search queries and documents

TF-IDF Example

Doc 1	I love cats. Cats are cute.
Doc 2	I love animals, animals are loyal.
Doc 3	I love birds and cats.

TF-IDF Example

Doc 1	I love cats. Cats are cute.
Doc 2	I love animals, animals are loyal.
Doc 3	I love birds and cats.

Term Frequency (Cats) in document 1:

$f(cat, doc1) = 2$

TF-IDF Example

Doc 1	I love cats. Cats are cute.
Doc 2	I love animals, animals are loyal.
Doc 3	I love birds and cats.

Term Frequency (Cats) in document 1:

f(cat, doc1) = 2

Inverse Document Frequency:
N = 3
df(cats) = 2

IDF = log(3/2)

TF-IDF Examples (2)

Doc 1: He loves to watch basketball and baseball but prefers basketball

Doc 2: Janet likes to play basketball

Doc 3: Julia loves to play baseball, and wishes she could play more often

TF-IDF Examples (2)

Doc 1: He loves to watch basketball and baseball but prefers basketball

Doc 2: Janet likes to play basketball

Doc 3: Julia loves to play baseball, and wishes she could play more often

1. Tf-idf of “basketball” in Doc 1 = ?
2. Tf-idf of “play” in Doc 2 = ?
3. Tf-idf of “she” in Doc 3 = ?
4. Tf-idf of “baseball” in Doc 3 = ?

TF-IDF Examples (2)

Doc 1: He loves to watch basketball and baseball but prefers basketball

Doc 2: Janet likes to play basketball

Doc 3: Julia loves to play baseball, and wishes she could play more often

1. Tf-idf of “basketball” in Doc 1 = (2/10) * log (3 / 2)
2. Tf-idf of “play” in Doc 2 = (⅕) * log (3 / 2)
3. Tf-idf of “she” in Doc 3 = (1/12) * log (3 / 1)
4. Tf-idf of “baseball” in Doc 3 = 0 (if you didn’t use nltk.word_tokenize() and just did a.split()!)

Logistic Regression

- Algorithm
- Simple workout examples
- Softmax function
- Back propagation and Gradient Descent

Logistic Regression

- Logistic Regression is an example of classification (instead of predicting a real number, i.e house price, age of child, etc), we'll predict **probabilities of a set of outcomes**

Logistic Regression

- Logistic Regression is an example of classification (instead of predicting a real number, i.e house price, age of child, etc), we'll predict **probabilities of a set of outcomes**
- Weight vector: β_i

Logistic Regression

- Logistic Regression is an example of classification (instead of predicting a real number, i.e house price, age of child, etc), we'll predict **probabilities of a set of outcomes**
- Weight vector: β_i
- Examples: X_i
- Bias term: β_0
- exp(x) → e^x
- Logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$ squashes numbers into [0,1]

Logistic Regression

- Logistic Regression is an example of classification (instead of predicting a real number, i.e house price, age of child, etc), we'll predict **probabilities of a set of outcomes**
- Weight vector: β_i
- Examples: X_i
- Bias term: β_0
- exp(x) → e^x
- Logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$ squashes numbers into [0,1]

Softmax

$$P(Y=0|X) = \frac{1}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}$$

$$P(Y=1|X) = \frac{\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}$$

Logistic Regression in Vector Form

- Logistic Regression is an example of classification (instead of predicting a real number, i.e house price, age of child, etc), we'll predict **probabilities of a set of outcomes**
- Weight vector: β_i
- Examples: X_i
- Bias term: β_0
- exp(x) → e^x
- Logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$ squashes numbers into [0,1]

Softmax

$$P(Y=0|X) = \frac{1}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}$$

$$P(Y=1|X) = \frac{\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}$$

Logistic Regression

Imagine we have feature vector $x_i = [1, 2, 2]$ and corresponding actual label $y_i = 1$ for the i^{th} example in our training set.

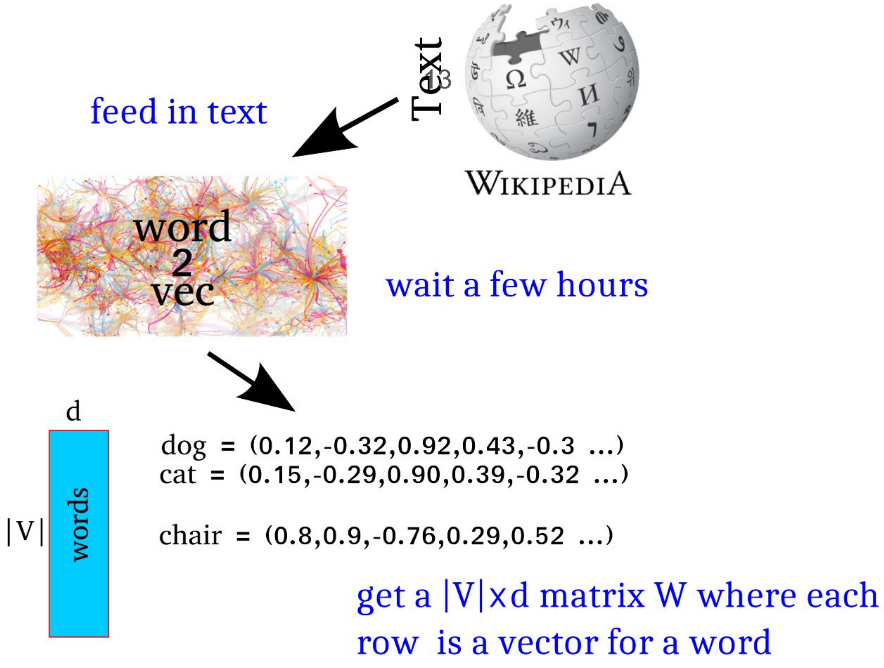
Suppose we have our current parameter vector be $\beta = [-1, 2, -1]$.

