Machine Learning: Chenhao Tan

University of Colorado Boulder

LECTURE 25

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Logistics

- HW5 is due this week.
- Project mid-point check-in on Wednesday!

Learning objectives

- Learn about formulation of topic models
- A preview of the learning theory

Outline

Topic models

PAC learnability

Topic models

- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

Topic models

Neat way to explore/understand corpus collections

- E-discovery
- Social media
- Scientific data

NLP Applications

- Word sense disambiguation
- Discourse segmentation

Psychology: word meaning, polysemy

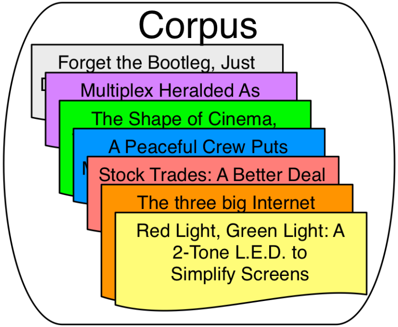
A general way to model count data and a general inference algorithm

Conceptual approach

Input: a text corpus and number of topics K

Output:

- Topic assignment for each document
- K topics, each topic is a list of words



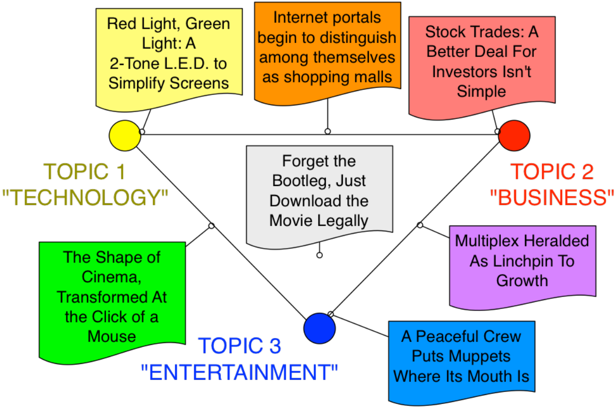
Conceptual approach

K topics, each topic is a list of words

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

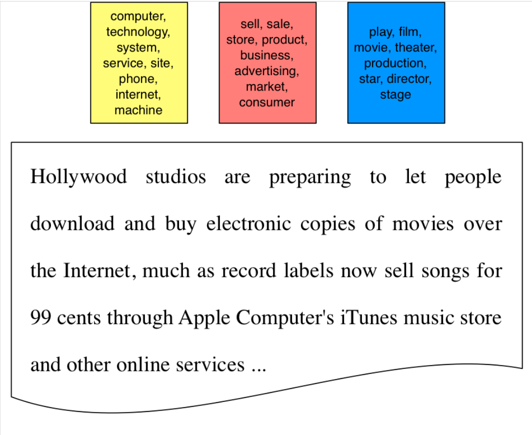
Conceptual approach

Topic assignment for each document



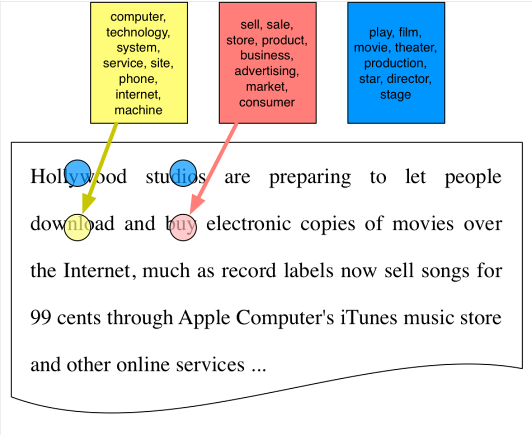
Conceptual approach

Generate each word



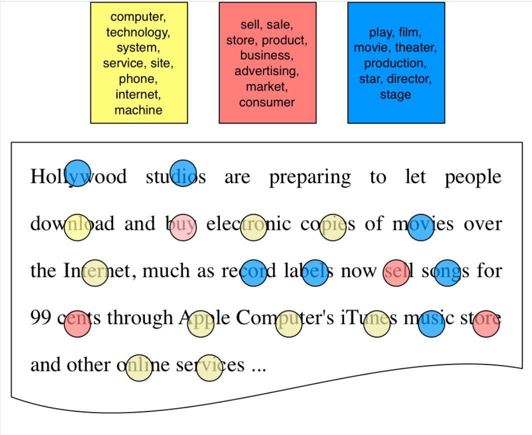
Conceptual approach

Generate each word



Conceptual approach

Generate each word



Conceptual approach

Real topics learned from Science

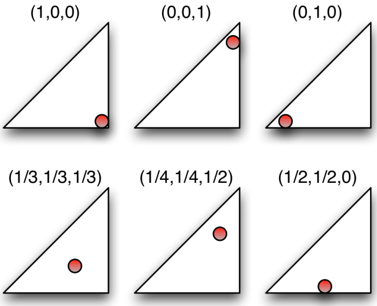
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Latent Dirichlet Allocation: Generative story

