# Johdatus variaatiomenetelmiin
# Osa II

**Antti Honkela, 2001-10-10**

Perustuu lähinnä lukuun Jordan, Ghahramani, Jaakkola ja Saul: *An Introduction to Variational Methods*

## Sisältö

- Blokkimenetelmä
- Esimerkkejä
- Yhteydet muihin menetelmiin
- Sekoitejakauman käyttö approksimaationa
  - Jaakkola ja Jordan: *Improving the mean field approximation via the use of mixture distributions*

## Blokkimenetelmä

- Muodostetaan approksimaatio kerralla joukolle noodeja
- Etsitään graafista helposti käsiteltävä osagraafi
  - Oletetaan muuttujille lisää riippumattomuuksia
  - Ts. approksimoidaan posteriorijakaumaa $P(H|E)$ yksinkertaisemmalla parametrisella jakaumalla $Q(H|E,\lambda)$, missä $\lambda$ kuvaa variaatioparametreja
- Tätä vastaava ongelma ratkaistaan normaalisti

## Blokkimenetelmä, jatkoa

- Variaatioparametrit $\lambda$ optimoidaan minimoimalla *Kullback-Leibler -divergenssi*

$$D(Q||P) = \sum_{\{H\}} Q(H|E,\lambda) \log \frac{Q(H|E,\lambda)}{P(H|E)} \tag{1}$$

- Optimia $\lambda^*$ vastaava jakauma $Q(H|E,\lambda^*)$ paras approksimaatio posteriorijakaumalle $P(H|E)$
- Kullback-Leibler -divergenssi tavallaan etäisyysmitta jakaumille
  - $D(Q||P) \geq 0$
  - $D(Q||P) = 0 \;\Leftrightarrow\; Q \equiv P$

## Blokkimenetelmä, jatkoa

- Todellisuudessa maksimoidaan lauseketta

$$
\begin{aligned}
C &= \sum_{\{H\}} Q(H|E,\lambda) \log \frac{P(H,E)}{Q(H|E,\lambda)} \\
&= \log P(E) - D(Q||P) \\
&\leq \log P(E)
\end{aligned}
\tag{2}
$$

$\Rightarrow$ alaraja evidenssille $P(E)$

- Optimi $\lambda^*$ antaa tiukimman rajan

## Blokkimenetelmä, jatkoa

- Alaraja on helppo todistaa Jensenin epäyhtälön avulla:

$$
\begin{aligned}
\log P(E) &= \log \sum_{\{H\}} P(H,E) \\
&= \log \sum_{\{H\}} Q(H|E) \frac{P(H,E)}{Q(H|E)} \\
&\geq \sum_{\{H\}} Q(H|E) \log \frac{P(H,E)}{Q(H|E)}
\end{aligned}
\tag{3}
$$

- Sama raja voitaisiin johtaa myös konveksista dualiteetista

## Esimerkki: Uskomusverkko

- Neuroverkkomainen graafinen malli, binääriset solmut
- Solmun aktivaation todennäköisyys:

$$
P(S_i = 1|S_{\pi(i)}) = \frac{1}{1 + \exp(-\sum_{j\in\pi(i)} \theta_{ij} S_j - \theta_{i0})} \tag{4}
$$

- Jakauma solmun tiloille:

$$
P(S_i|S_{\pi(i)}) = \frac{\exp(\sum_{j\in\pi(i)} \theta_{ij} S_j - \theta_{i0})^{S_i}}{1 + \exp(\sum_{j\in\pi(i)} \theta_{ij} S_j - \theta_{i0})} \tag{5}
$$

- Yhteisjakauma:

$$
P(S|\theta) = \prod_i P(S_i|S_{\pi(i)}) = \prod_i \left[ \frac{\exp(\sum_{j\in\pi(i)} \theta_{ij} S_j - \theta_{i0})^{S_i}}{1 + \exp(\sum_{j\in\pi(i)} \theta_{ij} S_j - \theta_{i0})} \right] \tag{6}
$$

## Uskomusverkko, jatkoa

- Kokeillaan täydellisesti tekijöihin jaettua approksimoivaa jakaumaa:

$$
Q(H|E,\mu) = \prod_{i\in H} \mu_i^{S_i} (1-\mu_i)^{(1-S_i)} \tag{7}
$$

- Lasketaan alaraja:

$$
\begin{aligned}
C &= \sum_{\{H\}} Q(H|E) \log \frac{P(H,E)}{Q(H|E)} \\
&= \sum_{\{H\}} Q(H|E) \log P(H,E) - \sum_{\{H\}} Q(H|E) \log Q(H|E) \\
&= C_P + C_Q
\end{aligned}
\tag{8}
$$

