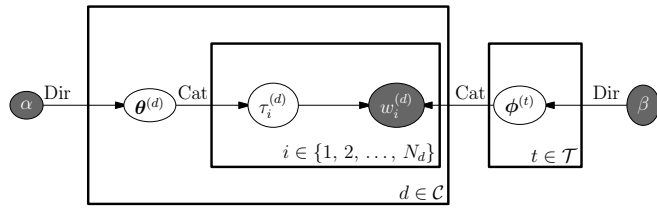## Advanced Machine Learning

### *Graphical Models*

Conditional Independence, Graphical models, LDA

Outline

Graphical Models

- If we want to build large probabilistic inference systems
  - ⋆ AI Doctor
  - ⋆ Fault diagnostic system for a computer

  we can describe this by introducing random variables, but it is helpful to graphically represent causal connections∎

- Graphical models allow us to do this∎

- It allows us to build a joint probability from which we can compute everything we want∎

Dependencies Between Variables

- In building a probabilistic model we want to know which random variables depend on each other directly and which don't∎

- Variables that don't will typically still be correlated∎

- If two random variables X and Y are correlated then
  - ⋆ X could affect Y
  - ⋆ Y could affect X
  - ⋆ X and Y could not influence each other, but both be affected by another random variable Z∎

Graphical Models

- **Bayesian Belief Networks** are a type of graphical models where we use a directed graphs to show causal relationships between random variables∎

- We could represent the three conditions described above by
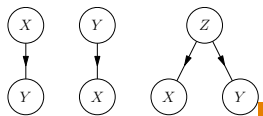
- We can use these graphical representations to work out how to efficiently average over latent variables∎

Statistical Independence

- Two random variables are statistically independent if

$$\mathbb{P}(X,Y) = \mathbb{P}(X)\mathbb{P}(Y)∎$$

- Equally this implies $\mathbb{P}(X|Y) = \mathbb{P}(X)$ and $\mathbb{P}(Y|X) = \mathbb{P}(Y)$∎

- Statistically independent variables are uncorrelated∎

- But statistical independence is often too powerful∎

Conditional Independence
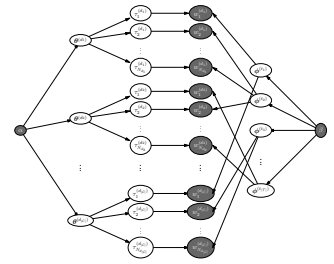
- A weaker notion is conditional independence

$$\mathbb{P}(X,Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)∎$$
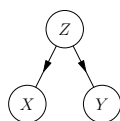
- Conditional independence implies that there is no direct causation∎

- But it doesn't imply zero correlation∎

- Conditional independence reduces computational complexity, e.g.

$$\mathbb{E}[XY] = \sum_{X,Y,Z} XY\mathbb{P}(X,Y,Z) = \sum_{Z} P(Z)\left(\sum_{X} XP(X|Z)\right)\left(\sum_{Y} YP(Y|Z)\right)∎$$

Outline
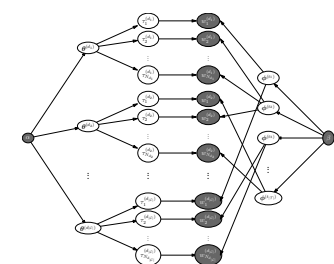
Let Them Eat Cakes

- I will go through a very simple example involving cakes

- It illustrates some simple principles

- In the subsidiary notes I present a very simple program for computing all the probabilities—I would encourage you to do this as it makes things much clearer

The Cake Scenario

- Abi and Ben both bake cakes and bring them into the coffee room

- Abi will bring in cakes 20% of the time: $\mathbb{P}(A = 1) = 0.2$

- Ben will bring in cakes 10% of the time: $\mathbb{P}(B = 1) = 0.1$

- 90% of the time if either Abi or Ben have put cakes in the coffee room there is some left when I enter
  $\mathbb{P}(C = 1|A = 1, B = 0) = \mathbb{P}(C = 1|A = 0, B = 1) = 0.9$

- If they both make cake then there is always cake left
  $\mathbb{P}(C = 1|A = 1, B = 1) = 1$

- If neither Abi or Ben has made cake there is still a 5% chance someone else has put cake in the coffee room
  $\mathbb{P}(C = 1|A = 0, B = 0) = 0.05$

Computing with Probabilities

- Other probabilities I can deduce, e.g.
  $\mathbb{P}(C = 0|A, B) = 1 - \mathbb{P}(C = 1|A, B)$

- I can depict the causal relationship as

$$A \longrightarrow C \longleftarrow B$$
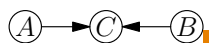
- The quantity that I really want is the joint probability

$$\mathbb{P}(A, B, C) = \mathbb{P}(C, B|A)\mathbb{P}(A)$$
$$= \mathbb{P}(C|A, B)\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(C|A, B)\mathbb{P}(B)\mathbb{P}(A)$$

