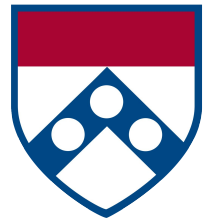
Wharton PhD Tech Camp

Session 8

Alex Miller

Ph.D. Student, Information Systems

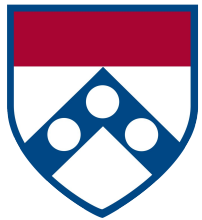
Wharton, OID



Goals for today

- Machine Learning Intro
- Where it fits into your research
- Brief walkthrough with word2vec

Machine Learning



What is it?

- Hard to say exactly…
  - Set of statistical techniques that have been developed over time to process data
  - Most ML techniques share the characteristic that some of the parameters/architecture of the model are “learned” from data
- No real universal though
  - Some forms of simple regression are considered ML

Causality v. Prediction

- Most interesting questions in business/economics are concerned with causality
  - Does x cause y?

$$y = X\beta + \varepsilon$$

- A lot of work in ML is *purely* concerned with prediction
  - Does x predict y? Correlations are very helpful!
- Machine learning is great if you need to be accurate
  - It is not very good at helping you test theories

Data Snooping in ML

- The t-tests/confidence intervals in traditional regressions if the set of variables has been determined beforehand
- Looking at the results of a model, fiddling with it, and re-running it can easily lead to p-hacking or “overfitting”
  - Overfitting means you found patterns in your exact dataset, but they don’t generalize to new datasets
- But model-fiddling is the bread and butter of machine learning!
  - Much of the problems in ML are about how to deal with overfitting -- “regularization”

“Supervised” Learning

- You are trying to learn the relationship between a set of X’s and a set of Y’s
  - Pictures on the internet
    - X = numeric pixel map; Y = “cat” or “not cat”
  - Movie reviews
    - X = raw text; Y = “favorable” or “unfavorable”
  - Monthly sales for some company
    - X = lots of monthly numbers; Y = numeric value of sales
- You are almost always trying to predict something
  - Doesn’t necessarily mean something in the future

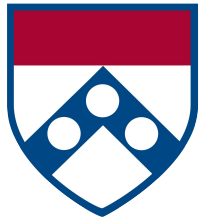
“Unsupervised” Learning

- You only have X’s, but want to cluster or transform them in some way
  - Cluster analysis
- Unsupervised output as input to supervised method
  - Principal Component Analysis

Important ML Methods

- Ridge regression
  - “Feature selection”
  - “Regularization”
- Support Vector Machine
- Random Forests
  - Almost always among the best, easiest methods for traditional X, Y data
- Neural Networks
  - Text data
  - Image data
- Semi/non-parametric Bayesian Methods

ML in Business Research



ML in Research

- It’s difficult to build an interesting paper 100% around machine learning
  - You may be able to use ML to solve very interesting problem, that could not be solved before
    - → Especially if there are interesting consequences
  - Even then, this is almost never sufficient for a truly interesting paper in business
    - → Lots of papers like this in CS

ML in Research

- ML *can* be useful for understanding your data or extracting covariates
  - More examples of this in business research
  - Still lots of low-hanging fruit
  - Especially with text, there is a lot of untapped data
- At the end of the day, your research question should be interesting even without ML
  - Depending on where your field is in the tech hype cycle, the cool factor may have already worn off

Examples of ML in IS/Marketing

Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics

Anindya Ghose, Panagiotis G. Ipeirotis, *Member, IEEE,*

Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook*

Dokyun Lee
Carnegie Mellon University

Kartik Hosanagar
The Wharton School

Harikesh S. Nair
Stanford GSB

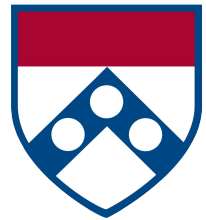
Machine Learning Workflow

- For feature extraction:
  - Obtain large dataset that is too large to label manually
  - Label a randomly drawn subsample
    - Could be X or Y variable
  - Build a model to predict the label you want
  - Apply your model to the entire dataset
    - It is important that your original data was randomly sampled!
    - We want an **honest** estimate of accuracy on this unlabeled data
  - Do a formal analysis using your predicted labels

Machine Learning Workflow

- Facebook Paper Example:
  - Scraped lots of Facebook posts/comments
  - Used AMT to apply labels to subsample
    - “Philanthropy”, “small talk”, “promotion”
  - Built ensemble ML model on subsample
    - Modern technique would be to use neural nets/word2vec
  - Predicted post labels for remainder of dataset
  - Used more traditional regression approach to analyze how various types of posts are affect comments/likes/shares

