150B/355B
Introduction to Machine Learning for Social Science
TA Section 5

Haemin Jee and Tongtong Zhang

February 22, 2018

Outline

Outline

Outline

Outline

Distinctive Words

Why do we care about distinctive words?

Distinctive Words

Why do we care about distinctive words?

- Create custom dictionaries if we want to do classification tasks

Distinctive Words

Why do we care about distinctive words?

- Create custom dictionaries if we want to do classification tasks
- Make interesting comparisons

Distinctive Words

Why do we care about distinctive words?

- Create custom dictionaries if we want to do classification tasks
- Make interesting comparisons
  - across genres

Distinctive Words

Why do we care about distinctive words?

- Create custom dictionaries if we want to do classification tasks
- Make interesting comparisons
  - across genres
  - across authors

Distinctive Words

Why do we care about distinctive words?

- Create custom dictionaries if we want to do classification tasks
- Make interesting comparisons
  - across genres
  - across authors
  - across politicans of different parties

Distinctive Words

Why do we care about distinctive words?

- Create custom dictionaries if we want to do classification tasks
- Make interesting comparisons
  - across genres
  - across authors
  - across politicans of different parties
  - across geographic regions

Distinctive Words

What does it mean to be distinctive? What are some metrics we could use?

- Unique usage

Distinctive Words

What does it mean to be distinctive? What are some metrics we could use?

- Unique usage *What are some problems with this?*

Distinctive Words

What does it mean to be distinctive? What are some metrics we could use?

- Unique usage *What are some problems with this?*
- Difference in frequency

Distinctive Words

What does it mean to be distinctive? What are some metrics we could use?

- Unique usage *What are some problems with this?*
- Difference in frequency *What are some problems with this?*

Distinctive Words

What does it mean to be distinctive? What are some metrics we could use?

- Unique usage *What are some problems with this?*
- Difference in frequency *What are some problems with this?*
    - Favors more frequent words

Distinctive Words

What does it mean to be distinctive? What are some metrics we could use?

- Unique usage *What are some problems with this?*
- Difference in frequency *What are some problems with this?*
  - Favors more frequent words
  - Ignores cases when one class of documents uses a word frequently and another class of documents barely uses it
- Difference in rates
- Standardized mean difference

Standardized Mean Difference

Formula:

Standardized Mean Difference

Formula:

$$\text{SMD} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Standardized Mean Difference

Formula:

$$\text{SMD} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

where μ_1 refers to the average in group 1, μ_2 refers to the average in group 2, and σ_1^2, σ_2^2 refer respectively to the variance in the two groups.

Standardized Mean Difference

Formula:

$$\text{SMD} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

where μ_1 refers to the average in group 1, μ_2 refers to the average in group 2, and σ_1^2, σ_2^2 refer respectively to the variance in the two groups.

In our example, μ_1 is the *average* times Author 1 uses the word "family", for instance, and μ_2 is the *average* times Author 2 uses that word.

Standardized Mean Difference

Formula:

$$\mathsf{SMD} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

where μ_1 refers to the average in group 1, μ_2 refers to the average in group 2, and σ_1^2, σ_2^2 refer respectively to the variance in the two groups.

In our example, μ_1 is the *average* times Author 1 uses the word “family”, for instance, and μ_2 is the *average* times Author 2 uses that word.

σ_1^2 is the variance of the usage of the word “family” among all of Author 1’s documents σ_2^2 is the variance of the usage of the same word among all of Author 2’s documents.

Back to the Goal

We may use all of these metrics (difference in rates, standardized mean difference) to create a "discriminitory" score for particular words.

Back to the Goal

We may use all of these metrics (difference in rates, standardized mean difference) to create a “discriminitory” score for particular words.

We can then use these scores and words to create a custom dictionary.

Back to the Goal

We may use all of these metrics (difference in rates, standardized mean difference) to create a "discriminitory" score for particular words.

We can then use these scores and words to create a custom dictionary.

Then we can use the dictionary for classification purposes!

Introduction to Clustering

Goal: Find documents that are similar to one another.