## Uskomusverkko, jatkoa

- Entropiatermi $(-Q \log Q)$ on helppo:

$$
\begin{aligned}
C_Q &= \sum_{\{H\}} -Q(H|E) \log Q(H|E) \\
&= -\sum_{i \in H} \sum_{\{H\}} Q(H|E) \left[ S_i \log \mu_i + (1 - S_i) \log(1 - \mu_i) \right] \\
&= -\sum_{i \in} \left[ \mu_i \log \mu_i + (1 - \mu_i) \log(1 - \mu_i) \right]
\end{aligned}
\tag{9}
$$

## Uskomusverkko, jatkoa

- Toinen termi $(Q \log P)$ on hieman hankalampi:

$$
\begin{aligned}
C_P &= \sum_{\{H\}} Q(H|E) \log P(H, E) = \sum_{i \in S} \sum_{\{H\}} Q(H|E) \log P(S_i | S_{\pi(i)}, \theta) \\
&= \sum_{i<j} \theta_{ij} \mu_i \mu_j + \sum_i \theta_{i0} \mu_i - \sum_i \langle \log \left[ 1 + \exp z_i \right] \rangle
\end{aligned}
\tag{10}
$$

- Tässä $\langle \cdot \rangle$ on odotusarvo jakauman $Q(H|E)$ yli ja $z_i = \sum_{j \in \pi(i)} \theta_{ij} S_j - \theta_{i0}$

## Uskomusverkko, jatkoa

- Odotusarvoa $\langle \log \left[ 1 + \exp z_i \right] \rangle$ ei voida helposti laskea täsmällisesti, mutta sille voidaan muodostaa variaatioapproksimaatio

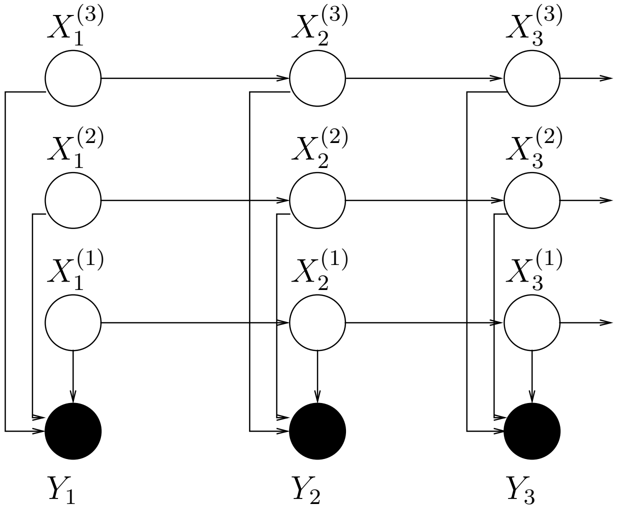
$$
\begin{aligned}
\langle \log \left[ 1 + \exp z_i \right] \rangle &= \langle \log \left[ \exp(\xi_i z_i) \exp(-\xi_i z_i)(1 + \exp z_i) \right] \rangle \\
&= \xi_i \langle z_i \rangle + \langle \log \left[ \exp(-\xi_i z_i) + \exp((1 - \xi_i) z_i) \right] \rangle \\
&\leq \xi_i \langle z_i \rangle + \log \langle \exp(-\xi_i z_i) + \exp((1 - \xi_i) z_i) \rangle
\end{aligned}
\tag{11}
$$

