

Introduction to Graphical Models

Mikaela Keller

IDIAP Research Institute
Martigny, Switzerland
mkeller[at]idiap.ch



July 10th, 2007

Overview

Inference

Comparison

Overview

Inference

Comparison

- ▶ **Directed Graphical Models**: Diagrams for representing probability distributions over a set of random variables.
- ▶ Graphical Models provide a graphical view on how to decompose the **joint probability** into a **product of factors** each depending only on a subset of variables.

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | pa_k),$$

- ▶ Graphical Models display the conditional independence of nodes.

- ▶ The **decomposition** of the joint probability provided by the Graphical Models is useful to:
 - ▶ **“Understand”** how the variables are connected to each others (conditional independence).
 - ▶ **Simplify** computations in inference and learning.
 - ▶ **Generalize** the understanding and simplification to any new model.

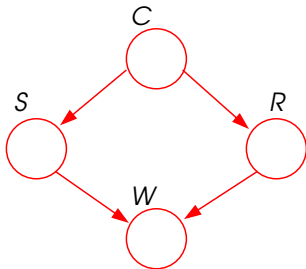
Overview

Inference

Comparison

Example

- ▶ Four random variables:
 - ▶ C : cloudy, $C \in \{0, 1\}$.
 - ▶ R : rain, $R \in \{0, 1\}$.
 - ▶ S : sprinkler on, $S \in \{0, 1\}$.
 - ▶ W : wet grass, $W \in \{0, 1\}$.



$$P(C, R, S, W) = P(C)P(R|C)P(S|C)P(W|R, S).$$

$$P(C = 1) = 0.5$$

$$P(S = 1|C = 1) = 0.1 \quad P(S = 1|C = 0) = 0.5$$

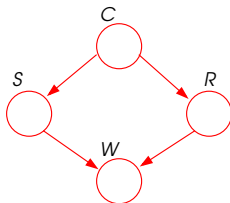
$$P(R = 1|C = 1) = 0.8, \quad P(R = 1|C = 0) = 0.2$$

$$P(W = 1|R = 1, S = 1) = 0.99$$

$$P(W = 1|R = 0, S = 1) = 0.9$$

$$P(W = 1|R = 1, S = 0) = 0.9$$

$$P(W = 1|R = 0, S = 0) = 0.0$$



$$P(S = 1|W = 1) = \frac{P(W = 1, S = 1)}{P(W = 1)}$$

$$= \frac{\sum_{R,C} P(W = 1, S = 1, R, C)}{\sum_{S,R,C} P(W = 1, S, R, C)}$$

$$= \frac{\sum_{R,C} P(C)P(S = 1|C)P(R|C)P(W = 1|R, S = 1)}{\sum_{S,R,C} P(C)P(S|C)P(R|C)P(W = 1|R, S)}$$

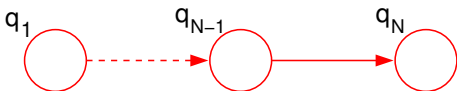
Inference in a chain

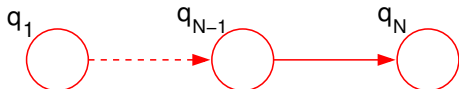
- ▶ Let $q_1^N = (q_1, \dots, q_N)$ be the random vector representing a sequence of states in a first order markov model. Each variable q_t takes value in $\{1, \dots, K\}$.
- ▶ The **first order markov** assumption states that for any time t :

$$p(q_t | q_1^{t-1}) = p(q_t | q_{t-1}).$$

- ▶ Thus, the **joint likelihood** of the sequence q_1^N :

$$p(q_1^N) = p(q_1) \prod_{t=1}^N p(q_t | q_{t-1}).$$





- ▶ Let assume that we are interested in the probability distribution of q_t .

$$p(q_t) = \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_N} p(q_1^N).$$

- ▶ Naive implementation: $\mathcal{O}(K^N)$.
- ▶ Let notice that the only factor in the joint distribution $p(q_1^N)$ depending on q_N is $p(q_N|q_{N-1})$, thus:

$$p(q_t) = \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \sum_{q_N} p(q_N|q_{N-1}).$$