- Discrete count data
- Gaussian distributions are not appropriate

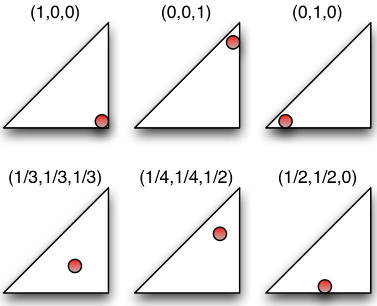
Latent Dirichlet Allocation: Generative story

- Generate a document, or a bag of words
- Blei, Ng, Jordan. Latent Dirichlet Allocation. JMLR, 2003.



Latent Dirichlet Allocation: Generative story

- Generate a document, or a bag of words Multinomial distribution
  - Distribution over discrete outcomes
  - Represented by non-negative vector that sums to one
  - Picture representation
  - Can be generated from a Dirichlet distribution



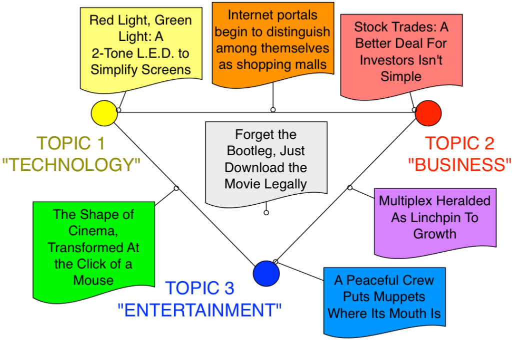
Latent Dirichlet Allocation: Generative story

Generate K topics: $\beta_k, k = 1, \dots, K; \sum_{i=1}^{V} \beta_{ki} = 1$ (Vocabulary size V)

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

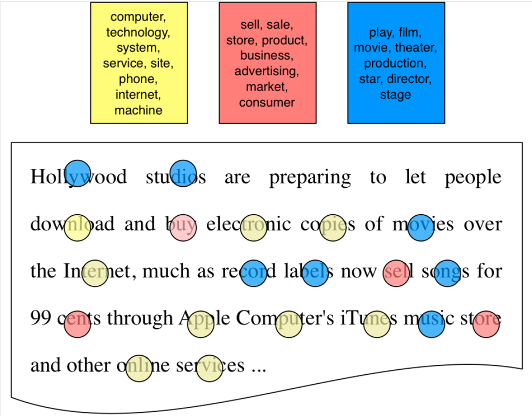
Latent Dirichlet Allocation: Generative story

Generate topic assignments for each document: $\theta_d, d = 1, \dots, M; \sum_{i=1}^{K} \theta_{dk} = 1$

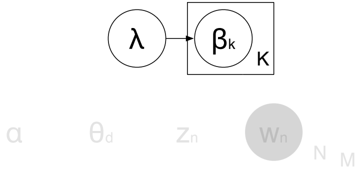


Latent Dirichlet Allocation: Generative story

Generate each word in a document by first sampling from $z \sim \text{Multinomial}(\theta_d)$ and then $w \sim \text{Multinomial}(\beta_z)$

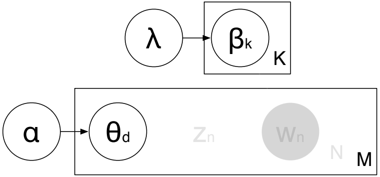


Making the generative story formal



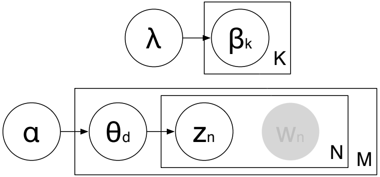
- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Making the generative story formal



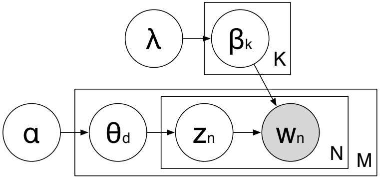
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Making the generative story formal



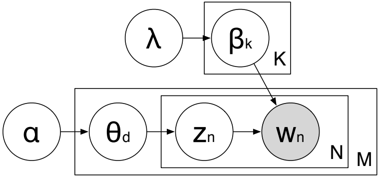
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

Making the generative story formal



- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

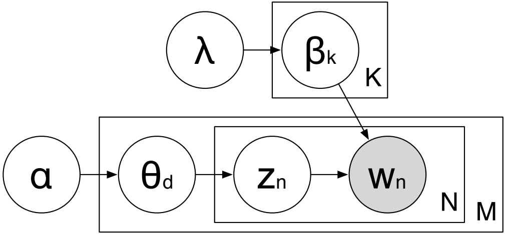
Making the generative story formal



- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

Learning is not required in this class.

