Inference and Representation

David Sontag

New York University

Lecture 1, September 2, 2014

One of the **most exciting advances** in machine learning (AI, signal processing, coding, control, ...) in the last decades

How can we gain **global insight** based on **local observations**?

Key idea

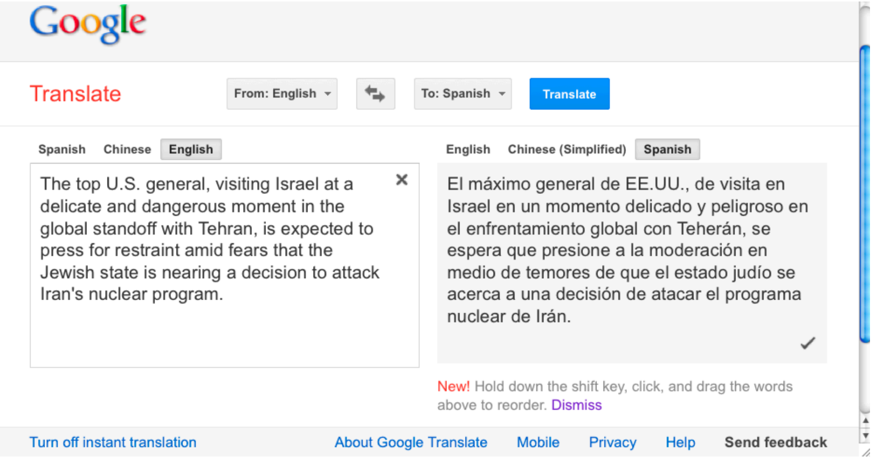
1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$
2. **Learn** the distribution from data
3. Perform "**inference**" (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)

Reasoning under uncertainty

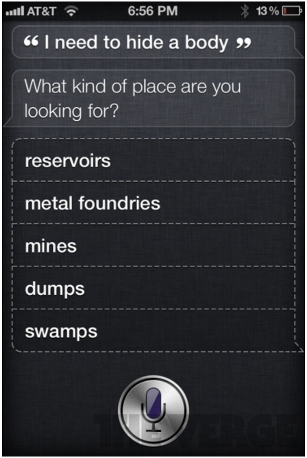
- As humans, we are continuously making predictions under uncertainty
- Classical AI and ML research ignored this phenomena
- Many of the most recent advances in technology are possible because of this new, *probabilistic*, approach

Applications: Deep question answering

Applications: Machine translation

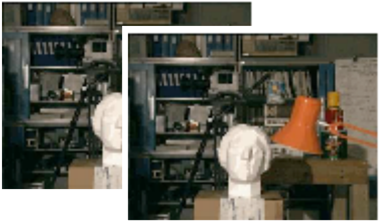


Applications: Speech recognition



Applications: Stereo vision

input: two images



output: disparity

Key challenges

1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$
   - How does one *compactly describe* this joint distribution?
   - Directed graphical models (Bayesian networks)
   - Undirected graphical models (Markov random fields, factor graphs)
2. **Learn** the distribution from data
   - Maximum likelihood estimation. Other estimation methods?
   - How much data do we need?
   - How much computation does it take?
3. Perform "**inference**" (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)

Syllabus overview

- We will study Representation, Inference & Learning
- First in the simplest case
  - Only discrete variables
  - Fully observed models
  - Exact inference & learning
- Then generalize
  - Continuous variables
  - Partially observed data during learning (hidden variables)
  - *Approximate* inference & learning
- Learn about algorithms, theory & applications

Logistics: class

- **Class webpage:**
  - `http://cs.nyu.edu/~dsontag/courses/inference14/`
  - Sign up for mailing list!
- **Book:** *Machine Learning: a Probabilistic Perspective* by Kevin Murphy, MIT Press (2012)
  - Required readings for each lecture posted to course website.
  - A good optional reference is *Probabilistic Graphical Models: Principles and Techniques* by Daphne Koller and Nir Friedman, MIT Press (2009)
- **Office hours:** Tuesdays 10:30-11:30am. 715 Broadway, 12th floor, Room 1204
- **Lab:** Thursdays, 5:10-6:00pm in Silver Center 401
  - Instructor: Yacine Jernite (jernite@cs.nyu.edu)
  - Required attendance; no exceptions.
- **Grader:** Prasoon Goyal (pg1338@nyu.edu)

