Sentiment Analysis

Week 12 Day 02

DS 3000 – Foundations of Data Science

1

Reminders

HW 8

Tuesday, November 26

Class on Tuesday, December 3

Attendance required (graded)

FP4 & FP5

Monday, December 9

2

Outline

Feature Extraction from Text

Sentiment Analysis

More Algorithms

3

Weighting Words Using tf-idf

Term frequency–inverse document frequency, tf-idf

Reflects how important a word is to a document or corpus

The tf-idf score for word ***w*** in document ***d*** is computed by

$$\text{tfidf}(w,\ d) = \text{tf}\log\left(\frac{N+1}{N_w+1}\right)+1$$

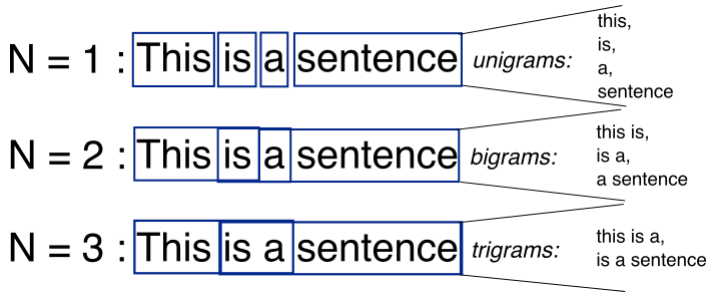
tf: the number of times the word ***w*** appears in the query document d
N: the number of documents in the corpus
$\mathbf{N_w}$: the number of documents in which the word ***w*** appears

4

N-Grams

N-gram:

A contiguous sequence of n tokens from a given piece of text



Provides more context

Addresses the problem with negations (“not good” vs. “not bad”)

5

N-Grams

For most text classification problems, unigrams are essential

Single words often capture a lot of meaning

Adding bigrams is helpful in most cases

Adds more context

Adding longer sequences, usually up to 5-grams, might be helpful too

Substantially increases the number of features

Risks overfitting

6

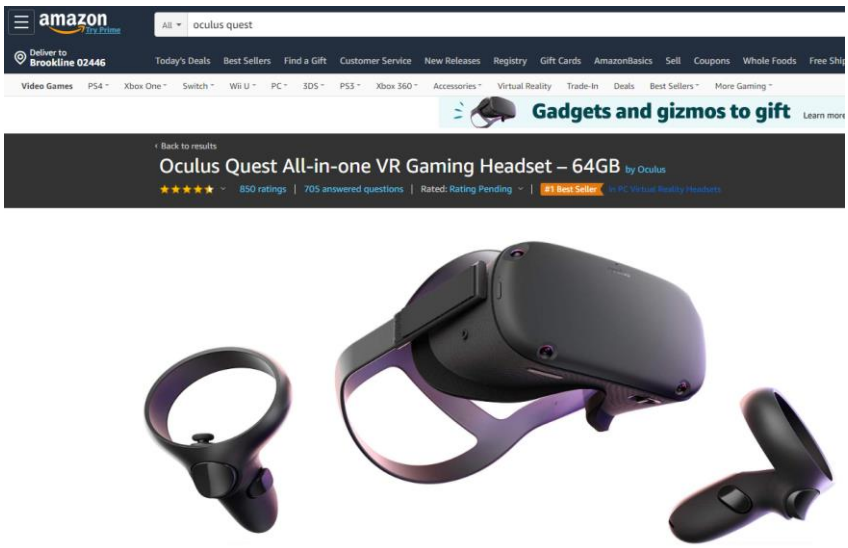
Sentiment Analysis

The use of natural language processing, text analysis, and computational linguistics to systematically identify and quantify affective states expressed in a piece of text

A common application of classification algorithms

7

Sentiment Analysis



8

Sentiment Analysis

Wow. I got more then I expected and I just can't stop playing. Beat saber, super hot, etc, etc, everything is thrilling, action packed, and amazing because you can move around, and it is almost like the system senses your body by what it can do in certain games. It's literally great because IM GETTING EXERCISE WHILE HAVING FUN! I

Customer rating: 

9

Sentiment Analysis

There seems to be a major problem with the Oculus Quest charging. I purchased 2 units. After 3 weeks, neither would charge anymore. Left to charge for 3 days it went from 3% to 11% with no use. Oculus does not have a customer support number. You have to send an email. Then, they don't get back to you. My 900 dollars worth of units are useless after 3 weeks. Google it. Everyone has the same issue. Nothing from Oculus. Very, very disappointing.

Customer rating: 

10

Sentiment Analysis

Can we predict whether a player will recommend a video game based on Steam reviews?

11

12

Logistic Regression

13

Logistic Regression

A common ML algorithm for binary classification tasks

- Despite its name

Similar to Ridge regression

- Linear model
- Applies L2 regularization (sklearn allows you to apply L1 too)

