Latent Variable Models for Text Documents

Bioinformatics
Course
Supplement

Probabilistic Latent Semantic Analysis (1)

- Proposed by T. Hofmann.
- Based on *aspect model*.
  - A latent variable model for general co-occurrence data.
  - Associates an unobserved class variable with each observation.
    - Latent $z \in Z = \{z_1, \mathrm{K}, z_k\}$
    - Observed

$$w \in W = \{w_1, \mathrm{K}, w_M\}$$

$$d \in D = \{d_1, \mathrm{K}, d_N\}$$

Probabilistic Latent Semantic Analysis (2)

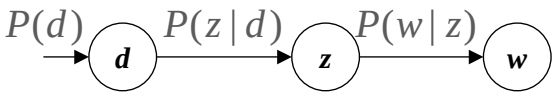
- A view as a generative model
  - Select a document d with probability $P(d)$.
  - Pick a latent class z with probability $P(z|d)$.
  - Generate a word w with probability $P(w|z)$.
- Probability model

$$P(d,w) = P(d)P(w|d)$$
$$= P(d)\sum_{z\in Z} P(w|z)P(z|d)$$
$$= \sum_{z\in Z} P(z)P(w|z)P(d|z)$$

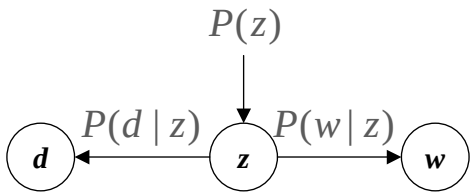
$$P(d)P(z|d) = P(z)P(d|z)$$

Probabilistic Latent Semantic Analysis (3)

- Graphical Representation
  - Asymmetric parameterization



  - Symmetric parameterization



Probabilistic Latent Semantic Analysis (4)

- A view as a statistical *mixture model*
  - Based on two independence assumption
    - Observation pairs (*d*, *w*) are generated independently.
    - Conditional independence
      - Conditioned on the latent class *z*, words *w* are generated independently of the specific document identity *d*.
  - $P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)$
    - P(*w*|*d*) are obtained by a convex combination of the aspects or factors P(*w*|*z*) with weights P(*z*|*d*).

Probabilistic Latent Semantic Analysis (5)

- Model Fitting
  - Objective Function : $L = \sum_{d \in D} \sum_{w \in W} n(d,w) \log P(d,w)$
  - EM Algorithm
    - E-Step $P(z|d,w) = \frac{P(z,d,w)}{P(d,w)} = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}$
    - M-Step

$$P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')} \qquad P(d|z) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{d',w} n(d',w)P(z|d',w)}$$

$$P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w) \qquad R = \sum_{d,w} n(d,w)$$

Probabilistic Latent Semantic Analysis (6)

- Experiment on text documents (T. Hofmann, SIGIR'99)
  - TDT-1 collection
    - 15,862 documents of broadcast news stories.

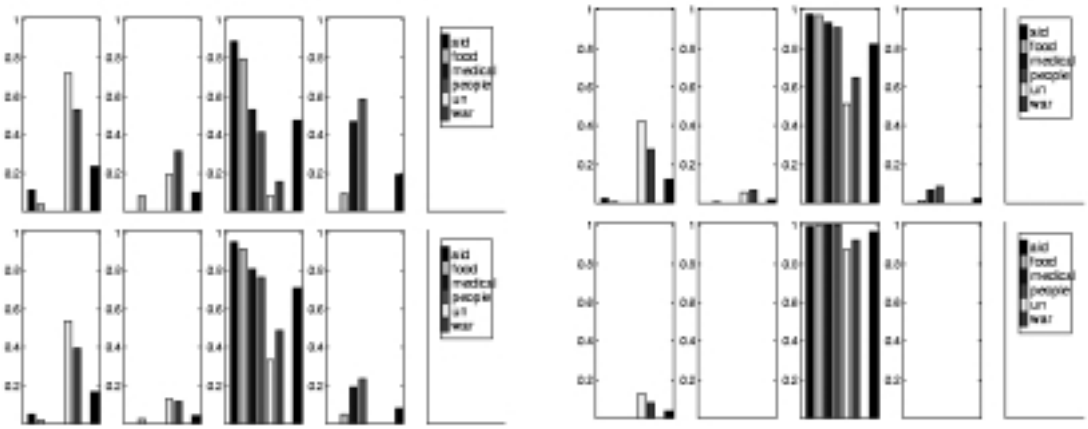
"plane"	"space shuttle"	"family"	"Hollywood"
plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	astronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
board	satellite	friends	entertainment
airline	earth	cnn	star