- Because $\mathbb{P}(B|A) = \mathbb{P}(B)$

Computing Expectations

- By using the joint probability and summing over all unknown quantities, we can compute expectations of anything we are interested in

- These sums are often sped up using knowledge of conditional independence

- To compute the probability of and event $\mathcal{E}$ we introduce an indicator function $[\![\mathcal{E}]\!]$ which is equal to 1 if the event happens and 0 otherwise

$$\mathbb{P}(\mathcal{E}) = \mathbb{E}[[\![\mathcal{E}]\!]]$$

- If E is a random variable equal to 1 if event $\mathcal{E}$ happens and 0 otherwise then $E = [\![\mathcal{E}]\!]$

Are There Any Cakes Left?

- We can use our model to compute the probabilities of there being cakes in the coffee room

$$\mathbb{P}(C = 1) = \sum_{A, B, C \in \{0, 1\}} [\![C = 1]\!]\mathbb{P}(A, B, C)$$
$$= \sum_{A, B \in \{0, 1\}} \mathbb{P}(C = 1|A, B)\mathbb{P}(A)\mathbb{P}(B) = 0.29$$

- The probability that Abi baked a cake is just 0.2 and for Ben its 0.1 (which is what we assume at the start)

- The probability of them both baking on a particular day is 0.02

Making Observation

- Making observations changes probabilities

- In graphical models observed random variables are shaded

$$A \longrightarrow C \longleftarrow B$$

- The probabilities conditioned on C is given by

$$\mathbb{P}(A, B|C) = \frac{\mathbb{P}(A, B, C)}{\mathbb{P}(C)}$$

where

$$\mathbb{P}(C) = \sum_{A, B \in \{0, 1\}} \mathbb{P}(A, B, C)$$

Who Made Those Cakes?

- If we observe there are cakes

$$\mathbb{P}(A, B|C = 1) = \mathbb{P}(A, B, C = 1)/\mathbb{P}(C = 1)$$

- A straightforward if tedious calculation shows

$$\mathbb{P}(A = 1|C = 1) = 0.628, \quad \mathbb{P}(B = 1|C = 1) = 0.317$$
$$\mathbb{P}(A = 1, B = 1|C = 1) = 0.069$$

- Note $\mathbb{P}(A = 1, B = 1|C = 1) \neq \mathbb{P}(A = 1|C = 1)\mathbb{P}(B = 1|C = 1)$

- When we observe C then A and B are no longer independent

Elaborate Cakes

- We can elaborate on our cake model

- We suppose that Dave likes cakes so if there is a cake in the coffee room there is a 80% chance that I will see him eating a cake: $\mathbb{P}(D = 1|C = 1) = 0.8$

- Even if there are no cakes in the coffee room there is a 10% chance that Dave has bought his own cake:
  $\mathbb{P}(D = 1|C = 0) = 0.1$

- Eli also likes cakes: there is a 60% chance that I will see her eating cakes if there are cakes in the coffee room:
  $\mathbb{P}(E = 1|C = 1) = 0.6$

- But she never buys herself cakes $\mathbb{P}(E = 1|C = 0) = 0$

Elaborate Graphical Model

- We can depict this situation as
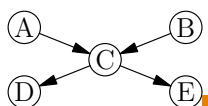
- This allows us to break down the joint probability

$$\mathbb{P}(A,B,C,D,E) = \mathbb{P}(C,D,E|A,B)\,\mathbb{P}(B)\,\mathbb{P}(A)$$
$$= \mathbb{P}(D|C)\,\mathbb{P}(E|C)\,\mathbb{P}(C|A,B)\,\mathbb{P}(B)\,\mathbb{P}(A)$$

- We use the conditional independence of D and E given C

Dependencies

- If we don't observe cakes then the probability of Dave and Eli eating cake are not independent

$$\mathbb{P}(D=1) = 0.303, \qquad \mathbb{P}(E=1) = 0.174$$
$$\mathbb{P}(D=1,E=1) = 0.1392$$

so $\mathbb{P}(D,E) \neq \mathbb{P}(D)\,\mathbb{P}(E)$

- This changes if we know there are cakes in the coffee room

$$\mathbb{P}(D=1|C=1) = 0.8 \qquad \mathbb{P}(E=1|C=1) = 0.6$$
$$\mathbb{P}(D=1,E=1|C=1) = 0.48$$

so $\mathbb{P}(D=1,E=1|C=1) = \mathbb{P}(D=1|C=1)\,\mathbb{P}(E=1|C=1)$

Observations and Independence

- Making observations changes the probabilities and in some case the dependencies of random variables on each other

$$A \not\!\perp\!\!\!\perp B$$
$$D \not\!\perp\!\!\!\perp E$$
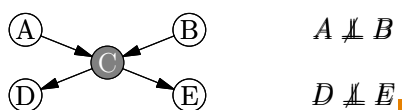
- There are rules to deduce the conditional independence from a graphical model given which variables have been observed—but these are details that you can look up if needed