Inference in a chain

- ▶ The decomposition of $p(q_1^N)$ into factors allow us to exploit the **distributivity** of the multiplication over addition:

$$ab + ac = a(b + c).$$

- ▶ Let proceed with the marginalisation:

$$\begin{aligned}
 p(q_t) &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \sum_{q_N} p(q_N | q_{N-1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \beta(q_{N-1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} p(q_1^{t+1}) \beta(q_{t+1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} p(q_1^t) \beta(q_t).
 \end{aligned}$$

Inference in a chain

- ▶ The decomposition of $p(q_1^N)$ into factors allow us to exploit the **distributivity** of the multiplication over addition:

$$ab + ac = a(b + c).$$

- ▶ Let proceed with the marginalisation:

$$\begin{aligned}
 p(q_t) &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \sum_{q_N} p(q_N | q_{N-1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \beta(q_{N-1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} p(q_1^{t+1}) \beta(q_{t+1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} p(q_1^t) \beta(q_t).
 \end{aligned}$$

Inference in a chain

- ▶ The decomposition of $p(q_1^N)$ into factors allow us to exploit the **distributivity** of the multiplication over addition:

$$ab + ac = a(b + c).$$

- ▶ Let proceed with the marginalisation:

$$\begin{aligned}
 p(q_t) &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \sum_{q_N} p(q_N | q_{N-1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_{N-1}} p(q_1^{N-1}) \beta(q_{N-1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} p(q_1^{t+1}) \beta(q_{t+1}) \\
 &= \sum_{q_1} \cdots \sum_{q_{t-1}} p(q_1^t) \beta(q_t).
 \end{aligned}$$

- ▶ We can proceed with a similar procedure beginning from the state q_1 :

$$\begin{aligned} p(q_t) &= \sum_{q_1} \cdots \sum_{q_{t-1}} p(q_1^t) \beta(q_t) \\ &= \sum_{q_{t-1}} \cdots \sum_{q_1} p(q_1) p(q_2|q_1) p(q_2^t) \beta(q_t) \\ &= \sum_{q_{t-1}} p(q_t|q_{t-1}) \alpha(q_{t-1}) \beta(q_t) \\ &= \alpha(q_t) \beta(q_t). \end{aligned}$$

- ▶ We can proceed with a similar procedure beginning from the state q_1 :

$$\begin{aligned} p(q_t) &= \sum_{q_1} \cdots \sum_{q_{t-1}} p(q_1^t) \beta(q_t) \\ &= \sum_{q_{t-1}} \cdots \sum_{q_1} p(q_1) p(q_2|q_1) p(q_2^t) \beta(q_t) \\ &= \sum_{q_{t-1}} p(q_t|q_{t-1}) \alpha(q_{t-1}) \beta(q_t) \\ &= \alpha(q_t) \beta(q_t). \end{aligned}$$

$$\begin{aligned}
 p(q_t) &= \sum_{q_1} \cdots \sum_{q_{t-1}} \sum_{q_{t+1}} \cdots \sum_{q_N} p(q_1^N) \\
 &= \alpha(q_t)\beta(q_t)
 \end{aligned}$$

- ▶ Complexity: $\mathcal{O}(N^2K^2)$
- ▶ Note that α and β are computed recursively:

$$\alpha(q_t) = \sum_{q_{t-1}} p(q_t|q_{t-1})\alpha(q_{t-1}), \quad \alpha(q_2) = \sum_{q_1} p(q_1)p(q_2|q_1).$$

$$\beta(q_t) = \sum_{q_{t+1}} p(q_{t+1}|q_t)\beta(q_{t+1}), \quad \beta(q_N) = \sum_{q_N} p(q_N|q_{N-1}).$$

- ▶ Note that we can also compute the joint probability of two consecutive states as:

$$p(q_{t-1}, q_t) = \alpha(q_{t-1})p(q_t|q_{t-1})\beta(q_t).$$

- ▶ Depending on the probability distributions chosen for the conditional probabilities, exact inference is not always possible.
- ⇒ **Approximate Inference:**
 - ▶ Variational methods.
 - ▶ Sampling methods.
 - ▶ etc.
- ▶ Learning:
 - ▶ when **all** random variables are **observed**: Maximum likelihood estimate, etc.
 - ▶ when **only some** random variables are **observed**: Expectation Maximization, Variational EM, etc.

