G. E. Hinton, B. Sallans ja Z. Ghahramani :

# A Hierarchical Community of Experts

(Hierarkkinen eksperttikommuuni)

Esittäjä : Jarkko Salojärvi

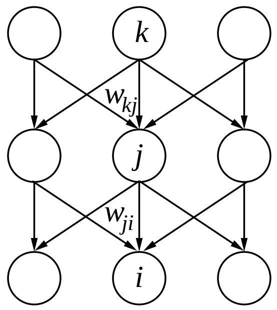


## Esitelmän rakenne

1. Monikerroksiset gaussiset verkot
   - Topologia
   - Edut ja haitat
   - Opetus
2. Monikerroksiset binääriset logistiset verkot
   - kuten kohta 1
3. Binääriyksikköjen käyttö gaussisten yksikköjen aktivaation määräämiseen
   - kuten kohta 1
4. Gibbsin otanta eksperttikommuunin opetuksessa ja miksi lyhytkin otanta riittää
5. Kokeet
6. Yhteenveto



## Monikerroksiset (lineaari)gaussiset verkot: topologia




- Ylimmän kerroksen yksiköiden $k$ tila on gaussisesti jakautunut datasta opitulla keskiarvolla ja varianssilla.
- Seuraavan kerroksen yksiköiden $j$ jakauma on $N(\hat{y}_j, \sigma_j^2)$, missä

$$\hat{y}_j = b_j + \sum_k w_{kj} y_k \tag{1}$$

⇒ Kunkin yksikön tila noudattaa gaussista jakaumaa. Keskiarvo on lineaarinen funktio sen vanhemmista $y_k$ (+bias $b_j$). Varianssit $\sigma_j^2$ ja painot $w_{kj}$ määritetään datasta (EM).



## Monikerroksiset (lineaari)gaussiset verkot: edut ja haitat

\+ Ovat usein hyviä malleja jatkuvalle datalle

\+ Sovitus onnistuu, vaikka puuttuvaa dataa olisi paljonkin (EM-algoritmilla).

\- Jättävät datasta huomiotta korkeamman asteen statistiikan.

## Monikerroksiset (lineaari)gaussiset verkot: opetus

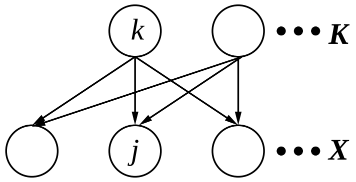
Opetus voidaan tehdä EM-algoritmilla :

- Oletetaan, että verkko jakautuu piilossa oleviin tiloihin H ja näkyviin tiloihin D. Tehtävänä on löytää ML-estimaatti verkon parametreille Θ. Algoritmissa iteroidaan seuraavia kahta askelta.
- Expectation : Pyritään estimoimaan piilossa olevat tilat jakaumalla $\tilde{P}_t(H) = P(H|D, \Theta_{t-1})$.
- Maximization : Puuttuvan datan estimaattia käyttämällä maksimoidaan näkyvän verkon log-uskottavuus, eli $\Theta_t = \max_\Theta[\log p(D|\Theta)] = E_{\tilde{P}_t(H)}[\log p(D, H|\Theta)]$.



## Monikerroksiset (lineaari)gaussiset verkot: opetus, esimerkki

- Faktorianalyysin generatiivinen malli koostuu yhdestä piilokerroksesta gaussisia yksiköitä (faktorit), joista lähtee painotetut yhteydet (factor loadings) näkyvään, gaussisista yksiköistä koostuvaan, kerrokseen.




eli $\mathbf{X}_i = \mathbf{W}\mathbf{K}_i + \mathbf{u}_i$

- oletukset: $\mathbf{K} \sim N(0, I)$, näytteet $p(\mathbf{X}|\mathbf{K})$ riippumattomia, eli siis kohina $\mathbf{u}$ on korreloimatonta.

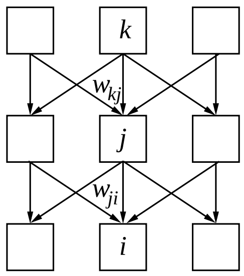
E: estimoidaan $E(\mathbf{K}|\mathbf{X})$ ja $E(\mathbf{K}\mathbf{K}'|\mathbf{X})$ jokaiselle datapisteelle $\mathbf{X}_i$ edellisen kierroksen parametrien arvoilla.