Logistics: prerequisites & grading

- **Prerequisite:**
  - DS-GA-1003/CSCI-GA.2567 (Machine Learning and Computational Statistics)
  - Exceptions to the prerequisite *must* be confirmed by me (via email), and are only likely to be granted to PhD students
- **Grading:** problem sets (55%) + in class midterm exam (20%) + in class final exam (20%) + participation (5%)
  - Class attendance is required.
  - 7-8 assignments (every 1–2 weeks). Both theory and programming.
  - First homework out **today**, due Monday Sept. 15 at 10pm (via email)
  - **Important:** See collaboration policy on class webpage
- Solutions to the theoretical questions require formal proofs.
- For the programming assignments, I recommend Python (Java or Matlab OK too). Do not use C++.

Example: Medical diagnosis

- Variable for each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "pneumonia", "flu", "common cold", "bronchitis", "tuberculosis")
- Diagnosis is performed by **inference** in the model:

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}
- **Estimation** of joint distribution would require a huge amount of data
- **Inference** of conditional probabilities, e.g.

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

  would require summing over exponentially many variables' values
- Moreover, defeats the purpose of probabilistic modeling, which is to make predictions with *previously unseen observations*

Structure through independence

- If $X_1, \ldots, X_n$ are independent, then

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n)$$

- 2^n entries can be described by just n numbers (if $|\text{Val}(X_i)| = 2$)!
- However, this is not a very *useful* model – observing a variable X_i cannot influence our predictions of X_j
- If $X_1, \ldots, X_n$ are *conditionally independent* given Y, denoted as $X_i \perp \mathbf{X}_{-i} \mid Y$, then

$$\begin{aligned} p(y, x_1, \ldots, x_n) &= p(y)p(x_1 \mid y) \prod_{i=2}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, y) \\ &= p(y)p(x_1 \mid y) \prod_{i=2}^{n} p(x_i \mid y). \end{aligned}$$

- This is a simple, yet *powerful*, model

Example: naive Bayes for classification

- Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $1 : n$ index the words in our vocabulary (e.g., English)
  - $X_i = 1$ if word i appears in an e-mail, and 0 otherwise
  - E-mails are drawn according to some distribution $p(Y, X_1, \ldots, X_n)$
- Suppose that the words are conditionally independent given Y. Then,

$$p(y, x_1, \ldots x_n) = p(y) \prod_{i=1}^{n} p(x_i \mid y)$$

Estimate the model with maximum likelihood. **Predict** with:

$$p(Y = 1 \mid x_1, \ldots x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i \mid Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i \mid Y = y)}$$

- Are the independence assumptions made here reasonable?
- Philosophy: Nearly all probabilistic models are "wrong", but many are nonetheless useful

Bayesian networks

Reference: Chapter 10

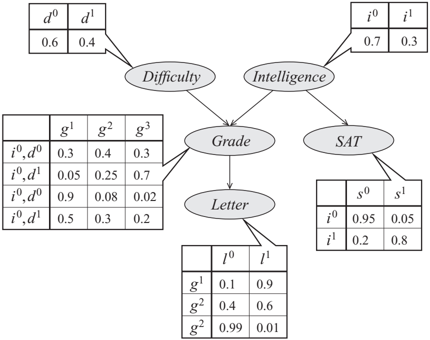
- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations
- Enables use of *prior knowledge* to specify (part of) model structure

Example

- Consider the following Bayesian network:



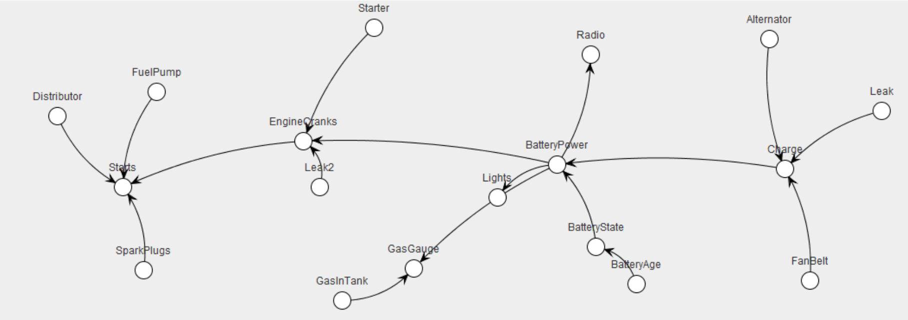
- What is its joint distribution?

$$
\begin{aligned}
p(x_1, \ldots x_n) &= \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)}) \\
p(d, i, g, s, l) &= p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)
\end{aligned}
$$

More examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

Will my car start this morning?



Heckerman *et al.*, Decision-Theoretic Troubleshooting, 1995

More examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

What is the differential diagnosis?

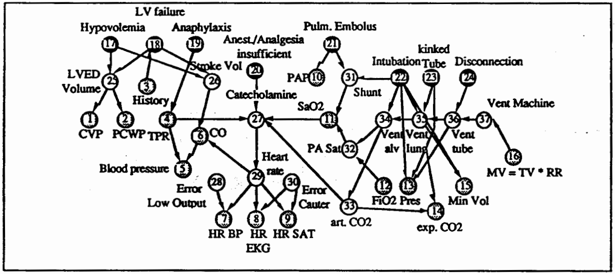
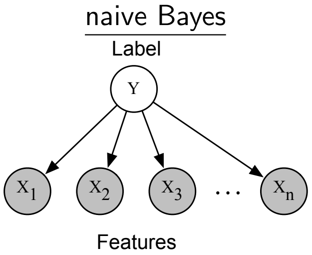


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (O) and measurement (⊙) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

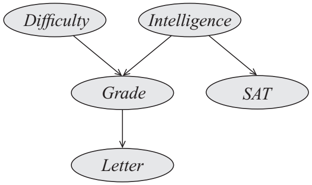
Beinlich *et al.*, The ALARM Monitoring System, 1989

Bayesian networks are *generative models*



- Evidence is denoted by shading in a node
- Can interpret Bayesian network as a **generative process**. For example, to *generate* an e-mail, we
  1. Decide whether it is spam or not spam, by samping $y \sim p(Y)$
  2. For each word $i = 1$ to n, sample $x_i \sim p(X_i \mid Y = y)$

Bayesian network structure implies conditional independencies!



- The joint distribution corresponding to the above BN factors as

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

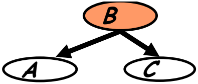
- However, by the chain rule, *any* distribution can be written as

$$p(d, i, g, s, l) = p(d)p(i \mid d)p(g \mid i, d)p(s \mid i, d, g)p(l \mid g, d, i, g, s)$$

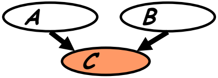
- Thus, we are assuming the following additional independencies:
$D \perp I$, $\quad S \perp \{D, G\} \mid I$, $\quad L \perp \{I, D, S\} \mid G$. $\quad$ What else?

Bayesian network structure implies conditional independencies!

- Generalizing the above arguments, we obtain that a variable is independent from its non-descendants given its parents

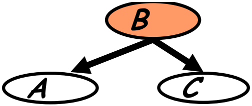


- **Common parent** – fixing B *decouples* A and C
- **Cascade** – knowing B *decouples* A and C



- **V-structure** – Knowing C *couples* A and B
  - This important phenomona is called **explaining away** and is what makes Bayesian networks so powerful

A simple justification (for common parent)



We'll show that $p(A, C \mid B) = p(A \mid B)p(C \mid B)$ for *any* distribution $p(A, B, C)$ that factors according to this graph structure, i.e.