Introduction to Clustering

Goal: Find documents that are similar to one another.

Problem: What does “similar” mean?

Introduction to Clustering

Goal: Find documents that are similar to one another.

Problem: What does “similar” mean?

- We assume “similar” means closeness

Introduction to Clustering

Goal: Find documents that are similar to one another.

Problem: What does "similar" mean?

- We assume "similar" means closeness
- We assume "closeness" means *geometrically close*.

Introduction to Clustering

Goal: Find documents that are similar to one another.

Problem: What does "similar" mean?

- We assume "similar" means closeness
- We assume "closeness" means *geometrically close*.

Why can we do this?

Introduction to Clustering

Goal: Find documents that are similar to one another.

Problem: What does “similar” mean?

- We assume “similar” means closeness
- We assume “closeness” means *geometrically close*.

Why can we do this?

- Because we've turned raw text $\rightarrow$ counts of words (dtm)!

Introduction to Clustering

Goal: Find documents that are similar to one another.

Problem: What does "similar" mean?

- We assume "similar" means closeness
- We assume "closeness" means *geometrically close*.

Why can we do this?

- Because we've turned raw text $\rightarrow$ counts of words (dtm)!
- **Q:** What are some pitfalls of this?

Euclidean distance

One metric of geometric distance: Euclidean distance.

Euclidean distance

One metric of geometric distance: Euclidean distance.

Formula:

$$d(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sqrt{\sum_{p=1}^{P} (x_{1p} - x_{2p})^2}$$

Where p refers to the dimensions (words) of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$.

Euclidean distance

One metric of geometric distance: Euclidean distance.

Formula:

$$d(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sqrt{\sum_{p=1}^{P} (x_{1p} - x_{2p})^2}$$

Where p refers to the dimensions (words) of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$.
Pitfalls?

Euclidean distance

One metric of geometric distance: Euclidean distance.

Formula:

$$d(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sqrt{\sum_{p=1}^{P} (x_{1p} - x_{2p})^2}$$

Where p refers to the dimensions (words) of $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$.
Pitfalls?

- Euclidean distance depends on document length!

Cosine similarity

Let's say we have a document $\boldsymbol{X_i}$ (one row in a document-term matrix).
The length of $\boldsymbol{X_i}$ is:

Cosine similarity

Let's say we have a document $\boldsymbol{X_i}$ (one row in a document-term matrix). The length of $\boldsymbol{X_i}$ is:

$$
\begin{aligned}
||X_i|| &= \sqrt{\boldsymbol{X_i} \cdot \boldsymbol{X_i}} \\
&= \sqrt{(X_i1^2 + X_i2^2 + X_i3^2 + ... + X_ik^2)} \\
&= \sqrt{\sum_{k=1}^{K} X_ik^2}
\end{aligned}
$$

Cosine similarity

Let's say we have a document $\boldsymbol{X_i}$ (one row in a document-term matrix). The length of $\boldsymbol{X_i}$ is:

$$
\begin{aligned}
||X_i|| &= \sqrt{\boldsymbol{X_i} \cdot \boldsymbol{X_i}} \\
&= \sqrt{(X_i1^2 + X_i2^2 + X_i3^2 + ... + X_ik^2)} \\
&= \sqrt{\sum_{k=1}^{K} X_ik^2}
\end{aligned}
$$

The cosine similarity of two vectors $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ is:

$$
\cos\theta = \left(\frac{X_1}{||X_1||}\right) \cdot \left(\frac{X_2}{||X_2||}\right)
$$

Cosine similarity

Cosine similarity:

Cosine similarity

Cosine similarity:

- Takes into account document length

Cosine similarity

Cosine similarity:

- Takes into account document length
- Measure cosine of the angle between the vectors

Cosine similarity

Cosine similarity:

- Takes into account document length
- Measure cosine of the angle between the vectors
- Ranges from 0-1

Cosine similarity

Cosine similarity:

- Takes into account document length
- Measure cosine of the angle between the vectors
- Ranges from 0-1
- To convert to distance, take 1 - cos θ