GENERAL ASSEMBLY

INTRO TO UNSUPERVISED LEARNING & TOPIC MODELING

Matt Brems

Data Science Immersive, GA DC

LEARNING OBJECTIVES

- By the end of this lesson, students should be able to:
  - Define supervised learning and unsupervised learning.
  - Identify strategies for unsupervised learning.
  - Describe topic modeling.
  - Implement and visualize Latent Dirichlet Allocation (LDA) in Python.

DATA SCIENCE PROCESS

- Step 0: Define your problem.
- Step 1: Gather data.
- Step 2: Clean your data and EDA.
- Step 3: Build your model.
- Step 4: Evaluate your model’s performance.
- Step 5: Answer your problem.

RECAP: SUPERVISED LEARNING

- Let’s list the modeling techniques we’ve learned about.

WHAT IS OUR Y VARIABLE IN THESE CASES?

- I want to predict who is likely to vote in the 2020 election.
- I want to group stores by the demographic profiles of their consumers.
- I want to organize tweets by their topic.

UNSUPERVISED LEARNING

- **Unsupervised learning** is where we have, as part of our training data, no observed Y values.

- **Supervised learning** is where we have observed Y values as part of our training data.

STRATEGIES IN UNSUPERVISED LEARNING

1. Pick a proxy Y, then do supervised learning.

2. Use unsupervised learning as a stepping stone to get to supervised learning.

3. Try to organize observations by features.

STRATEGIES IN UNSUPERVISED LEARNING

1. Pick a proxy Y, then do supervised learning.

STRATEGIES IN UNSUPERVISED LEARNING

2. Use unsupervised learning as a stepping stone to get to supervised learning.

STRATEGIES IN UNSUPERVISED LEARNING

3. Try to organize observations by features.

EVALUATING UNSUPERVISED LEARNING MODELS

- In regression, we often use MSE and/or R^2 to evaluate models.
- In classification, we often use accuracy, sensitivity, specificity, precision, and/or AUC ROC to evaluate models.
- All of these rely on knowing our actual Y and predicted Y.
  - …so how do we evaluate models for unsupervised learning?

UNSUPERVISED LEARNING TAKEAWAYS

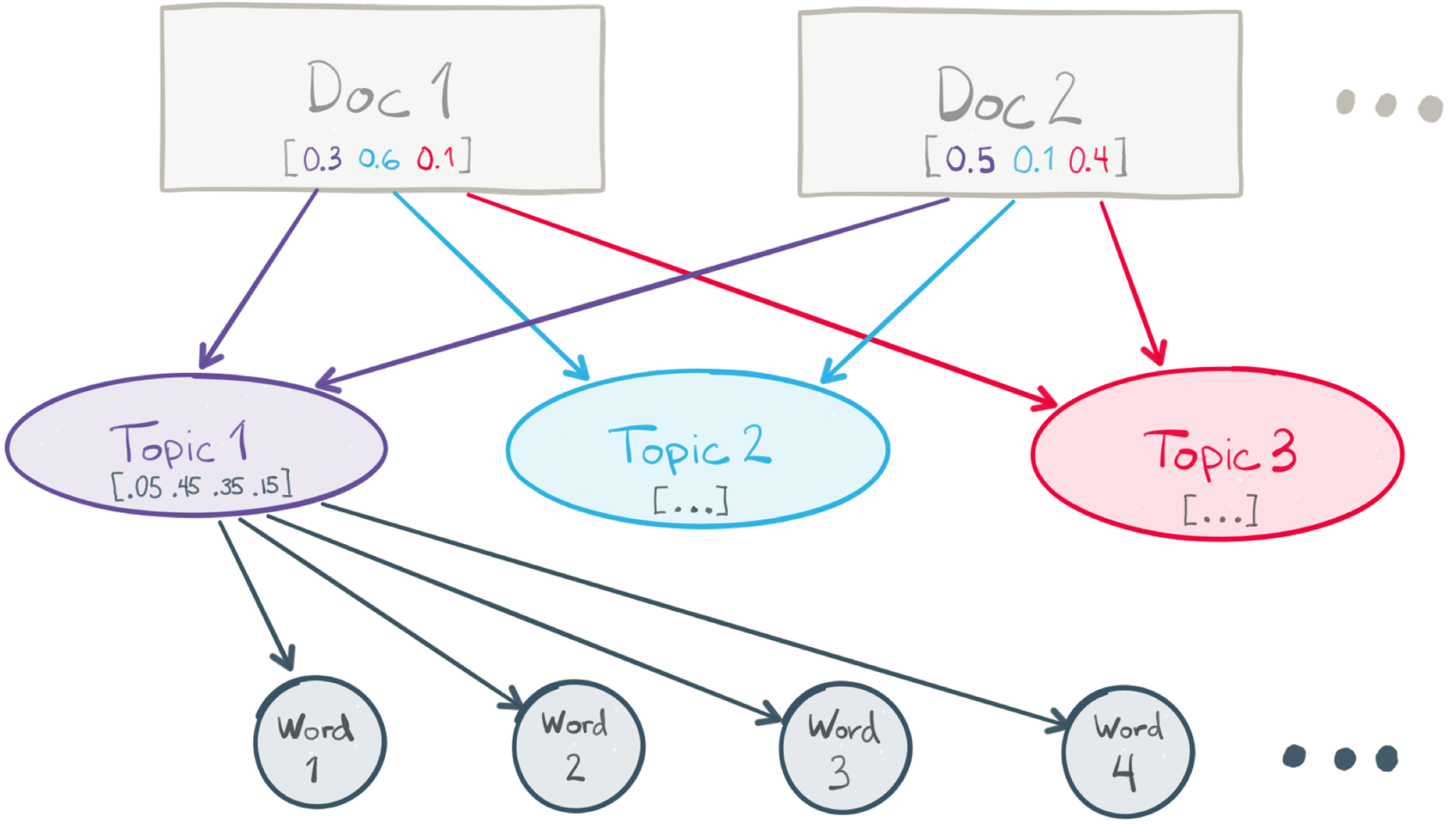
- Unsupervised learning will almost always underperform supervised learning methods.