Overview

Inference

Comparison

Text Representation Setting

- ▶ We are given a database of N documents.
- ▶ After some preprocessing steps (stemming, stopping), we have a dictionary of M different words.
- ▶ Find a good text representation for these words/documents...
- ▶ Typical applications:
 - ▶ Document retrieval
 - ▶ Text categorization
 - ▶ Text filtering
 - ▶ Text entailment
 - ▶ etc

The Vector Space Model

- ▶ Represent document d as a **bag-of-words**:

$$d = (\alpha_1, \alpha_2, \dots, \alpha_M)$$

- ▶ α_m is function of the frequency of m^{th} word (w_m) in d and in the database.
- ▶ In this representation:
 - “A boat in the sea” \neq “A ship in the ocean”.
 - “Surfing a wave” \approx “Surfing the Internet”.

The Vector Space Model

- ▶ Represent document d as a **bag-of-words**:

$$d = (\alpha_1, \alpha_2, \dots, \alpha_M)$$

- ▶ α_m is function of the frequency of m^{th} word (w_m) in d and in the database.
- ▶ In this representation:
 - “A boat in the sea”* \neq *“A ship in the ocean”*.
 - “Surfing a wave”* \approx *“Surfing the Internet”*.

- We are looking for a **representation** $\phi(d)$, where the mapping $\phi(\cdot)$ takes into account knowledge about the links between words.
- Gain this knowledge from the huge amount of **available documents** in digital format. *Ex.*: Internet.

Graphical Models for Text Representation

- ▶ Several probabilistic models of the relationships between words and documents have been proposed, among which:
- ▶ Probabilistic Latent Semantic Analysis (**PLSA**, Hofmann, 2001),
- ▶ Latent Dirichlet Allocation (**LDA**, Blei *et al.*, 2003) and
- ▶ Theme Topic Mixture Model (**TTMM**, Keller and Bengio, 2004).

Common main assumption

The distribution of words in a document is **independent** of the document given a hidden variable K , which partitions the **word space** into \mathcal{K} **Topics**.

Graphical Models for Text Representation

- ▶ Several probabilistic models of the relationships between words and documents have been proposed, among which:
- ▶ Probabilistic Latent Semantic Analysis (**PLSA**, Hofmann, 2001),
- ▶ Latent Dirichlet Allocation (**LDA**, Blei *et al.*, 2003) and
- ▶ Theme Topic Mixture Model (**TTMM**, Keller and Bengio, 2004).

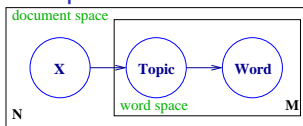
Common main assumption

- ▶ $P(w_m|X) = \sum_{k=1}^{\mathcal{K}} P(K = k|X)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ X is a generic random variable in the document space.

Graphical Models for Text Representation

- ▶ Several probabilistic models of the relationships between words and documents have been proposed, among which:
- ▶ Probabilistic Latent Semantic Analysis (**PLSA**, Hofmann, 2001),
- ▶ Latent Dirichlet Allocation (**LDA**, Blei *et al.*, 2003) and
- ▶ Theme Topic Mixture Model (**TTMM**, Keller and Bengio, 2004).

Common main assumption



Definitions

PLSA (Hofmann, 2001)

PLSA models the **word-document** co-occurrences, for each document d_δ and each word w_m :

$$P(d_\delta, w_m) = P(\delta)P(w_m|d_\delta).$$

Common main assumption

- ▶ $P(w_m|d_\delta) = \sum_{k=1}^{\mathcal{K}} P(K = k|d_\delta)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ δ values $\in \{1, \dots, N\}$, $P(\delta)$ proportional to the document length.

Definitions

PLSA (Hofmann, 2001)

PLSA models the **word-document** co-occurrences, for each document d_δ and each word w_m :

$$P(d_\delta, w_m) = P(\delta) \sum_{k=1}^{\mathcal{K}} P(K = k | d_\delta) P(w_m | K = k).$$

Common main assumption

- ▶ $P(w_m | d_\delta) = \sum_{k=1}^{\mathcal{K}} P(K = k | d_\delta) P(w_m | K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ δ values $\in \{1, \dots, N\}$, $P(\delta)$ proportional to the document length.