Q1. Which class will the logistic regression classifier predict at this stage?

Class Example: https://users.umiacs.umd.edu/~ying/teaching/CMSC_470/lr_ex.pdf

Word2Vec

- Represent words with their meaning (semantics)



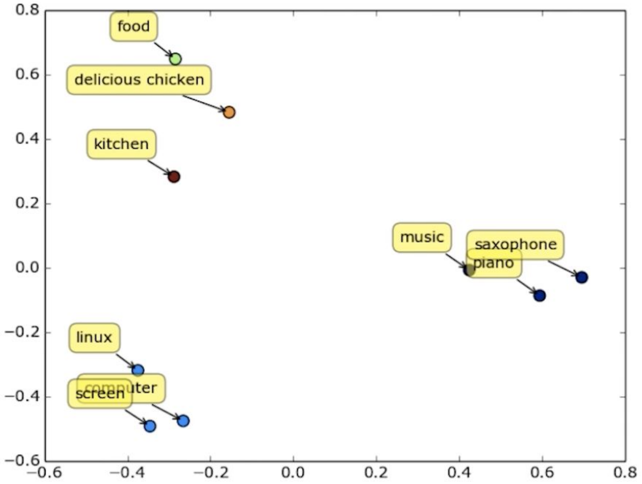
https://users.umiacs.umd.edu/~jbg/teaching/CMSC_470/06b_word2vec.pdf

Word2Vec

- Distributional hypothesis: Learn something about a meaning of a word based on the other words it appears with
- Encode words with similar context to be close in some vector space

How to measure similarity?

cosine similarity!

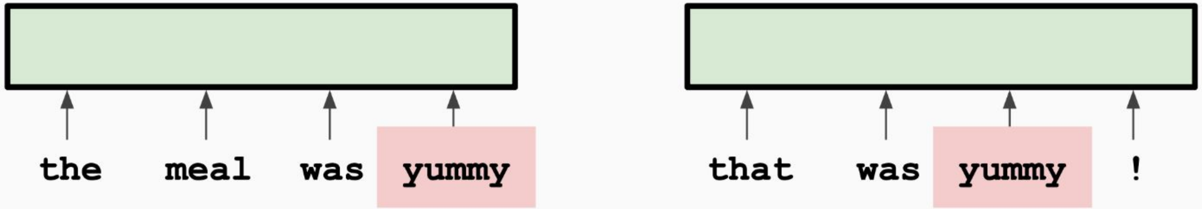


https://users.umiacs.umd.edu/~jbg/teaching/CMSC_470/06b_word2vec.pdf

RNN

- Network Architecture

- Why RNNs? Why not feedforward neural networks (or FFNNs)?
  - Variable length input – sequences are naturally varying in length
  - With FFNNs, each position in the input embedding has some fixed semantics



  - Ideally, we can process these tokens in a uniform manner
  - Exploit context!

Adapted from Greg

RNN Computation

- For each time step computation, the hidden unit computation is:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1})$$

$$y_t = W_{hy}h_t$$

- f is activation function. (tanh)

RNN Computation (Example)

- For each time step computation, the hidden unit computation is:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1})$$

$$y_t = \mathrm{softmax}(W_{hy}h_t)$$

x1	x2	x3	
I	like	eating	___

$$x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$W_{xh} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}, \quad W_{hh} = \begin{bmatrix} 0.4 & 0.1 \\ 0.2 & 0.5 \end{bmatrix}, \quad h_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad W_{hy} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 0.5 & 0.5 \end{bmatrix} \begin{array}{l} \text{donut} \\ \text{fish} \\ \text{burger} \end{array}$$

What is next word?

Concepts Need to know

TF-IDF / Information Retrieval

- What is TF? IDF?
- How are TF-IDF terms computed?
- How does a TF-IDF system work in practice?
- What does TF-IDF frequency and Rank plot look like?
- What are some of the drawbacks of TF-IDF systems?

Distributional Semantics

- What is distributional semantics?
- What is word2vec, how does it work?
- What are context vectors and Weight vectors? How are they computed?

Regression

- What is linear regression?
- Logistic regression?
  - What is the logistic function?
- How to interpret logistic regression weights?
- Evaluation: how to interpret confusion matrix for binary classification

Recurrent Neural Networks

- What is Embedding from Language Models? How is it used in RNN?
- How do you initialize weights for a neural network?

More Concepts Need to know

Byte Pair Encoding (BPE)

- How does it work?
- How does BPE differ from traditional word-level tokenization?
- How to handle new (unseen) tokens?
- Go through homework BPE implementation.

Dependency Parsing and Part-of-Speech

- What is the meaning of parsing objective?
- What is POS?
- What models are usually used to train POS?
- Evaluation: how to evaluate POS? What are some metrics?

Hidden Markov Models

- What is HMM used for?
- Describe how it works.

Adam

- Basic computation of Adam