$$p(A, B, C) = p(B)p(A \mid B)p(C \mid B)$$

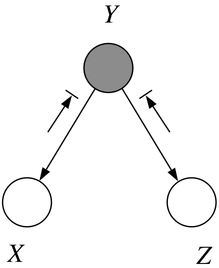
Proof.

$$p(A, C \mid B) = \frac{p(A, B, C)}{p(B)} = p(A \mid B)p(C \mid B)$$

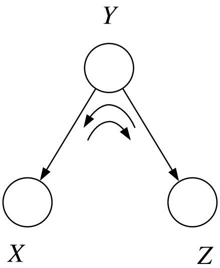
$\square$

D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



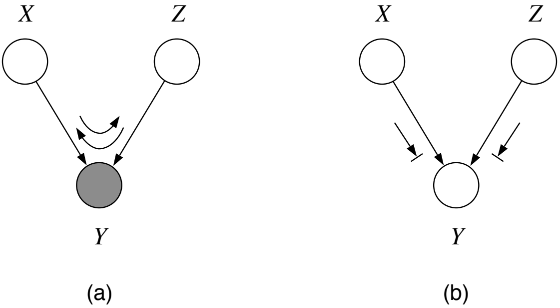
(a)



(b)

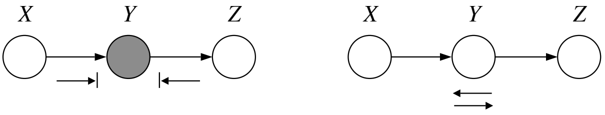
D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



D-separation ("directed separated") in Bayesian networks

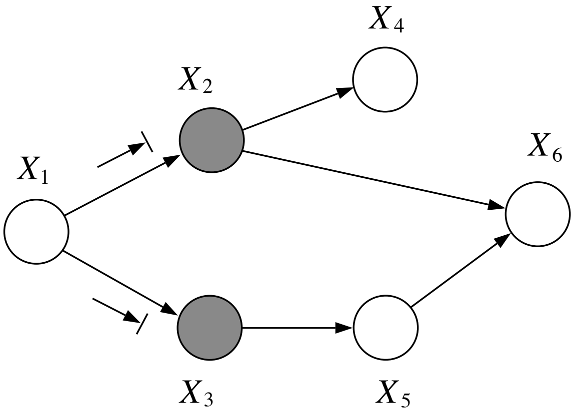
- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



(a) (b)

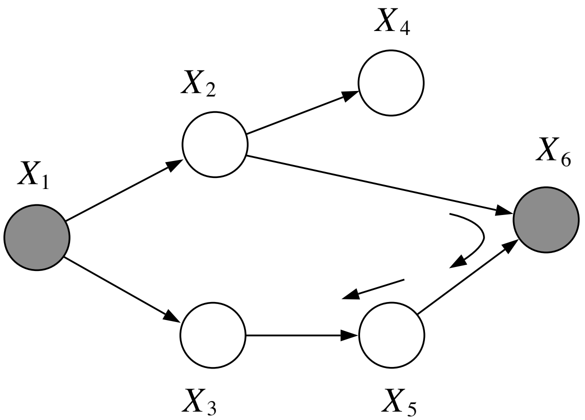
- If no such path, then X and Z are **d-separated** with respect to $\mathbf{Y}$
- d-separation reduces statistical independencies (hard) to connectivity in graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query

D-separation example 1



Is $X_6 \perp X_5 \mid X_2, X_3$? Is $X_4 \perp X_5 \mid X_2, X_3$?

D-separation example 2



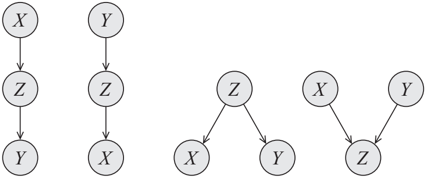
Is $X_4 \perp X_5 \mid X_1, X_6$? What about is X_6 is not observed? I.e., is $X_4 \perp X_5 \mid X_1$?