M: Päivitetään uudet parametrien arvot $w_{kj}, \sigma_j$. [a]

[a] Johto ja päivityskaavat artikkelissa Z. Ghahramani & G.E. Hinton : The EM Algorithm for Mixtures of Factor Analyzers (Z.G:n kotisivulla)



## Monikerroksiset (binääri)logistiset verkot [a]: topologia




- Topologia on samanlainen kuin (lineaari)gaussisessa verkossa, mutta yksiköiden toiminta on erilaista.
- Yksikön $j$ tila voi olla joko 0 tai 1.
- Tilan 1 todennäköisyys yksikölle $j$ määräytyy vanhempien lineaarisena summana, joka viedään sigmoidin läpi, eli

$$p(s_j = 1|\{s_k : k \in pa_j\}) = \sigma(\hat{s}_j) = \frac{1}{1 + e^{-\hat{s}_j}}, \tag{2}$$

missä $\hat{s}_j = b_j + \sum_k w_{kj} s_k$ kuten edellä.

[a] Logistic Belief Nets



## Monikerroksiset (binääri)logistiset verkot: edut ja haitat

\+ Vapaiden parametrien vähäisyys : Ei tarvita erillistä varianssiparametriä, sillä $\hat{s}_j$ riittää määrittelemään Bernoullin jakauman.

\- Eksaktin posteriorin laskeminen skaalautuu eksponentiaalisesti. Tämä voidaan kiertää käyttämällä opetuksessa Gibbsin otannalla saatuja näytteitä posteriorista.

- Neal (1992) : Toimii ainakin paremmin kuin Bolzmannin kone ja noisy-OR -verkko.

## Monikerroksiset (binääri)logistiset verkot: opetus Gibbsin otannalla

Tehdään jokaiselle datapisteelle erikseen :

- Fiksataan näkyvät tilat datavektorin arvoihin.
- Käydään piiloyksiköissä $u$ yksi kerrallaan. Arvotaan joka kierroksella sen tila todennäköisyydellä, joka on verrannollinen verkon Bernoullin jakauman antamien todennäköisyyksien suhteeseen, kun $u$ on tilassa 1 tai 0. (Energioiden erotus vietynä sigmoidin läpi.)
- Kun otantaa jatketaan, saadaan piilotilojen konfiguraatioita, jotka on valittu posterioritodennäköisyyksiensä mukaan.
- Laskemalla uskottavuuden derivaatta painojen suhteen ja käyttämällä näytteitä piilotilojen jakaumasta, voidaan päivittää verkon painoja $w$ gradienttinousumenetelmällä.



## Hierarkkinen eksperttikommuuni, eli binääriyksikköjen käyttö gaussisten yksikköjen aktivaation määräämiseen

Gaussisia lineaarimalleja voidaan yleistää käyttämällä niiden sekoitusta (esim. mixture of factor analyzers).

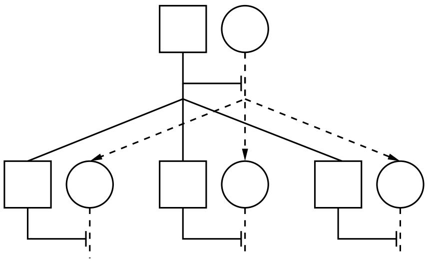
+ Posteriori osataan edelleen laskea : Lasketaan posteriori jokaisen mallin yli ja normalisoidaan se.

- Mikstuurilla ei voida kuitenkaan esittää objektien kombinaatioita. Kullekin kombinaatiolle tarvittaisiin oma lineaarimallinsa.

Ratkaisuna on hajautettu esitys, jossa kukin lineaariyksikkö mallintaa oman osansa datasta. $\Rightarrow$ yhdistetään edelliset verkot niin, että lineaariyksikön toimintaa säädellään binääriyksiköllä.



## Hierarkkinen eksperttikommuuni



Mikäli binääriyksikkö toimii gaussisen yksikön porttina, tulee ulostulosta

$$\hat{y}_j = b_j + \sum_k w_{kj} y_k s_k. \tag{3}$$

Tästä aiheutuu riippuvuuksia näiden kahden verkon välille, minkä takia verkkoja ei voi suoraan opettaa erikseen kuten edellä.



