

# CS145: INTRODUCTION TO DATA MINING

## Text Data: Topic Model

---

**Instructor: Yizhou Sun**

[yzsun@cs.ucla.edu](mailto:yzsun@cs.ucla.edu)


March 8, 2019

# Methods to be Learnt

	Vector Data	Set Data	Sequence Data	Text Data
Classification	Logistic Regression; Decision Tree; KNN; SVM; NN			Naïve Bayes for Text
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models			PLSA
Prediction	Linear Regression GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

# 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

c1: *Human machine interface* for Lab ABC *computer* applications  
c2: A *survey* of *user* opinion of *computer system response time*  
c3: The *EPS user interface* management *system*  
c4: *System* and *human system* engineering testing of *EPS*  
c5: Relation of *user-perceived response time* to error measurement

m1: The generation of random, binary, unordered *trees*  
m2: The intersection *graph* of paths in *trees*  
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering  
m4: *Graph minors*: A *survey*



	c1	c2	c3	c4	c5	m1	m2	m3	m4
<i>human</i>	1	0	0	1	0	0	0	0	0
<i>interface</i>	1	0	1	0	0	0	0	0	0
<i>computer</i>	1	1	0	0	0	0	0	0	0
<i>user</i>	0	1	1	0	1	0	0	0	0
<i>system</i>	0	1	1	2	0	0	0	0	0
<i>response</i>	0	1	0	0	1	0	0	0	0
<i>time</i>	0	1	0	0	1	0	0	0	0
<i>EPS</i>	0	0	1	1	0	0	0	0	0
<i>survey</i>	0	1	0	0	0	0	0	0	1
<i>trees</i>	0	0	0	0	0	1	1	1	0
<i>graph</i>	0	0	0	0	0	0	1	1	1
<i>minors</i>	0	0	0	0	0	0	0	1	1

Vector space model

# Topics

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

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148
bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824



# Topic Models


- Topic modeling
  - Get topics automatically from a corpus
  - Assign documents to topics automatically
- Most frequently used topic models
  - pLSA
  - LDA

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

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.

# Text Data: Topic Models

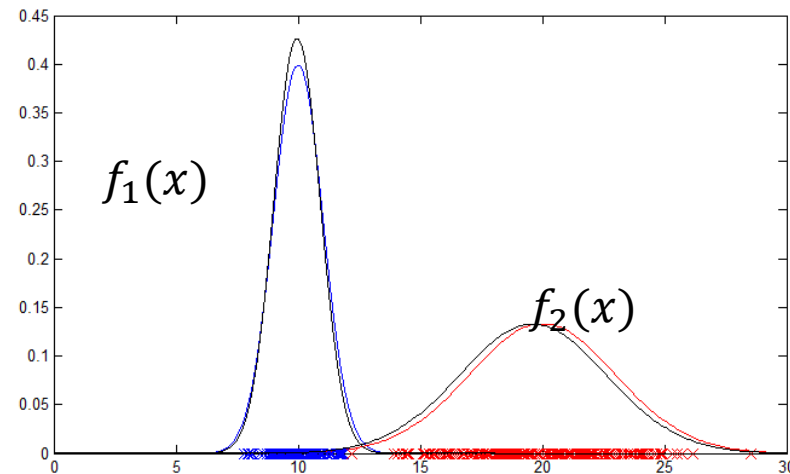
---

- 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  $C_1, \dots, C_k$ 
  - probability density/mass functions:  $f_1, \dots, f_k$ ,
  - Cluster prior probabilities:  $w_1, \dots, w_k, \sum_j w_j = 1$
- 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  $D = \{x_1, \dots, x_n\}$ , 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:  
 $Z \sim \text{Categorical}(w_1, \dots, w_k)$
    - 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
  - $\mathbf{x}_d = (x_{d1}, x_{d2}, \dots, x_{dN})$ ,  $x_{dn}$  is the number of words for nth word in the vocabulary
- Generative model
  - For each document
    - Sample its cluster label  $z \sim \text{Categorical}(\boldsymbol{\pi})$ 
      - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$ ,  $\pi_k$  is the proportion of kth cluster
      - $p(z = k) = \pi_k$
    - Sample its word vector  $\mathbf{x}_d \sim \text{multinomial}(\boldsymbol{\beta}_z)$ 
      - $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \dots, \beta_{zN})$ ,  $\beta_{zn}$  is the parameter associate with nth word in the vocabulary
      - $p(\mathbf{x}_d | z = k) = \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \prod_n \beta_{kn}^{x_{dn}}$

# Likelihood Function

---

- For a set of  $M$  documents

$$\begin{aligned} L &= \prod_d p(\mathbf{x}_d) = \prod_d \sum_k p(\mathbf{x}_d, z = k) \\ &= \prod_d \sum_k p(\mathbf{x}_d | z = k) p(z = k) \\ &= \prod_d \frac{(\sum_n x_{dn})!}{\prod_n x_{dn}!} \sum_k p(z = k) \prod_n \beta_{kn}^{x_{dn}} \end{aligned}$$