Graphical Model Frameworks

- There are sophisticated frameworks for computing probabilities in Bayesian Belief Networks efficiently

- If our graph is a tree then we can evaluate probabilities efficiently

- When there are loops (so that a random variable both influences and is influenced by another random variables) then exact evaluation of expectations requires exhaustive summing over variables (which is often not tractable)

- There are various message passing algorithms designed to obtain approximations of expectations

Outline

1. Graphical Models
2. Cakes!
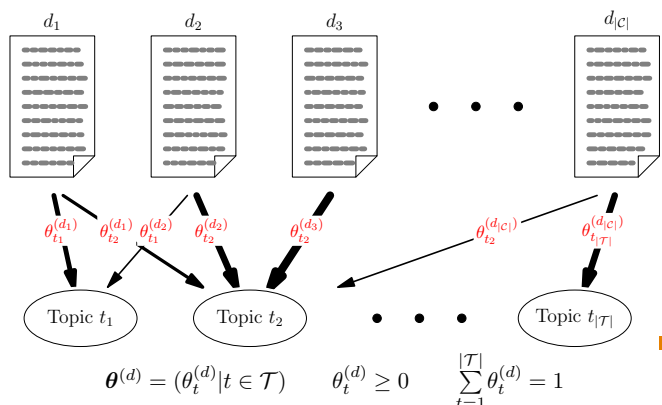3. **Latent Dirichlet Allocation**
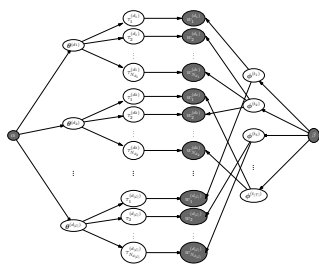
Model for Documents

- We consider a model for the words in a set of documents (we ignore word order)

- We consider a corpus $\mathcal{C} = \{d_i | i = 1, 2, \ldots |\mathcal{C}|\}$

- With documents consisting of words

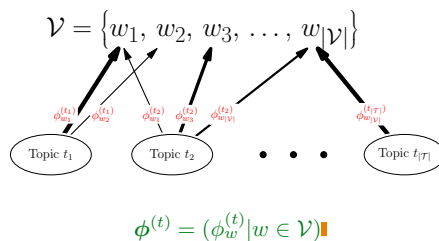$$d = \left(w_1^{(d)}, w_2^{(d)}, \ldots, w_{N_d}^{(d)} \right)$$

- We assume that there is a set of topics $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$

- We associate a probability, $\theta_t^{(d)}$, that a word in document d relates to a topic t

Documents and Topic

$$\boldsymbol{\theta}^{(d)} = (\theta_t^{(d)} | t \in \mathcal{T}) \qquad \theta_t^{(d)} \geq 0 \qquad \sum_{t=1}^{|\mathcal{T}|} \theta_t^{(d)} = 1$$

Words and Topic

- We associate a probability $\phi_w^{(t)}$ that a word, w, is related to a topic t

$$\mathcal{V} = \left\{ w_1, w_2, w_3, \ldots, w_{|\mathcal{V}|} \right\}$$

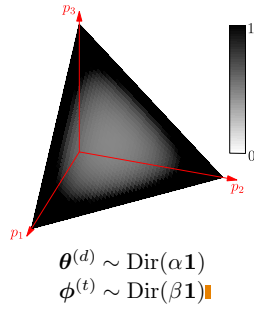$$\boldsymbol{\phi}^{(t)} = (\phi_w^{(t)} | w \in \mathcal{V})$$

Dirichlet Allocation

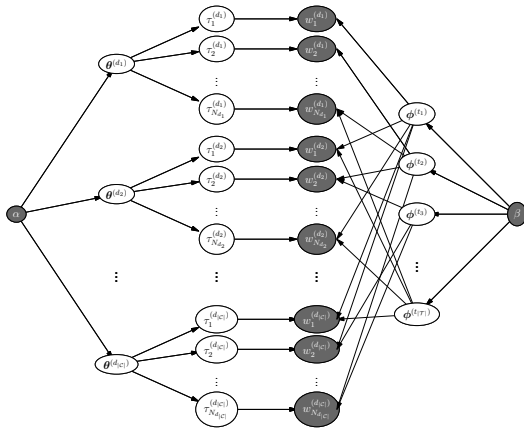- Most documents are predominantly about a few topics and most topic have a small number of words associated to them

