Text Data: Topic Model

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

December 3, 2018
# Methods to be Learnt

<table>
<thead>
<tr>
<th></th>
<th>Vector Data</th>
<th>Set Data</th>
<th>Sequence Data</th>
<th>Text Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Logistic Regression; Decision Tree; KNN; SVM; NN</td>
<td></td>
<td></td>
<td>Naïve Bayes for Text</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>K-means; hierarchical clustering; DBSCAN; DBSCAN; Mixture Models</td>
<td></td>
<td></td>
<td>PLSA</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Linear Regression GLM*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Frequent Pattern Mining</strong></td>
<td>Apriori; FP growth</td>
<td>GSP; PrefixSpan</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Search</strong></td>
<td></td>
<td></td>
<td></td>
<td>DTW</td>
</tr>
</tbody>
</table>
Text Data: Topic Models

- Text Data and Topic Models
- Revisit of Mixture Model
- Probabilistic Latent Semantic Analysis (pLSA)
- Summary
Text Data

- Word/term
- Document
  - A sequence of words
- Corpus
  - A collection of documents
Represent a Document

• Most common way: Bag-of-Words
  • Ignore the order of words
  • keep the count

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>m1</th>
<th>m2</th>
<th>m3</th>
<th>m4</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interface</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>response</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>survey</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>trees</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>graph</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Vector space model
Topics

- **Topic**
  - A topic is represented by a word distribution
  - Relate to an issue
### Topic Models

- **Topic modeling**
  - Get topics automatically from a corpus
  - Assign documents to topics automatically

- Most frequently used topic models
  - pLSA
  - LDA

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts [as these] grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Text Data: Topic Models

• Text Data and Topic Models

• Revisit of Mixture Model

• Probabilistic Latent Semantic Analysis (pLSA)

• Summary
Mixture Model-Based Clustering

• A set $C$ of $k$ probabilistic clusters $C_1, ..., C_k$
  • probability density/mass functions: $f_1, ..., f_k$,
  • Cluster prior probabilities: $w_1, ..., w_k$, $\sum_j w_j = 1$

• Joint Probability of an object $i$ and its cluster $C_j$ is:
  • $P(x_i, z_i = C_j) = w_j f_j(x_i)$
  • $z_i$: hidden random variable
• Probability of $i$ is:
  • $P(x_i) = \sum_j w_j f_j(x_i)$
Maximum Likelihood Estimation

- Since objects are assumed to be generated independently, for a data set $D = \{x_1, \ldots, x_n\}$, we have,

$$P(D) = \prod_{i} P(x_i) = \prod_{i} \sum_{j} w_j f_j(x_i)$$

$$\Rightarrow \log P(D) = \sum_{i} \log P(x_i) = \sum_{i} \log \sum_{j} w_j f_j(x_i)$$

- Task: Find a set $C$ of $k$ probabilistic clusters s.t. $P(D)$ is maximized
Gaussian Mixture Model

• Generative model
  • For each object:
    • Pick its cluster, i.e., a distribution component:
      \( Z \sim \text{Multinoulli}(w_1, ..., w_K) \)
    • Sample a value from the selected distribution:
      \( X | Z \sim N(\mu_Z, \sigma_Z^2) \)

• Overall likelihood function
  • \( L(D | \theta) = \prod_i \sum_j w_j p(x_i | \mu_j, \sigma_j^2) \)
    s.t. \( \sum_j w_j = 1 \) and \( w_j \geq 0 \)
Multinomial Mixture Model

- For documents with bag-of-words representation
  - \( \mathbf{x}_d = (x_{d1}, x_{d2}, \ldots, x_{dN}) \), \( x_{dn} \) is the number of words for nth word in the vocabulary

- Generative model
  - For each document
    - Sample its cluster label \( z \sim \text{Multinoulli}(\mathbf{\pi}) \)
      - \( \mathbf{\pi} = (\pi_1, \pi_2, \ldots, \pi_K) \), \( \pi_k \) is the proportion of kth cluster
      - \( p(z = k) = \pi_k \)
    - Sample its word vector \( \mathbf{x}_d \sim \text{multinomial}(\mathbf{\beta}_z) \)
      - \( \mathbf{\beta}_z = (\beta_{z1}, \beta_{z2}, \ldots, \beta_{zN}) \), \( \beta_{zn} \) is the parameter associate with nth word in the vocabulary
      - \( p(\mathbf{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}} \propto \prod_n \beta_{kn}^{x_{dn}} \)
Likelihood Function

• For a set of M documents

\[
L = \prod_{d} p(x_d) = \prod_{d} \sum_{k} p(x_d, z = k)
\]

\[
= \prod_{d} \sum_{k} p(x_d | z = k) p(z = k)
\]

\[
\propto \prod_{d} \sum_{k} p(z = k) \prod_{n} \beta_{kn}^{x_{dn}}
\]
Mixture of Unigrams

• For documents represented by a sequence of words

\[ \mathbf{w}_d = (w_{d1}, w_{d2}, \ldots, w_{dN_d}), \quad N_d \text{ is the length of document } d, \quad w_{dn} \text{ is the word at the nth position of the document} \]