Independence maps

- Let $I(G)$ be the set of all conditional independencies implied by the directed acyclic graph (DAG) G
- Let $I(p)$ denote the set of all conditional independencies that hold for the joint distribution p.
- A DAG G is an **I-map** (independence map) of a distribution p if $I(G) \subseteq I(p)$
  - A fully connected DAG G is an I-map for *any* distribution, since $I(G) = \emptyset \subseteq I(p)$ for all p
- G is a **minimal I-map** for p if the removal of even a single edge makes it not an I-map
  - A distribution may have several minimal I-maps
  - Each corresponds to a specific node-ordering
- G is a **perfect map** (P-map) for distribution p if $I(G) = I(p)$

Equivalent structures

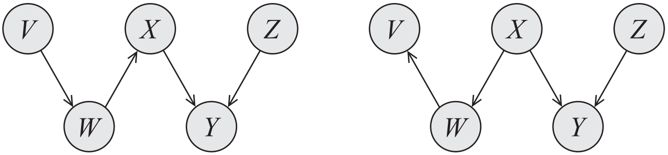
- Different Bayesian network structures can be **equivalent** in that they encode precisely the same conditional independence assertions (and thus the same distributions)
- Which of these are equivalent?



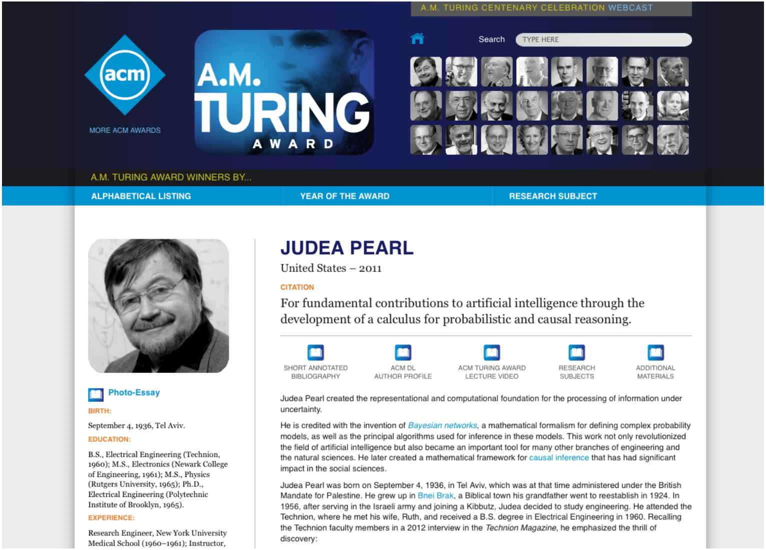
(a) (b) (c) (d)

Equivalent structures

- Different Bayesian network structures can be **equivalent** in that they encode precisely the same conditional independence assertions (and thus the same distributions)
- Are these equivalent?

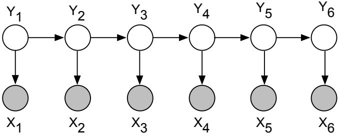


2011 Turing Award was for Bayesian networks



What are some frequently used graphical models?

Hidden Markov models

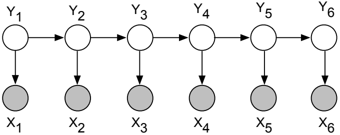


- Frequently used for speech recognition and part-of-speech tagging
- Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1 \mid y_1) \prod_{t=2}^{T} p(y_t \mid y_{t-1})p(x_t \mid y_t)$$

  - $p(y_1)$ is the distribution for the starting state
  - $p(y_t \mid y_{t-1})$ is the *transition* probability between any two states
  - $p(x_t \mid y_t)$ is the *emission* probability
- What are the conditional independencies here? For example, $Y_1 \perp \{Y_3, \ldots, Y_6\} \mid Y_2$

Hidden Markov models



- Joint distribution factors as:

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)p(x_1 \mid y_1) \prod_{t=2}^{T} p(y_t \mid y_{t-1}) p(x_t \mid y_t)$$

- A **homogeneous** HMM uses the same parameters (β and α below) for each transition and emission distribution (**parameter sharing**):

$$p(\mathbf{y}, \mathbf{x}) = p(y_1)\alpha_{x_1, y_1} \prod_{t=2}^{T} \beta_{y_t, y_{t-1}} \alpha_{x_t, y_t}$$

How many parameters need to be learned?