- Aiheesta lisää myöhemmin kurssilla

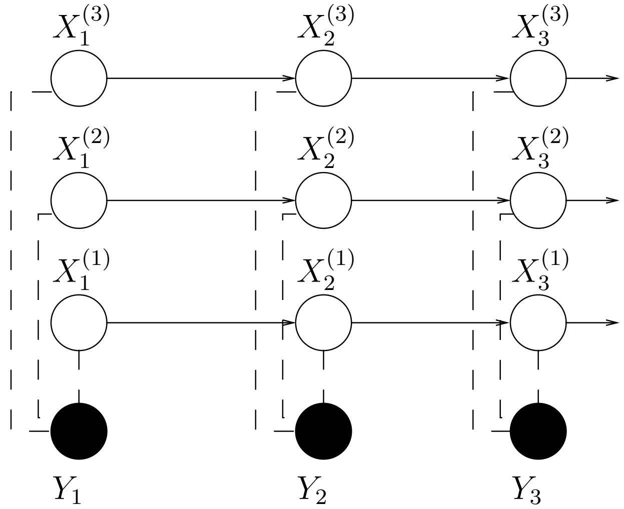
## Esimerkki: Faktoriaalinen HMM

- Ei tarvita täysin tekijöihin jaettua approksimaatiota, pelkkä yksinkertaisempaa osaverkkoa vastaava jakauma riittää
  - Katkotaan vain ketjujen suorat riippuvuudet havainnoista
  - Tämä erottaa ketjut, jolloin kuhunkin voidaan käyttää erikseen tavallisia HMM-algoritmeja

## Faktoriaalinen HMM, jatkoa



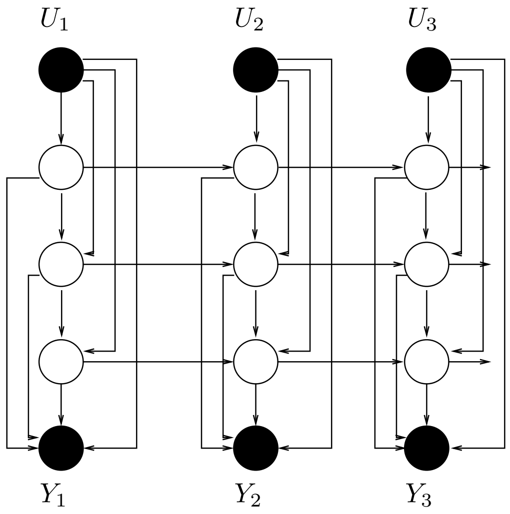

## Faktoriaalinen HMM, jatkoa



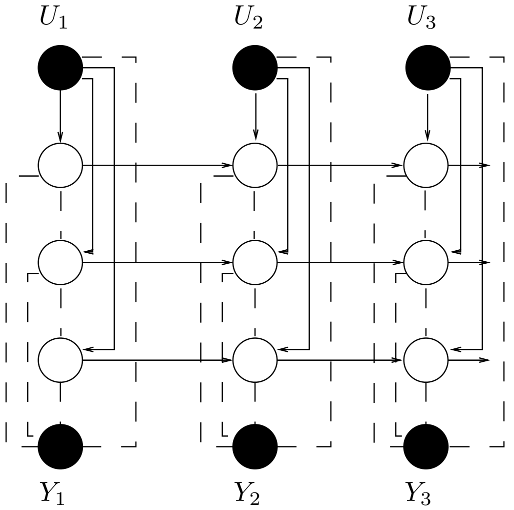

## Esimerkki: Kätketyt Markov-päättelypuut

- Jälleen voidaan käyttää monimutkaisempaa approksimaatiota, mutta nyt selkeitä vaihtoehtoja onkin useampi
  - Katkotaan ketjujen suorat riippuvuudet havainnoista kuten edellä
  - Katkotaan "aikariippuvuudet" ketjujen sisällä

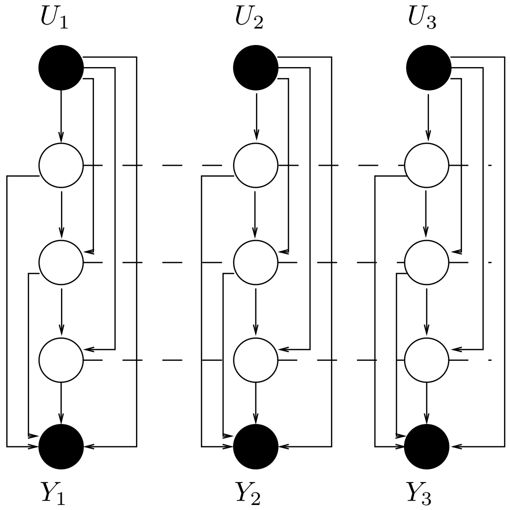
## Kätketyt Markov-päättelypuut, jatkoa



## Kätketyt Markov-päättelypuut, jatkoa




## Kätketyt Markov-päättelypuut, jatkoa




## Parametrien estimoinnista

- Noodeihin liittyy erillisiä parametreja $\theta$
- Myöhemmin kurssilla näytetään, että suurimman uskottavuuden estimoinnissa käytetty EM-algoritmi voidaan tulkita vastaavan lausekkeen optimoinniksi
- Tällöin optimoidaan vuoroin piilomuuttujien posteriorijakaumaa $P(H|E,\theta)$, vuoroin parametreja $\theta$
- Variaatioapproksimaatio posteriorille $\Rightarrow$ kevyempi approksimatiivinen variantti EM:lle