## Eksperttikommuunin opetus opetus Gibbsin otannalla(1/2)

Periaatteessa Gibbsin otanta tehdään kuten logististen verkkojen tapauksessa, eli :

- Vieraillaan yksi kerrallaan kussakin yksikössä ja arvotaan yksikön tila sen posteriorijakaumasta kaikkien muiden yksiköiden tilojen ollessa kiinnitettyjä.
  - Binääriyksikölle lasketaan verkon energia ($-\log P$), kun yksikkö on tilassa 0 tai 1.
  - Gaussiselle yksikölle pitää laskea neliöllinen funktio, joka määrittää kuinka verkon energia riippuu yksikön tilasta.

## Eksperttikommuunin opetus Gibbsin otannalla(2/2)

Verkkojen riippuvuuksista aiheutuneet ongelmat ratkaistaan seuraavasti :

- Binääriyksikön tila riippuu gaussisen yksikön arvoista, joten päätettäessä binääriyksikön tilaa $s_j$, integroidaan yli kaikkien mahdollisten gaussisen yksikön arvojen $y_j$.
- Vieraillaan siis ensin binääriyksikössä. Jos se kytkeytyy päälle, arvotaan gaussisen yksikön tila.
- Sellaiset gaussiset yksiköt, jotka ovat pois päältä, integroidaan pois kokonaisenergiasta.

Kun Gibbsin otannalla on lopulta saatu generoitua näytteet posteriorijakaumasta, voidaan painoja päivittää yllättävän yksinkertaisilla kaavoilla :

$$\Delta w_{ji}^{(gauss)} = \epsilon y_j s_j (y_i - \hat{y}_i) s_i / \sigma_i^2 \quad (4)$$

$$\Delta w_{ji}^{(binary)} = \epsilon s_j (s_i - \hat{s}_i) \quad (5)$$

$$\Delta \sigma_j^2 = \epsilon s_j [(y_j - \hat{y}_j)^2 - \sigma_j^2] \quad (6)$$



## Miksi lyhyt Gibbsin otanta riittää?

Kokeissa tehtiin vain 16 kierrosta Gibbsin otantaa (joista 4 ensimmäistä heitettiin pois). Miksi ?

- EM - algoritmi voidaan tulkita vapaan energian minimoinniksi (Neal & Hinton,1993)

$$F = -\log p(visible) + \sum_{\alpha} Q_\alpha \log \frac{Q_\alpha}{P_\alpha} \quad (7)$$

- Jokainen E- ja M -askel pienentää vapaata energiaa. M-askel pienentää uskottavuutta ja E-askel vie piilotilojen konfiguraatiota $Q_\alpha$ kohti tasapainotilaa $P_\alpha$.
- Tällöin ainoa vaatimus piilotilojen konfiguraation määrityksessä käytettävälle Gibbsin otannalle on, että se lähestyy tasapainotilaa joka iteraatiolla.
- Lyhyttä otantaa käytettiin, ettei lähdettäisi jahtaamaan tasapainotilaa, joka on kaukana alkupisteestä (joka oli aina sama).



## Koetulokset

- Paperissa esitetty kaksi (luokittelu)esimerkkiä, joissa ongelmanratkaisussa voisi olla hyötyä hierarkiasta, palkkien generointi ja numeroiden erottelu.
- Ensimmäinen esimerkki oli leludatalla. Demonstroitiin, että eksperttien kommuunilla saadaan mallinnettua hierarkkisella mallilla tuotettua dataa.
- Toisessa esimerkissä käytettiin "oikean maailman" dataa, yritettiin erottaa käsinkirjoitetut kakkoset kolmosista.
- Kokeiden perusteella malli tuntui toimivan.
- Kyseessä oli kuitenkin vain proof of concept -tyyppinen tarkastelu, ei verrokkimenetelmiä.
  - Tuloksia olisi voinut verrata ainakin perinteiseen eksperttimikstuuriin tai edes gaussiseen mikstuurimalliin.



## Yhteenveto

- Esiteltiin eksperttikommuuni, jossa kunkin lineaarigaussisen yksikön aktiivisuutta säädeltiin binäärisellä portilla.
- Eksperttimikstuurista järjestelmä eroaa siinä, että jokainen gaussinen yksikkö on oma eksperttinsä.

\+ Uusi, varsin joustava konsepti, erityisesti resurssien jakamisen kannalta.

\- Missään ei puhuta siitä, kuinka nopeasti systeemi konvergoi, jotain voi päätellä artikkelien julkaisutiheydestä.