• Generative model

• For each document

  • Sample its cluster label \( z \sim \text{Multinoulli}(\pi) \)
    - \( \pi = (\pi_1, \pi_2, \ldots, \pi_K), \quad \pi_k \text{ is the proportion of kth cluster} \)
    - \( p(z = k) = \pi_k \)

  • For each word in the sequence
    - Sample the word \( w_{dn} \sim \text{Multinoulli}(\beta_z) \)
    - \( p(w_{dn}|z = k) = \beta_{kw_{dn}} \)
Likelihood Function

For a set of M documents

\[ L = \prod_{d} p(w_d) = \prod_{d} \sum_{k} p(w_d, z = k) \]

\[ = \prod_{d} \sum_{k} p(w_d | z = k) p(z = k) \]

\[ = \prod_{d} \sum_{k} p(z = k) \prod_{n} \beta_{kw_dn} \]
Question

- Are multinomial mixture model and mixture of unigrams model equivalent? Why?
Text Data: Topic Models

• Text Data and Topic Models

• Revisit of Mixture Model

• Probabilistic Latent Semantic Analysis (pLSA)

• Summary
Notations

• Word, document, topic
  • $w$, $d$, $z$

• Word count in document
  • $c(w, d)$

• Word distribution for each topic ($\beta_z$)
  • $\beta_{zw}: p(w|z)$

• Topic distribution for each document ($\theta_d$)
  • $\theta_{dz}: p(z|d)$ (Yes, soft clustering)
Issues of Mixture of Unigrams

• All the words in the same documents are sampled from the same topic

• In practice, people switch topics during their writing
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Generative Model for pLSA

• Describe how a document $d$ is generated probabilistically

  • For each position in $d$, $n = 1, \ldots, N_d$
    • Generate the topic for the position as $z_n \sim \text{Multinoulli}(\theta_d)$, i.e., $p(z_n = k) = \theta_{dk}$
      (Note, 1 trial multinomial, i.e., categorical distribution)
    • Generate the word for the position as $w_n | z_n \sim \text{Multinoulli}(\beta_{z_n})$, i.e., $p(w_n = w | z_n) = \beta_{znw}$
Note: Sometimes, people add parameters such as $\theta$ and $\beta$ into the graphical model.
The Likelihood Function for a Corpus

• Probability of a word

\[ p(w|d) = \sum_k p(w, z = k|d) = \sum_k p(w|z = k)p(z = k|d) = \sum_k \beta_{kw}\theta_{dk} \]

• Likelihood of a corpus

\[
\prod_{d=1} P(w_1, \ldots, w_{N_d}, d|\theta, \beta, \pi) \\
= \prod_{d=1} P(d) \left\{ \prod_{n=1}^{N_d} \left( \sum_k P(z_n = k|d, \theta_d)P(w_n|\beta_k) \right) \right\} \\
= \prod_{d=1} \pi_d \left\{ \prod_{n=1}^{N_d} \left( \sum_k \theta_{dk}\beta_{kw_n} \right) \right\}
\]

\( \pi_d\) is usually considered as uniform, i.e., 1/M
Re-arrange the Likelihood Function

• Group the same word from different positions together

\[
\max \log L = \sum_{dw} c(w, d) \log \sum_{z} \theta_{dz} \beta_{zw}
\]

\[s.t. \sum_{z} \theta_{dz} = 1 \text{ and } \sum_{w} \beta_{zw} = 1\]
Optimization: EM Algorithm

• Repeat until converge

  • E-step: for each word in each document, calculate its conditional probability belonging to each topic
    \[ p(z|w,d) \propto p(w|z,d)p(z|d) = \beta_{zw}\theta_{dz} \text{ (i.e., } p(z|w,d) \text{)} \]
    \[ = \frac{\beta_{zw}\theta_{dz}}{\sum_{z'} \beta_{z'w}\theta_{dz'}} \]

  • M-step: given the conditional distribution, find the parameters that can maximize the expected complete log-likelihood
    \[ \beta_{zw} \propto \sum_d p(z|w,d)c(w,d) \text{ (i.e., } \beta_{zw} = \frac{\sum_d p(z|w,d)c(w,d)}{\sum_{w',d} p(z|w',d)c(w',d)} \) \]
    \[ \theta_{dz} \propto \sum_w p(z|w,d)c(w,d) \text{ (i.e., } \theta_{dz} = \frac{\sum_w p(z|w,d)c(w,d)}{N_d} \)} \]
Example

- Two documents, two topics
  - Vocabulary: \{1: data, 2: mining, 3: frequent, 4: pattern, 5: web, 6: information, 7: retrieval\}