Definitions

TTMM (Keller and Bengio, 2004)

In TTMM the **document space** is partitioned into \mathcal{H} **Themes**:

$$P(d) = \sum_{h=1}^{\mathcal{H}} P(H = h)P(d|H = h)$$

Common main assumption

- ▶ $P(w_m|H = h) = \sum_{k=1}^{\mathcal{K}} P(K = k|H = h)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ Topics K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ Themes H values $\in \{1, \dots, \mathcal{H}\}$.

Definitions

TTMM (Keller and Bengio, 2004)

In TTMM the **document space** is partitioned into \mathcal{H} **Themes**:

$$P(d) = \sum_{h=1}^{\mathcal{H}} P(H = h)P(d|H = h)$$

Common main assumption

- ▶ $P(w_m|H = h) = \sum_{k=1}^{\mathcal{K}} P(K = k|H = h)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ Topics K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ Themes H values $\in \{1, \dots, \mathcal{H}\}$.

Definitions

TTMM (Keller and Bengio, 2004)

In TTMM the **document space** is partitioned into \mathcal{H} **Themes**:

$$P(d) = \sum_{h=1}^{\mathcal{H}} P(H = h) \prod_{w_m \in d} [P(w_m | H = h)]^{\text{tf}_m(d)}$$

Common main assumption

- ▶ $P(w_m | H = h) = \sum_{k=1}^{\mathcal{K}} P(K = k | H = h) P(w_m | K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ Topics K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ Themes H values $\in \{1, \dots, \mathcal{H}\}$.

Definitions

TTMM (Keller and Bengio, 2004)

In TTMM the **document space** is partitioned into \mathcal{H} **Themes**:

$$P(d) = \sum_{h=1}^{\mathcal{H}} P(H = h) \prod_{w_m \in d} [P(w_m | H = h)]^{\text{tf}_m(d)}$$

Common main assumption

- ▶ $P(w_m | H = h) = \sum_{k=1}^{\mathcal{K}} P(K = k | H = h) P(w_m | K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ Topics K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ Themes H values $\in \{1, \dots, \mathcal{H}\}$.

Definitions

TTMM (Keller and Bengio, 2004)

In TTMM the **document space** is partitioned into \mathcal{H} **Themes**:

$$P(d) = \sum_{h=1}^{\mathcal{H}} P(H = h) \prod_{w_m \in d} \left[\sum_{k=1}^{\mathcal{K}} P(K = k | H = h) P(w_m | K = k) \right]^{tf_m(d)}$$

Common main assumption

- ▶ $P(w_m | H = h) = \sum_{k=1}^{\mathcal{K}} P(K = k | H = h) P(w_m | K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ Topics K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ Themes H values $\in \{1, \dots, \mathcal{H}\}$.

Definitions

LDA (Blei *et al.*, 2003)

In LDA a document is sampled from a **random mixture** over latent topics:

$$P(d) = \int P(\tau|\zeta)P(d|\tau)d\tau$$

Common main assumption

- ▶ $P(w_m|\tau) = \sum_{k=1}^{\mathcal{K}} P(K = k|\tau)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ $\tau = (\tau_1, \dots, \tau_{\mathcal{K}})$, $\tau_k = P(K = k|\tau)$
 $\tau \sim \text{Dirichlet}(\zeta)$, $\zeta \in \mathbb{R}^{+\mathcal{K}}$.

Definitions

LDA (Blei *et al.*, 2003)

In LDA a document is sampled from a **random mixture** over latent topics:

$$P(d) = \int P(\tau|\zeta)P(d|\tau)d\tau$$

Common main assumption

- ▶ $P(w_m|\tau) = \sum_{k=1}^{\mathcal{K}} P(K = k|\tau)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ $\tau = (\tau_1, \dots, \tau_{\mathcal{K}})$, $\tau_k = P(K = k|\tau)$
 $\tau \sim \text{Dirichlet}(\zeta)$, $\zeta \in \mathbb{R}^{+\mathcal{K}}$.