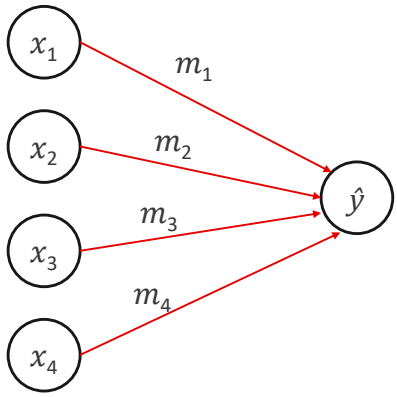
Regularization parameter: C

- Higher values of C increases the complexity of the model

14

Linear Regression

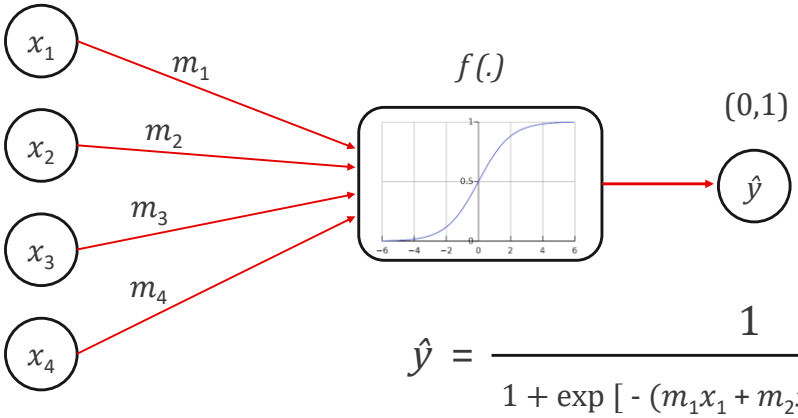
$$\hat{y} = m_1x_1 + m_2x_2 + \ldots + m_nx_n + b$$



15

Logistic Regression

$$\hat{y} = \text{logistic}(m_1x_1 + m_2x_2 + \ldots + m_nx_n + b)$$



$$\hat{y} = \frac{1}{1 + \exp\left[- (m_1x_1 + m_2x_2 + \ldots + m_nx_n + b)\right]}$$

16

Neural Networks

17

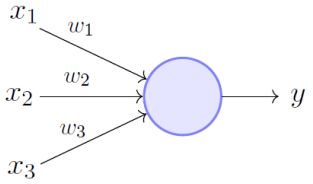
Neural Networks

Deep learning

A family of algorithms modeled loosely after the human brain, but not the same thing at all

Multilayer perceptrons (MLPs)

Simple, feed-forward neural networks

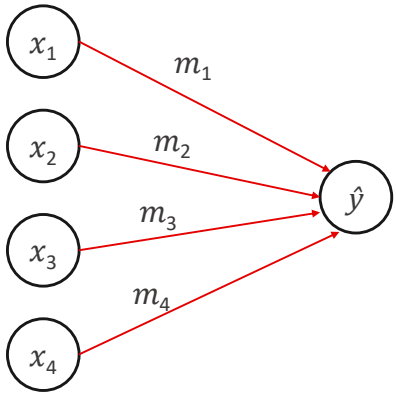


Perceptron Model (Minsky-Papert in 1969)

18

Neural Networks

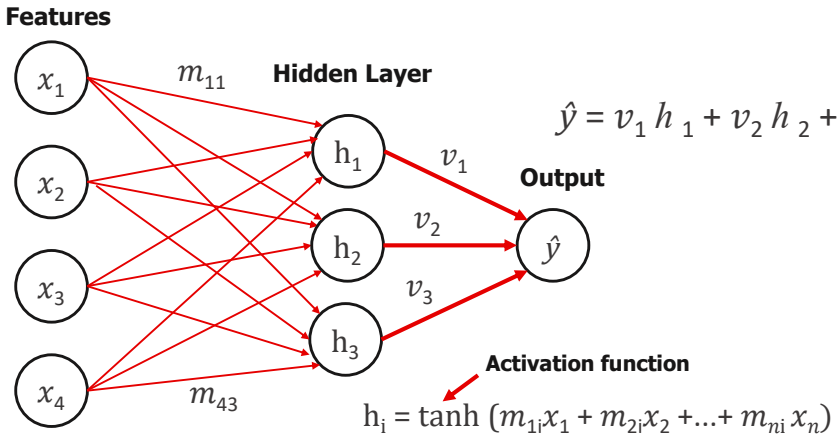
Linear models:



19

Neural Networks: Multilayer Perceptrons

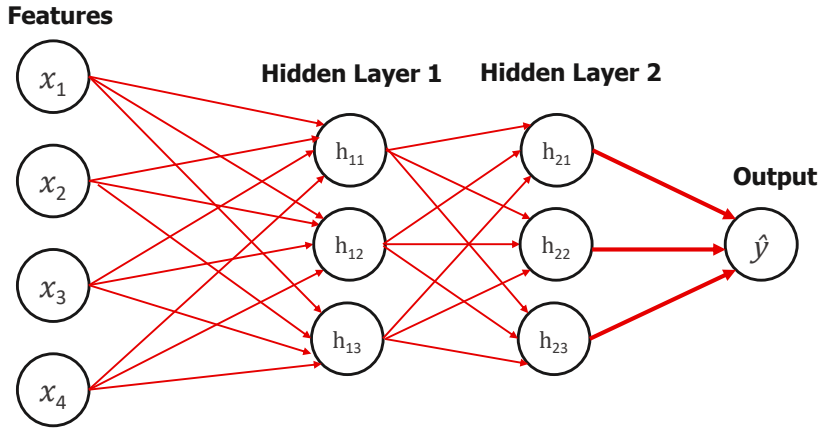
Generalized linear models that perform multiple stages of processing to come to a decision



20

Neural Networks: Multilayer Perceptrons

Deep learning is inspired by the idea of having large neural networks made up many hidden layers of computation



21

Neural Networks: Tuning

hidden_layer_sizes: sets the number of hidden layers and number of hidden units per layer (each list element).

Default: (100)

alpha: controls weight on the regularization penalty that shrinks weights to zero.

Default: alpha = 0.0001

activation: controls the nonlinear function used for the activation function, including: 'relu' (default), 'logistic', 'tanh'

22