Parameters to estimate



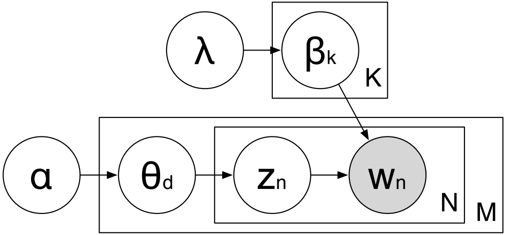
A. β

B. θ

C. β, θ

D. β, θ, z

Parameter size



Given M documents, each document N_d words, vocabulary size V, what is the size of the parameters if we are going to learn K topics?

- β
- θ
- z

Outline

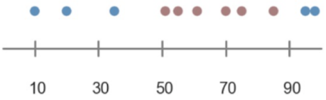
Topic models

PAC learnability

A motivating example

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Alien has a perfect alien thermometer
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have concept $c(x)$ of *nice*
- Alien wants to learn hypothesis $h(x)$

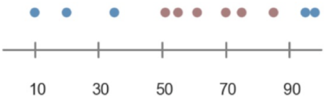
What does it mean that Alien has learned?



A motivating example

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Alien has a perfect alien thermometer
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have concept $c(x)$ of *nice*
- Alien wants to learn hypothesis $h(x)$

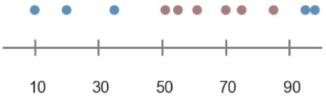
How many locals does he need to ask to get $h(x)$ that is 99% accurate?



A motivating example

- Alien moves to Colorado
- Want to talk to locals about weather
- Specifically about when weather is *nice*
- Alien has a perfect alien thermometer
- Asks a bunch of locals if it's *nice* out
- Gets labeled observations $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{m}$
- Coloradans have concept $c(x)$ of *nice*
- Alien wants to learn hypothesis $h(x)$

How many locals does he need to ask to get $h(x)$ that is 99% accurate about 99% of the time?



PAC learnability

Assumptions:

- Data comes from distribution $\mathcal{D}$
- Concept $c : X \rightarrow Y$ comes from concept class C
- Hypothesis $h : X \rightarrow Y$ comes from hypothesis class H

Generalization Error

$$R(h) = Pr_{x \sim D}\left[h(x) \neq c(x)\right] = E_{x \sim D}[\ I[h(x) \neq c(x)]\ ]$$

Goal: Given a set of data S of size m, can we learn a hypothesis h that we can say is **accurate** with high **confidence**?

PAC learnability

We say that a concept is PAC-Learnable if we can find a hypothesis that is **P**robably **A**pproximately **C**orrect using a training set S of size m where m isn't too large

$$R(h_S) \leq \epsilon$$

- Approximately correct: Accuracy is $1 - \epsilon$

PAC learnability

We say that a concept is PAC-Learnable if we can find a hypothesis that is **P**robably **A**pproximately **C**orrect using a training set S of size m where m isn't too large

$$Pr_{S \sim \mathcal{D}^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

- Approximately correct: Accuracy is $1 - \epsilon$
- Probably: Confidence in hypothesis is $1 - \delta$

PAC = Probably Approximately Correct

PAC learnability

PAC Learnability

A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$

$$Pr_{S \sim \mathcal{D}^m}\left[R(h_S) \leq \epsilon\right] \geq 1 - \delta$$

for any $c \in C$ and any distribution $\mathcal{D}$ for any sample size $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$.

PAC learnability

PAC Learnability

A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$

$$Pr_{S \sim \mathcal{D}^m} \left[R(h_S) \leq \epsilon \right] \geq 1 - \delta$$

for any $c \in C$ and any distribution $\mathcal{D}$ for any sample size $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$.

- S: The training set we learn from
- $\mathcal{D}$: The distribution the data comes from
- h_S: The hypothesis we learn from training set

PAC learnability

PAC Learnability

A concept from class C is PAC-Learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function f such that for any $\epsilon > 0$ and any $\delta > 0$

$$Pr_{S \sim \mathcal{D}^m} \left[R(h_S) \leq \epsilon \right] \geq 1 - \delta$$

for any $c \in C$ and any distribution $\mathcal{D}$ for any sample size $m \geq f\left(1/\epsilon, 1/\delta, n, |C|\right)$.

- $R(h_S)$: The generalization error of h_S
- $1 - \epsilon$: The accuracy of h_s
- $1 - \delta$: The confidence the accuracy $1 - \epsilon$ is realized