## Läheisiä menetelmiä: Helmholtzin kone

- Ohjaamattomassa oppimessa mallit usein generatiivisia
  - Piilomuuttujien estimointi vaikeaa
- Muodostetaan erillinen malli käänteiseen suuntaan
- Menetetään iteratiivisen menetelmän esiintuomat riippuvuudet
- Kääntävä malli alkuarvona iteraatiolle?

## Läheisiä menetelmiä: Otanta

- Toinen merkittävä vaihtoehto approksimatiiviseen päättelyyn ja oppimiseen Bayes-verkoissa (ja yleensäkin)
- Arvotaan realisaatioita, jotka lopulta tulevat todellisesta posteriorijakaumasta
- Käsitellään tarkemmin ensi viikolla

## Otanta, jatkoa

- Suurin ero menetelmissä viestinnässä
  - Otannassa noodit viestivät arvottuja realisaatioita
  - Variaatiomenetelmissä välitetään approksimatiivisia jakaumia
- Otanta konvergoi varmasti mutta hidas
- Marginaalijakaumien estimointi otannalla vaikeaa
- Yhdistäminen mielenkiintoinen ongelma

## Läheisiä menetelmiä: Bayesiläiset menetelmät

- Variaatiomenetelmiä voidaan soveltaa yleiseen bayesiläiseen tilastolliseen analyysiin
- Ensemble-oppiminen vastaa edellä esitettyä blokkimenetelmää ja optimoi Kullback-Leibler -divergenssiä approksimoidun ja todellisen posteriorin välillä jatkuville malleille
- Monia menestyksellisiä käytännön sovelluksia

## Laajennus: Sekoitejakaumat

- Seuraava luku kirjassa (Jaakkola ja Jordan)
- Parempi approksimaatio monihuippuisille jakaumille sekoitemallilla:

$$Q_{\mathsf{mix}}(H) = \sum_m \alpha_m Q_{\mathsf{mf}}(H|m) \tag{12}$$

- Tätä vastaava Kullback-Leibler -divergenssin antama alaraja:

$$\begin{aligned} C &= \sum_m \alpha_m C_m + \sum_{m,H} \alpha_m Q_{\mathsf{mf}}(H|m) \log \frac{Q_{\mathsf{mf}}(H|m)}{Q_{\mathsf{mix}}(H)} \\ &= \sum_m \alpha_m C_m + I(m;S) \end{aligned} \tag{13}$$

## Sekoitejakaumat, jatkoa

- Summan ensimmäinen termi on konveksi kombinaatio eri mallien antamista rajoista, toinen indeksin $m$ ja muuttujien $S$ välinen yhteisinformaatio $I(m; S)$
- Informaatioteoriasta: $I(m; S) \leq \log M$, $M$ sekoitteen komponenttien määrä
- Yhteisinformaation tarkka määrittäminen vaikeaa (kytkee kaikki muuttujat sekoitejakauman kautta)
- Variaatiomenetelmin yhteisinformaatiolle heikompi, helposti laskettava alaraja

## Yhteenveto variaatiomenetelmistä

- Approksimatiivisia laskentamenetelmiä Bayes-verkoille
- Yleinen viitekehys approksimaatioiden muodostamiseen, ei vain yhtä menetelmää
- Helpotetaan ongelmaa erilaisin muunnoksin, kunnes voidaan käyttää täsmällisiä menetelmiä
- Muunnoksia noodi kerrallaan tai koko verkko kerralla

## Yhteenveto, jatkoa

- Globaalissa approksimaatiossa yleinen alaraja evidenssille, ylärajasta ei tietoa
- Palautuu epälineaariseen optimointitehtävään, ei aina takeita ratkaisusta
- Antaa kuitenkin usein hyviä tuloksia nopeasti
- Monet teoreettiset kysymykset vielä avoimia, menetelmien käyttö vaatii hieman käsityötä