# Mixture of Unigrams

---

- For documents represented by a sequence of words
  - $\mathbf{w}_d = (w_{d1}, w_{d2}, \dots, w_{dN_d})$ ,  $N_d$  is the length of document  $d$ ,  $w_{di}$  is the word at the  $i$ th position of the document
- Generative model
  - For each document
    - Sample its cluster label  $z \sim \text{Categorical}(\boldsymbol{\pi})$ 
      - $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$ ,  $\pi_k$  is the proportion of  $k$ th cluster
      - $p(z = k) = \pi_k$
    - For each word in the sequence
      - Sample the word  $w_{di} \sim \text{Categorical}(\boldsymbol{\beta}_z)$
      - $p(w_{di} | z = k) = \beta_{kw_{di}}$

# Likelihood Function

---

- For a set of  $M$  documents

$$\begin{aligned} L &= \prod_d p(\mathbf{w}_d) = \prod_d \sum_k p(\mathbf{w}_d, z = k) \\ &= \prod_d \sum_k p(\mathbf{w}_d | z = k) p(z = k) \\ &= \prod_d \sum_k p(z = k) \prod_i \beta_{kw_{di}} \end{aligned}$$

# 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

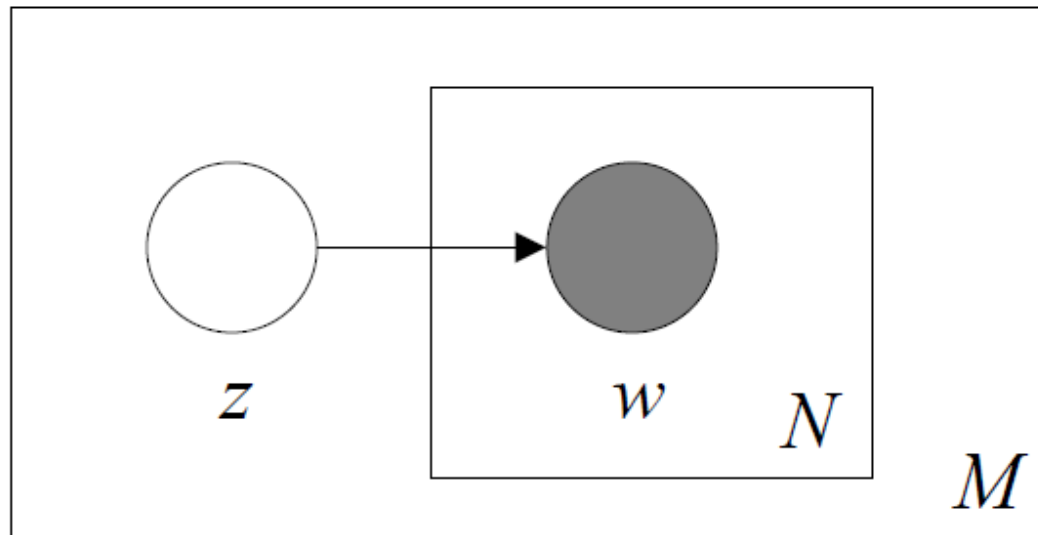
# Notations

---

- 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

# Illustration of pLSA

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

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  $d$ ,  $n = 1, \dots, N_d$

- Generate the topic for the position as

$$z_n \sim \text{Categorical}(\boldsymbol{\theta}_d), \text{ i. e. }, p(z_n = k) = \theta_{dk}$$