Probabilistic Latent Semantic Analysis (7)

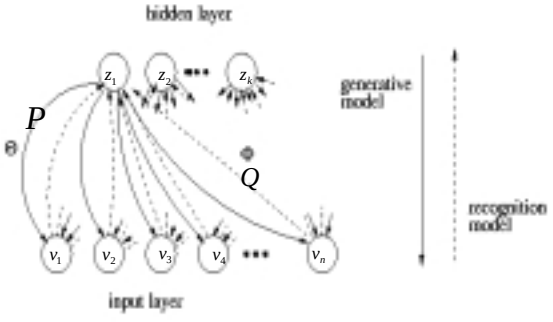
"Bosnia"	"Iraq"	"Rwanda"	"Kobe"
un	iraq	refugees	building
bosnian	iraqi	aid	city
serbs	sanctions	rwanda	people
bosnia	kuwait	relief	rescue
serb	un	people	buildings
sarajevo	council	camps	workers
nato	gulf	zaire	kobe
peacekeepers	saddam	camp	victims
nations	baghdad	food	area
peace	hussein	rwandan	earthquake

Probabilistic Latent Semantic Analysis (8)

- Folding-in query



Helmholtz Machine (1)



- Each node is probabilistic binary node.
- Given values of the parent nodes, the nodes in the same layer are conditionally independent.

- **Generative network**

$$\log p(\mathbf{v}\mid\mathbf{z}) = \sum_{i=1}^{n}(v_i \log(p(v_i=1)) + (1-v_i)(1-\log(1-p(v_i=1)))$$

$$p(v_i=1) = \frac{1}{1+\exp(-b_i - \sum_j z_j\theta_{ji})}$$

$$p(z_i=1) = \frac{1}{1+\exp(-b_i)}$$

- **Recognition network**

To ease learning and inference

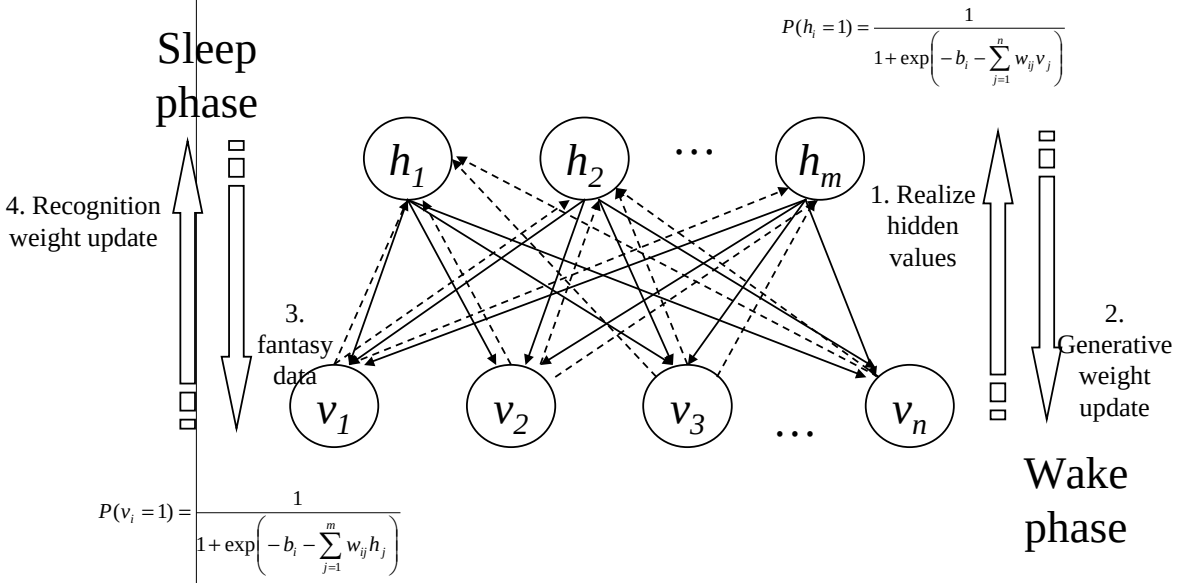
$$\log q(\mathbf{z}\mid\mathbf{v}) = \sum_{i=1}^{n}(z_i \log(q(z_i=1)) + (1-z_i)(1-\log(1-q(z_i=1)))$$