Mixture of Gaussians

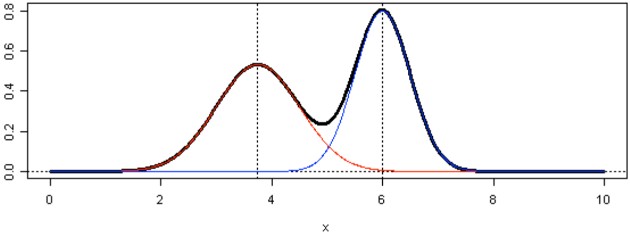
- The N-dim. multivariate normal distribution, $\mathcal{N}(\mu, \Sigma)$, has density:

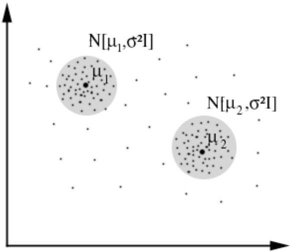
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\Big)$$

- Suppose we have k Gaussians given by μ_k and Σ_k, and a distribution θ over the numbers $1, \ldots, k$
- Mixture of Gaussians distribution $p(y, \mathbf{x})$ given by
  1. Sample $y \sim \theta$ (specifies which Gaussian to use)
  2. Sample $x \sim \mathcal{N}(\mu_y, \Sigma_y)$

Mixture of Gaussians

- The marginal distribution over **x** looks like:





Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics	religion	sports
president	hindu	baseball
obama	judiasm	soccer
washington	ethics	basketball
religion	buddhism	football
…	…	…

- Many applications in information retrieval, document summarization, and classification

New document

Words $w_1, \ldots, w_N$

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

- LDA is one of the simplest and most widely used topic models

Generative model for a document in LDA

1. Sample the document's **topic distribution** θ (aka topic vector)

$$\theta \sim \mathrm{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^{T}$ are fixed hyperparameters. Thus θ is a distribution over T topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1$ to N, sample the **topic** z_i of the i'th word

$$z_i | \theta \sim \theta$$

3. ... and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^{T}$ are the *topics* (a fixed collection of distributions on words)

Generative model for a document in LDA

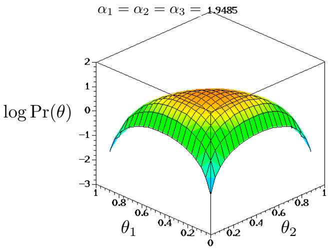
1. Sample the document's **topic distribution** θ (aka topic vector)

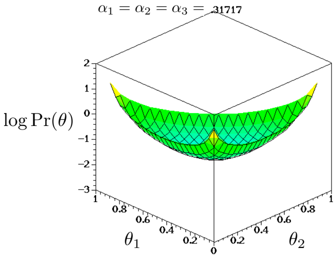
$$\theta \sim \mathrm{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^{T}$ are hyperparameters.The Dirichlet density, defined over $\Delta = \{\vec{\theta} \in \mathbb{R}^T : \forall t\ \theta_t \geq 0, \sum_{t=1}^{T} \theta_t = 1\}$, is:

$$p(\theta_1, \ldots, \theta_T) \propto \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$$

For example, for T=3 ($\theta_3 = 1 - \theta_1 - \theta_2$):





Generative model for a document in LDA

3. … and then sample the actual **word** w_i from the z_i'th topic

$$w_i | z_i \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^{T}$ are the *topics* (a fixed collection of distributions on words)