- We can generate sparse vectors $\boldsymbol{\theta}^{(d)}$ and $\phi^{(t)}$ from a Dirichlet distribution with small parameters $\boldsymbol{\alpha}$

$$\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \Gamma\left(\sum_i \alpha_i\right)\prod_{i=1}^n \frac{p_i^{\alpha_i-1}}{\Gamma(\alpha_i)}$$

$$\boldsymbol{\theta}^{(d)} \sim \mathrm{Dir}(\alpha\mathbf{1})$$
$$\boldsymbol{\phi}^{(t)} \sim \mathrm{Dir}(\beta\mathbf{1})$$

- $\sum_i p_i = 1$

Generating Document

- To generate a document we choose a topic for each word and a word for each topic

$$\forall d \in \mathcal{C} \quad \boldsymbol{\theta}^{(d)} \sim \mathrm{Dir}(\alpha\mathbf{1})$$
$$\forall t \in \mathcal{T} \quad \boldsymbol{\phi}^{(t)} \sim \mathrm{Dir}(\beta\mathbf{1})$$
$$\forall d \in \mathcal{C} \ \wedge \ \forall i \in \{1,2,...,N_d\} \quad \tau_i^{(d)} \sim \mathrm{Cat}(\boldsymbol{\theta}^{(d)}), w_i^{(d)} \sim \mathrm{Cat}(\boldsymbol{\phi}^{(\tau_i^{(d)})})$$

- Where $\mathrm{Cat}(i|\boldsymbol{p}) = p_i$ is the categorical distribution (we choose one of a number of options)

- This model is known as **Latent Dirichlet Allocation**

LDA Graphical Model (version 1)

Plate Diagrams

- Drawing every random variable is tedious (and not really possible)

- A short-hand is to draw a box (plate) meaning repeat

- That is we generate vectors $\boldsymbol{\theta}^d$ from a Dirchelet distribution $\mathrm{Dir}(\boldsymbol{\theta}|\alpha\mathbf{1})$ for all documents in corpus $\mathcal{C}$

LDA Graphical Model (version 2)

- This is a lot more compact

- Personally, I find it hard to read, but you get used to it

Probabilistic Model

- The graphical Model is shorthand for the variables

$$\boldsymbol{W} = (\boldsymbol{w}^{(d)}|d \in \mathcal{C}) \quad \text{with} \quad \boldsymbol{w}^{(d)} = (w_1^{(d)}, w_2^{(d)}, ..., w_{N_d}^{(d)}), \quad \text{and} \quad w_i^{(d)} \in \mathcal{V}$$
$$\boldsymbol{T} = (\tau_i^{(d)}|d \in \mathcal{C} \ \wedge \ i \in \{1,2,...,N_d\}) \quad \text{with} \quad \tau_i^{(d)} \in \mathcal{T}$$
$$\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(d)}|d \in \mathcal{C}) \quad \text{with} \quad \boldsymbol{\theta}^{(d)} = (\theta_t^{(d)}|t \in \mathcal{T}) \in \Lambda^{|\mathcal{T}|}$$
$$\boldsymbol{\Phi} = (\boldsymbol{\phi}^{(t)}|t \in \mathcal{T}) \quad \text{with} \quad \boldsymbol{\phi}^{(t)} = (\phi_w^{(t)}|w \in \mathcal{V}) \in \Lambda^{|\mathcal{V}|}$$

- Distributed according to

$$\mathbb{P}(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\alpha, \beta) = \left(\prod_{t \in \mathcal{T}} \mathrm{Dir}\left(\boldsymbol{\phi}^{(t)}|\beta\mathbf{1}\right)\right)$$
$$\left(\prod_{d \in \mathcal{C}} \mathrm{Dir}\left(\boldsymbol{\theta}^{(d)}|\alpha\mathbf{1}\right)\prod_{i=1}^{N_d} \mathrm{Cat}\left(\tau_i^{(d)}|\boldsymbol{\theta}^{(d)}\right)\mathrm{Cat}\left(w_i^{(d)}|\boldsymbol{\phi}^{(\tau_i^{(d)})}\right)\right)$$

Finding Topics

- We are given the set of words $\boldsymbol{W}$ and don't really care about τ_i^d the topic associated with word i in document d

- But we are interested in the words associated with each topic $\phi^{(t_i)}$

- And the topics associated with each document $\boldsymbol{\theta}^{(d)}$

- To compute them we need to sample the probability distribution

- One way to do this is using Monte Carlo methods (see next lecture)

Summary

- Building probabilistic models is an intricate process

- Graphical models provide a representation showing the causal relationship between random variables

- This allows us to break down the joint probability of all the variables into conditional probabilities

- This is useful for building the model, but also can speed up evaluating expectations

- Making observations changes the probabilities of random variables

- It is possible to generate very rich models such as Latent Dirichlet Allocation (LDA)