$$q(z_i=1) = \frac{1}{1+\exp(-b_i - \sum_j v_j\theta_{ji})}$$

- Both networks are estimated using ***wake-sleep*** algorithm.

Helmholtz Machine (2)

- Wake-sleep algorithm



Helmholtz Machine (3)

- ***Wake*** phase

1. An data object is clamped on input nodes.
2. Probabilistically determine the values (***on***(1)/***off***(0)) of the hidden nodes through recognition network.

$$P(h_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{n} w_{ij} v_j\right)}$$

3. While estimating the probability of being '***on***' of each node in top-down way, update the weights of generative network.

$$w_{uv}^{new} = w_{uv}^{old} + \Delta w_{uv}$$
$$\Delta w_{uv} = \gamma\, s_u (s_v - p(s_v = 1))$$

- ***Sleep*** phase

1. Probabilistically determine the values of nodes from topmost layer to input layer through generative network. That is, a ***fantasy data*** is generated.

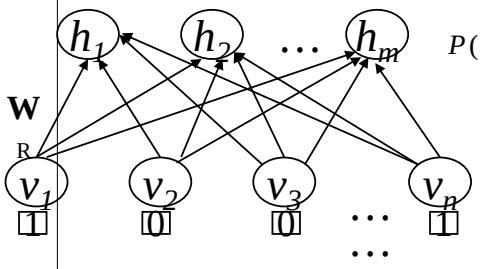
$$P(v_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{m} w_{ij} h_j\right)}$$

2. While estimating the probability of being 'on' of each node in bottom-up way through recognition network, update the weights of recognition network.

$$w_{vu}^{new} = w_{vu}^{old} + \Delta w_{vu}$$
$$\Delta w_{vu} = \gamma\, s_v (s_u - p(s_u = 1))$$

Helmholtz Machine (4)

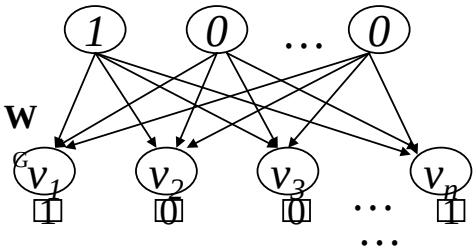
- Wake Phase



$$P(h_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{n} w_{ij} v_j\right)}$$

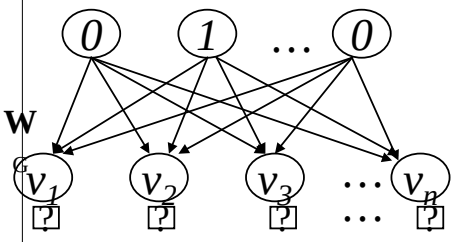
$$w_{uv}^{new} = w_{uv}^{old} + \Delta w_{uv}$$

$$\Delta w_{uv} = \gamma\, s_u (s_v - p(s_v = 1))$$



Helmholtz Machine (5)