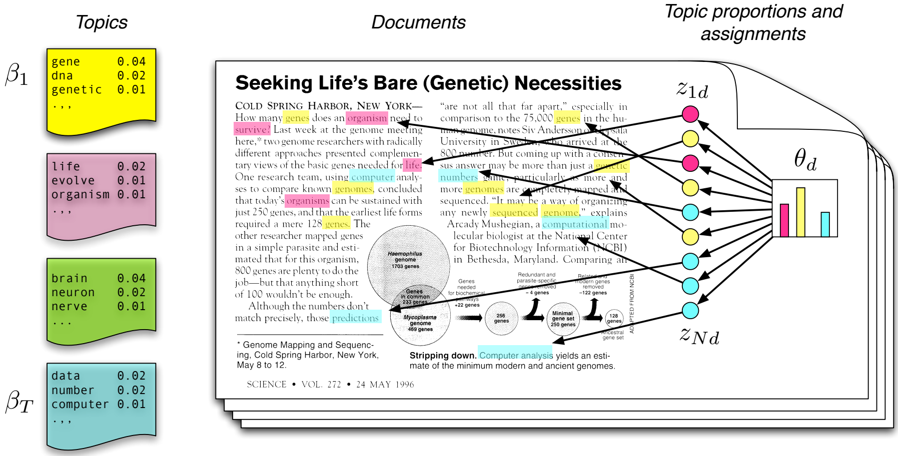
Documents

Topics

politics .0100	religion .0500	sports .0105
president .0095	hindu .0092	baseball .0100
obama .0090	judiasm .0080	soccer .0055
washington .0085	ethics .0075	basketball .0050
religion .0060	buddhism .0016	football .0045
…	…	…

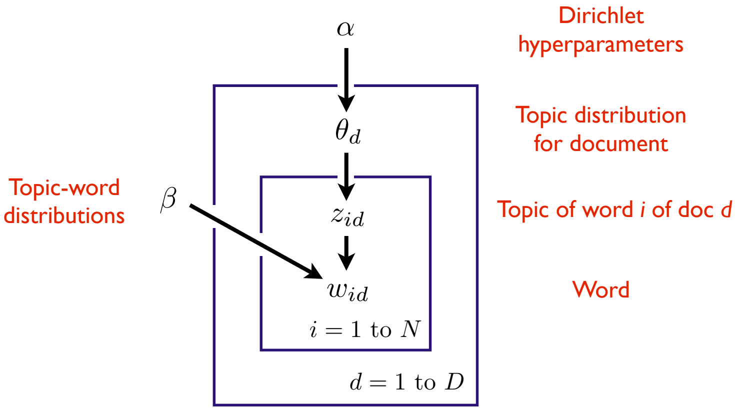
$$\beta_t = \{ p(w \mid z = t) \}$$

Example of using LDA



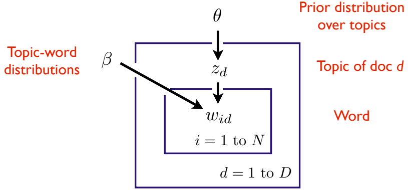
(Blei, *Introduction to Probabilistic Topic Models*, 2011)

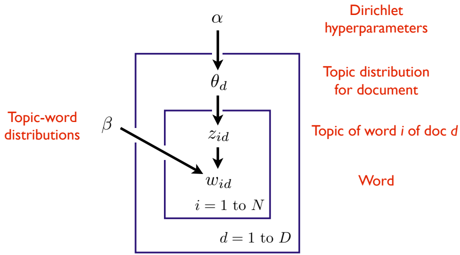
"Plate" notation for LDA model



Variables within a plate are replicated in a conditionally independent manner

Comparison of mixture and admixture models





- Model on left is a **mixture model**
  - Called *multinomial* naive Bayes (a word can appear multiple times)
  - Document is generated from a single topic
- Model on right (LDA) is an **admixture model**
  - Document is generated from a distribution over topics

Summary

- **Bayesian networks** given by (G, P) where P is specified as a set of local **conditional probability distributions** associated with G's nodes
- One interpretation of a BN is as a **generative model**, where variables are sampled in topological order
- Local and global independence properties identifiable via **d-separation** criteria
- Computing the probability of any assignment is obtained by multiplying CPDs
  - **Bayes' rule** is used to compute conditional probabilities
  - Marginalization or **inference** is often computationally difficult
- Examples (will show up again): **naive Bayes**, **hidden Markov models**, **latent Dirichlet allocation**