ML Methods for Text

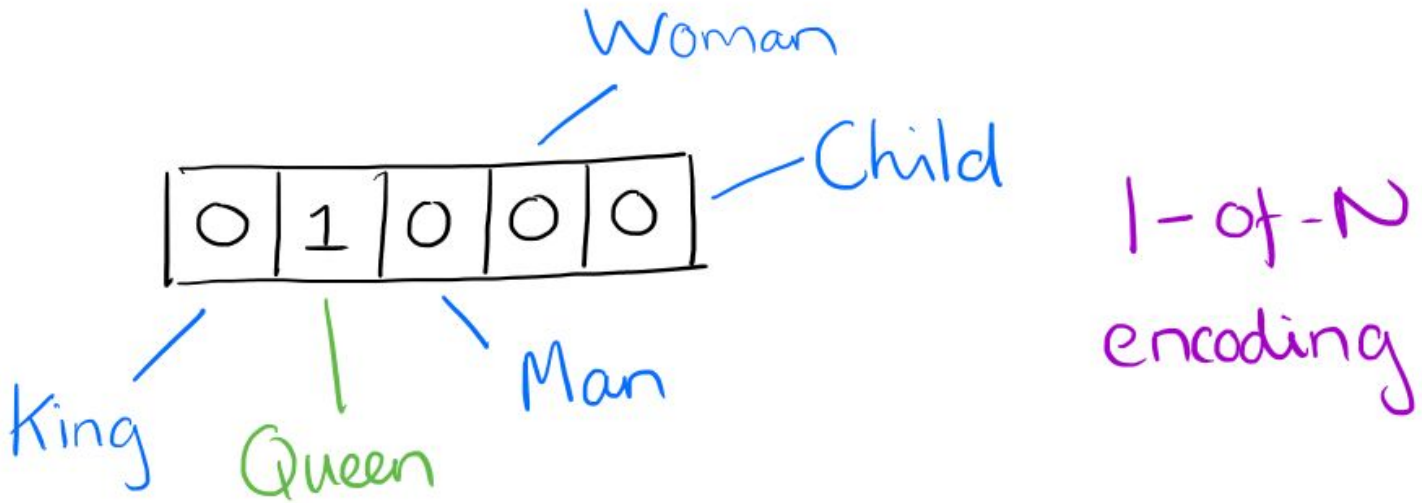


Intro to Word Embeddings

- As mentioned, text data is very high-dimensional
  - For large corpora, potentially 100,000+ of features
  - You need LOTS of data to get multiple examples for a large % of your features
    - → Leads to lots of noise
- What if we could “project” words into a low-dimensional “latent” vector space?
  - … what is latent word vector space?

Word Vectors: Traditional

1-hot encoding (“bag of words”) representation



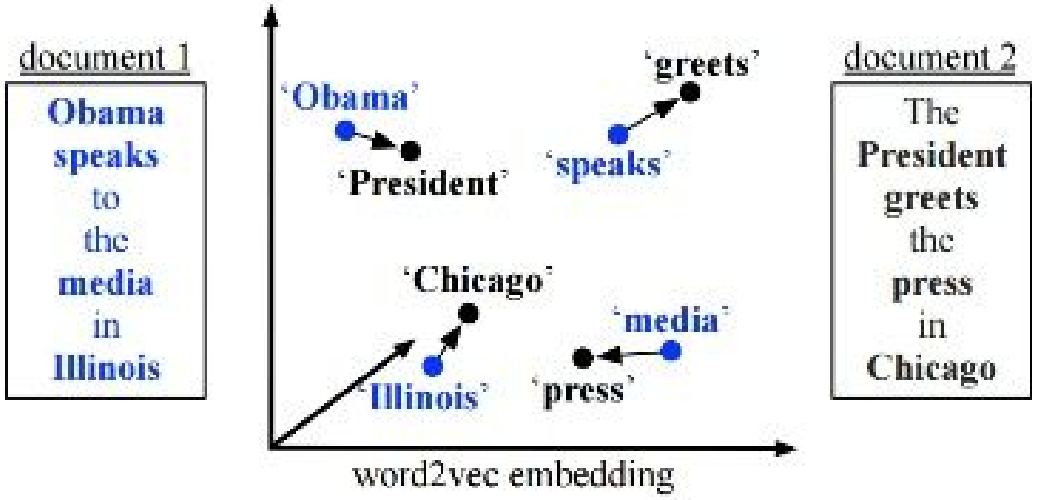
Many images from this great post: https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Word Vectors

Using “latent” attributes of “word space”

Why?