- It is quite difficult to learn without access to our Y variable… but the tools this week will be geared toward learning as much as we can from the data that we have!
  - Topic Modeling
  - Network Analysis
  - Clustering
  - Principal Component Analysis

TOPIC MODELING

- One example of an unsupervised learning problem is **topic modeling**.
- A topic model is where we use a statistical model to attempt to identify topics or categories in a set of documents.
- We'll use a specific method called Latent Dirichlet Allocation (LDA) to model topics.
  - There exist plenty of other topic modeling algorithms, but we'll stick with LDA for today.

TOPIC MODELING



TOPIC MODELING

- Each topic is a combination of words.

- Each document is a combination of topics.

TOPIC MODELING

- Suppose you have a set of documents and you want to be able to group them into topics.
  - Why is this an unsupervised learning problem?

LATENT DIRICHLET ALLOCATION

- How might you summarize these?

1. I like to eat broccoli and bananas.
2. I ate a banana and spinach smoothie for breakfast.
3. Chinchillas and kittens are cute.
4. My sister adopted a kitten yesterday.
5. Look at this cute hamster munching on a piece of broccoli.

LATENT DIRICHLET ALLOCATION

1. I like to eat broccoli and bananas.
2. I ate a banana and spinach smoothie for breakfast.
3. Chinchillas and kittens are cute.
4. My sister adopted a kitten yesterday.
5. Look at this cute hamster munching on a piece of broccoli.

- It looks like sentences 1 and 2 are about **food**.
- It looks like sentences 3 and 4 are about **animals**.
- It looks like sentence 5 is about **food** and **animals**.

- We want a computer to be able to discover these. How?

LATENT DIRICHLET ALLOCATION: ALGORITHM

- Select some fixed number of topics T.

- Randomly assign each word to a topic.
  - This arrangement of words is presently meaningless.

- For each document d, word w, and topic t, calculate $P(t|w,d)$.
  - For every word across every document, we now have the probability that that word falls into topic 1, topic 2, and so on.
  - Randomly reassign word w to a topic based on these probabilities.
  - Repeat this a large number of times “until convergence.”

LATENT DIRICHLET ALLOCATION: ALGORITHM

- How does this work?
  - We calculate $P(t|w,d)$ using Bayes' Theorem.
    - If you want more details, check out the OneNote notebook.
  - Convergence will happen because this iterative method of assigning words to topics (known as a Markov Chain Monte Carlo method) has specific properties.
    - "Convergence" means arriving at the optimal assignments of words to topics.