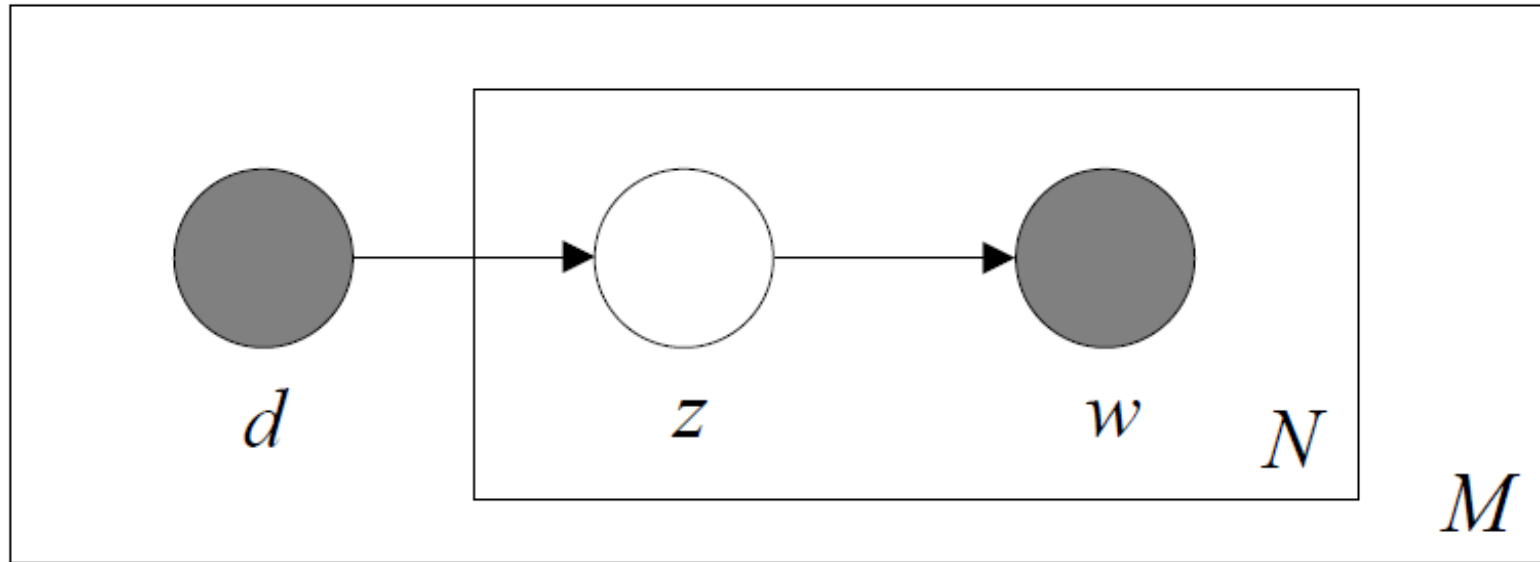
(Note, 1 trial multinomial)

- Generate the word for the position as

$$w_n | z_n \sim \text{Categorical}(\boldsymbol{\beta}_{z_n}), \text{ i. e. }, p(w_n = w | z_n) = \beta_{z_n w}$$

# Graphical Model

---



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

$$\begin{aligned} & \prod_{d=1}^D P(w_1, \dots, w_{N_d}, d | \theta, \beta, \pi) \\ &= \prod_{d=1}^D P(d) \left\{ \prod_{n=1}^{N_d} \left( \sum_k P(z_n = k | d, \theta_d) P(w_n | \beta_k) \right) \right\} \\ &= \prod_{d=1}^D \pi_d \left\{ \prod_{n=1}^{N_d} \left( \sum_k \theta_{dk} \beta_{kw_n} \right) \right\} \end{aligned}$$

*$\pi_d$  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
$$p(z|w, d) \propto p(w|z, d)p(z|d) = \beta_{zw}\theta_{dz} \text{ (i.e., } p(z|w, d) = \frac{\beta_{zw}\theta_{dz}}{\sum_{z'} \beta_{z'w}\theta_{dz'}})$$
  - 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}

- A 

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

$$\beta_{11} = \frac{0.8 * 5 + 0.5 * 2}{11.8 + 5.8} = 5/17.6$$

$$\beta_{12} = \frac{0.8 * 4 + 0.5 * 3}{11.8 + 5.8} = 4.7/17.6$$

$$\beta_{13} = 3/17.6$$

$$\beta_{14} = 1.6/17.6$$

$$\beta_{15} = 1.3/17.6$$

$$\beta_{16} = 1.2/17.6$$


$$\beta_{17} = 0.8/17.6$$

$$\theta_{11} = \frac{11.8}{17}$$

$$\theta_{12} = \frac{5.2}{17}$$

# 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

# Learnt Algorithms

	Vector Data	Set Data	Sequence Data	Text Data
Classification	Logistic Regression; Decision Tree; KNN; SVM; NN			Naïve Bayes for Text
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models			PLSA
Prediction	Linear Regression GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

# Deadlines ahead

---

- Homework 6 (optional)
  - Thursday, Mar. 14, 2019, 11:59pm
  - We will pick 5 highest score
- Course project
  - Mar. 20, 2019, 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
  - Tuesday, 3/19, 8am-11am
- Location
  - BROAD 2100A
- 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 midt

	Vector Data	Set Data	Sequence Data	Text Data
Classification	Logistic Regression; Decision Tree; KNN; SVM; NN			Naïve Bayes for Text
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models			PLSA
Prediction	Linear Regression GLM*			
Frequent Pattern Mining		Apriori; FP growth	GSP; PrefixSpan	
Similarity Search			DTW	

# 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:
  - $S1 = \langle 1, 2, 5, 3, 2, 1, 7 \rangle$
  - $S2 = \langle 2, 3, 2, 1, 7, 4, 3, 0, 2, 5 \rangle$
- 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 2020

# Thank you and good luck!

---

- Give us feedback by submitting evaluation form (counted as bonus for participation)
- 1-2 undergraduate research intern positions are available