- Need fewer dimensions to represent essentially the same information
  - Traditionally “distant” words now appear close



Word2Vec

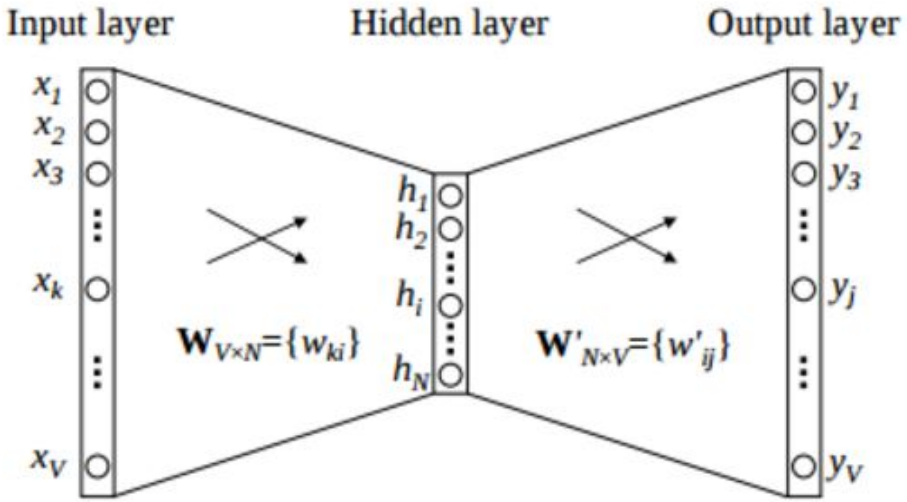
- Widely-used technique for “projecting” words into a low-dimensional vector space
- Very effective at grouping similar words, while retaining important differences/relationships between them
- Quite recent development
  - 2013 paper out of Google: https://arxiv.org/abs/1301.3781

Word2Vec: How does it work?

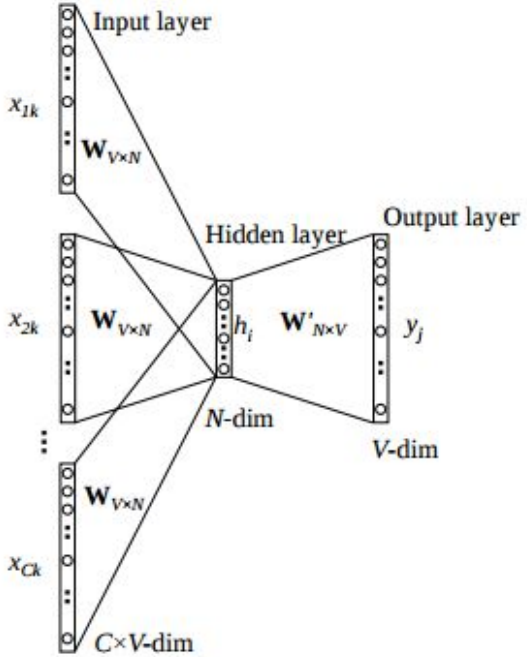
- The gist:
  - You are the company you keep
  - We can infer underlying/hidden/latent meanings from words from their context

Word2Vec: Behind the scenes

Continuous Bag of Words:
Context → Word



Skip-gram:
Word → Context



Quick Aside...

- Is any of this actually “NLP”?
  - If you ask a linguist, not really
  - But these models perform much better than “rules” or syntax-based algorithms that try to understand parts-of-speech and sentence structure

Using Word2Vec

- Train your own model (if you have enough data)
  - You can use ALL the text in your corpus (both labeled and unlabeled examples) when training a word2vec model
- Or use a prebuilt model
  - https://github.com/3Top/word2vec-api#where-to-get-a-pretrained-models
- Notebook Walkthrough
  - Will be posted after class

Resources

- Other useful links
  - Word2Vec Explainer; slighlty more technical
  - Word2Vec Python Example
- ML Courses
  - STAT974 Modern Regression
    - → Good for understanding the consequences of “data snooping”
  - CIS 519 Intro to Machine Learning
  - CIS 520 Machine Learning
- Your resident CS/ML expert