- At some iteration of EM algorithm, E-step

| word \((w)\) | word count in Document 1 \((c(w,d_1))\) | \(p(z = 1|w,d_1)\) |
|-------------|-----------------------------------|-------------------|
| data        | 5                                 | 0.8               |
| mining      | 4                                 | 0.8               |
| frequent    | 3                                 | 0.6               |
| pattern     | 2                                 | 0.8               |
| web         | 2                                 | 0.5               |
| information | 1                                 | 0.2               |

| word \((w)\) | word count in Document 2 \((c(w,d_2))\) | \(p(z = 1|w,d_2)\) |
|-------------|-----------------------------------|-------------------|
| information | 5                                 | 0.2               |
| retrieval   | 4                                 | 0.2               |
| web         | 3                                 | 0.1               |
| mining      | 3                                 | 0.5               |
| frequent    | 2                                 | 0.6               |
| data        | 2                                 | 0.5               |
Example (Continued)

• M-step

\[
\beta_{11} = \frac{0.8 \times 5 + 0.5 \times 2}{11.8 + 5.8} = \frac{5}{17.6}
\]

\[
\beta_{12} = \frac{0.8 \times 4 + 0.5 \times 3}{11.8 + 5.8} = \frac{4.7}{17.6}
\]

\[
\beta_{13} = \frac{3}{17.6}
\]

\[
\beta_{14} = \frac{1.6}{17.6}
\]

\[
\beta_{15} = \frac{1.3}{17.6}
\]

\[
\beta_{16} = \frac{1.2}{17.6}
\]

\[
\beta_{17} = \frac{0.8}{17.6}
\]

\[
\theta_{11} = \frac{11.8}{17}
\]

\[
\theta_{12} = \frac{5.2}{17}
\]
Text Data: Topic Models

• Text Data and Topic Models

• Revisit of Mixture Model

• Probabilistic Latent Semantic Analysis (pLSA)

• Summary
Summary

• Basic Concepts
  • Word/term, document, corpus, topic

• Mixture of unigrams

• pLSA
  • Generative model
  • Likelihood function
  • EM algorithm
CS145: INTRODUCTION TO DATA MINING

Final Review

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

December 5, 2018
# Learnt Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Vector Data</th>
<th>Set Data</th>
<th>Sequence Data</th>
<th>Text Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Logistic Regression; Decision Tree; KNN; SVM; NN</td>
<td></td>
<td></td>
<td>Naïve Bayes for Text</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>K-means; hierarchical clustering; DBSCAN; Mixture Models</td>
<td></td>
<td></td>
<td>PLSA</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Linear Regression GLM*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Frequent Pattern Mining</strong></td>
<td>Apriori; FP growth</td>
<td></td>
<td>GSP; PrefixSpan</td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Search</strong></td>
<td></td>
<td></td>
<td></td>
<td>DTW</td>
</tr>
</tbody>
</table>
Deadlines ahead

- Homework 6 (optional)
  - Sunday, Dec. 9, 2018, 11:59pm
  - We will pick 5 highest score

- Course project
  - Tuesday, Dec. 11, 2018, 11:59 pm.
  - Don’t forget peer evaluation form:
    - One question is: “Do you think some member in your group should be given a lower score than the group score? If yes, please list the name, and explain why.”
Final Exam

• Time
  • 12/13, 11:30am-2:30pm

• Location
  • Franz Hall: 1260

• Policy
  • Closed book exam
  • You can take two “reference sheets” of A4 size, i.e., one in addition to the midterm “reference sheet”
  • You can bring a simple calculator
## Content to Cover

- All the content learned so far
  - ~20% before midterm
  - ~80% after midterm

<table>
<thead>
<tr>
<th></th>
<th>Vector Data</th>
<th>Set Data</th>
<th>Sequence Data</th>
<th>Text Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Logistic Regression; Decision Tree; KNN; SVM; NN</td>
<td></td>
<td></td>
<td>Naïve Bayes for Text</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>K-means; hierarchical clustering; DBSCAN; DBSCAN; Mixture Models</td>
<td></td>
<td></td>
<td>PLSA</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Linear Regression GLM*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Frequent Pattern Mining</strong></td>
<td>Apriori; FP growth</td>
<td></td>
<td>GSP; PrefixSpan</td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Search</strong></td>
<td></td>
<td></td>
<td></td>
<td>DTW</td>
</tr>
</tbody>
</table>
Type of Questions

• Similar to Midterm
  • True or false

• Conceptual questions

• Computation questions
Sample question on DTW

• Suppose that we have two sequences S1 and S2 as follows:
  • S1 = <1, 2, 5, 3, 2, 1, 7>
  • S2 = <2, 3, 2, 1, 7, 4, 3, 0, 2, 5>

• Compute the distance between two sequences according to the dynamic time warping algorithm.
What’s next?

• Following-up courses

  • **CS247: Advanced Data Mining**
    • Focus on Text, Recommender Systems, and Networks/Graphs
    • Will be offered in Spring 2019

  • **CS249: Probabilistic Graphical Models for Structured Data**
    • Focus on Probabilistic Models on text and graph data
    • Research seminar: You are expected to read papers and present to the whole class; you are expected to write a survey or conduct a course project
    • Will be offered in Winter 2019
Thank you and good luck!

• Give us feedback by submitting evaluation form
• 1-2 undergraduate research intern positions are available
  • Please refer to piazza: https://piazza.com/class/jmpc925zfo92qs?cid=274