Definitions

LDA (Blei *et al.*, 2003)

In LDA a document is sampled from a **random mixture** over latent topics:

$$P(d) = \int P(\tau|\zeta) \prod_{w_m \in d} P(w_m|\tau) d\tau$$

Common main assumption

- ▶ $P(w_m|\tau) = \sum_{k=1}^{\mathcal{K}} P(K = k|\tau)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ $\tau = (\tau_1, \dots, \tau_{\mathcal{K}})$, $\tau_k = P(K = k|\tau)$
 $\tau \sim \text{Dirichlet}(\zeta)$, $\zeta \in \mathbb{R}^{+\mathcal{K}}$.

Definitions

LDA (Blei *et al.*, 2003)

In LDA a document is sampled from a **random mixture** over latent topics:

$$P(d) = \int P(\tau|\zeta) \prod_{w_m \in d} P(w_m|\tau) d\tau$$

Common main assumption

- ▶ $P(w_m|\tau) = \sum_{k=1}^{\mathcal{K}} P(K = k|\tau)P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ $\tau = (\tau_1, \dots, \tau_{\mathcal{K}})$, $\tau_k = P(K = k|\tau)$
 $\tau \sim \text{Dirichlet}(\zeta)$, $\zeta \in \mathbb{R}^{+\mathcal{K}}$.

Definitions

LDA (Blei *et al.*, 2003)

In LDA a document is sampled from a **random mixture** over latent topics:

$$P(d) = \int P(\tau|\zeta) \prod_{w_m \in d} \left[\sum_{k=1}^{\mathcal{K}} P(K = k|\tau) P(w_m|K = k) \right]^{tf_m(d)} d\tau$$

Common main assumption

- ▶ $P(w_m|\tau) = \sum_{k=1}^{\mathcal{K}} P(K = k|\tau) P(w_m|K = k)$,
 - ▶ w_m values $\in \{0, 1\}$ model the presence of absence of word m ,
 - ▶ K values $\in \{1, \dots, \mathcal{K}\}$,
 - ▶ $\tau = (\tau_1, \dots, \tau_{\mathcal{K}})$, $\tau_k = P(K = k|\tau)$
 $\tau \sim \text{Dirichlet}(\zeta)$, $\zeta \in \mathbb{R}^{+\mathcal{K}}$.

Comparison

- ▶ PLSA (Hofmann, 2001)

$$P(d_\delta, w_m) = P(\delta) \sum_{k=1}^{\mathcal{K}} P(K = k | d_\delta) P(w_m | K = k).$$

- ▶ TTMM (Keller and Bengio, 2004)

$$P(d) = \sum_{h=1}^{\mathcal{H}} P(H = h) \prod_{w_m \in d} \left[\sum_{k=1}^{\mathcal{K}} P(K = k | H = h) P(w_m | K = k) \right]^{\text{tf}_m(d)}.$$

- ▶ LDA (Blei *et al.*, 2003)

$$P(d) = \int P(\tau) \prod_{w_m \in d} \left[\sum_{k=1}^{\mathcal{K}} P(K = k | \tau) P(w_m | K = k) \right]^{\text{tf}_m(d)} d\tau.$$

► Inference

PLSA	LDA	TTMM
exact - EM	approx. - variational EM	exact - EM

► Complexity

PLSA	LDA	TTMM
$\mathcal{O}(NK\bar{n}[\bar{n} + M])$	$\mathcal{O}(NK\bar{n}[\bar{n} + M])$	$\mathcal{O}(NK\mathcal{H}[\bar{n} + M])$

► Number of parameters

PLSA	LDA	TTMM
$\mathcal{K}N + \mathcal{K}M$	$\mathcal{K} + \mathcal{K}M$	$(1 + \mathcal{K})\mathcal{H} + \mathcal{K}M$

- \mathcal{K} : Number of topics.
- \mathcal{H} : Number of themes.
- M : Number of words.

- N : Number of documents.
- \bar{n} : average length of documents.

- ▶ Sources of inspiration:
- ▶ Christopher M. Bishop's book "Pattern Recognition and Machine Learning" .
- ▶ Kevin Murphy Introduction to Graphical Models (on the web).
- ▶ My PhD thesis...