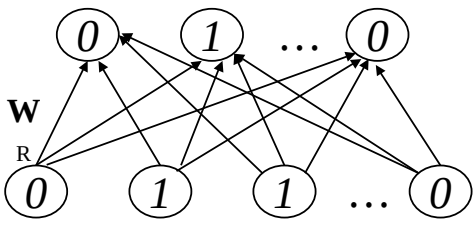
- Sleep phase



$$P(v_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{m} w_{ij} h_j\right)}$$

$$w_{vu}^{new} = w_{vu}^{old} + \Delta w_{vu}$$

$$\Delta w_{vu} = \gamma\, s_v (s_u - p(s_u = 1))$$

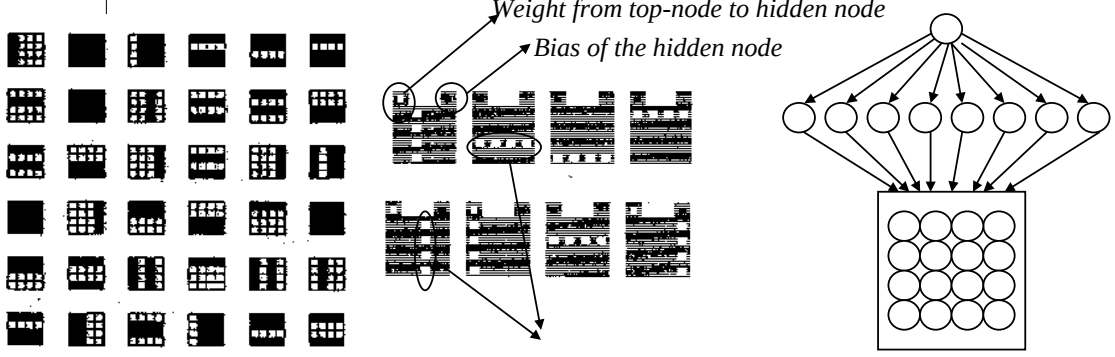


Helmholtz Machine (6)

- Experiments on “bar image” analysis
  - Data set : 4×4 binary images of which the number is 2,000,000.
    - Select an orientation (horizontal or vertical) with equal probability.
    - Randomly instantiate each of the four possible bars with that orientation with probability 0.5.
    - *All-on* images are ignored.

Helmholtz Machine (7)

- Model Fitting
  - A *Helmholtz machine* is fitted using *wake-sleep* algorithm.
    - The network has three layers of binary variables (1 ×8 ×16).
    - The variables in adjacent layers are fully-connected.
    - The conditional distributions are modeled using logistic regression.



Helmholtz Machine (8)

- Experiments on topic word extraction
  - Subset of TREC-8 adhoc data

Topics	*20 most probable words according to* $\boldsymbol{P(w|z)}$
434	estonia, economic, foreign, trade, estonia, year, states, government, country, state, russia, tallinn, union, european, economy, baltic, market, investment, goods, inflation
439	company, process, technology, time, high, development, market, production, developed, research, make, patent, work, cost, institute, chemical, group, materials, industrial, products
450	king, jordan, peace, israel, jordanian, israeli, talks, arab, minister, east, al, palestinian, middle, president, process, agreement, husayn, majesty, plo, palestine, washinton
401	germany, german, asylum, party, government, foreigners, political, wing, minister, social, foreign, state, union, interior, turkish, europe, workers, seekers, extremists, nazi

Helmholtz Machine (9)

- Subset of TDT-2 collection

Skating	Ice hockey	General
skating	**team**	**won**
figure	**hockey**	**olympics**
program	**ice**	**winter**
olympic	**canada**	**games**
champion	**game**	**nagano**
skate	**olympic**	**world**
short	**players**	**race**
lipinski	**goal**	**medal**
judges	**tournament**	**gold**
tara	**league**	**silver**

Financial crisis	Israel & Palestine	Nuclear race	Winter Olympics
asia	**israel**	**india**	**won**
economy	**palestinian**	**nuclear**	**olympics**
percent	**peace**	**pakistan**	**winter**
financial	**netanyahu**	**tests**	**games**
market	**process**	**arms**	**nagano**
crisis	**arafat**	**RACE**	**world**
currency	**BANK**	**test**	**RACE**
dollar	**benjamin**	**weapons**	**medal**
japan	**yasser**	**indian**	**gold**
…	**…**	**…**	**silver**
BANK	**talk**	**security**	**…**
…	**…**	**…**	**athletes**