

Ryhmäfaktorianalyysi (GFA) (aihe-esittely)

Sami Remes

12.03.2012

Ohjaaja: Arto Klami

Valvoja: Harri Ehtamo

Työn saa tallentaa ja julkistaa Aalto-yliopiston avoimilla verkkosivuilla. Muilta osin kaikki oikeudet pidätetään.

Yleinen ongelma: datajoukkojen väliset riippuvuudet (1/3)

- Tarkoituksena löytää jotain jaettuja ominaisuuksia useista datajoukoista
- Klassisia menetelmiä mm. kanoninen korrelaatio (CCA) ja faktorianalyysi
- CCA etsii projektiot avaruuteen, jossa projisoidut muuttujat korreloivat maksimaalisesti

Yleinen ongelma: datajoukkojen väliset riippuvuudet (2/3)

- CCA-esimerkki
 - Datajoukko 1: oppilaan pituus, paino, ÄO ym. ominaisuuksia
 - Datajoukko 2: arvosana eri kouluaineissa
- CCA saattaisi löytää esimerkiksi, että jokin summa ominaisuuksia korreloi joidenkin arvosanojen summan kanssa
- Useita “komponentteja”, ts. eri joukkoja painoja eri joukoille kouluaineita

Yleinen ongelma: datajoukkojen väliset riippuvuudet (3/3)

- CCA vain kahdelle datajoukolle
- Joissakin sovelluksissa ongelmana datan suuri dimensio; näytteitä paljon vähemmän
- Ryhmäfaktorianalyysi (Group Factor Analysis, GFA) yleistää menetelmän M:lle datajoukolle ja myös tapauksiin joissa dimensio suurempi kuin näytteiden määrä
 - Etsii myös “komponentteja”, jotka ovat aktiivisia vain jossakin osajoukossa datajoukkoja (toisin kuin monet muut CCA:n yleistykset)

Ryhmäfaktorianalyysi (1/2)

- Malli datajoukoille Y1 ja Y2, $Y = [Y1 \ Y2]$
- $Y = Z*W^T + E$, missä Z on piilomuuttujat ja E virhetermi
- W:llä rakenne

$$W = \begin{pmatrix} W_1 & V_1 & 0 \\ W_2 & 0 & V_2 \end{pmatrix}$$

- Tällainen matriisin rakenne saa aikaan komponentteja jotka voivat olla aktiivisia mielivaltaisessa datajoukkojen osajoukossa (kun datajoukkoja enemmän kuin kaksi)

Ryhmäfaktoriansalyysi (2/2)

- Bayesilainen malli
 - Parametreille priorit ja hyperpriorit
- Posteriorin approksimointi variaatiomenetelmällä
 - Oletetaan että posteriori faktoroituu tietyllä tavalla parametrien suhteen, esim. $p(A,B) = p(A)p(B)$
 - Minimoidaan tietty “etäisyys” approksimaation ja oikean posteriorin välillä
- Toteutus R-kielellä

Työn tavoitteet ja raja

- Katsaus menetelmän taustoihin:
 - Klassiset menetelmät
 - Klassisten menetelmien probabilistinen tulkinta
 - Bayesilaiset menetelmät ja approksimaatiot
- Ryhmäfaktorianalyysi
- Numeerisia kokeita
 - Simuloitua dataa
- Soveltaminen neurotieteeseen
 - Aivojen eri osat ja ärsyke datajoukkoina

Aikataulu

- Työ on jo aloitettu, tarkoitus saada pikimmiten valmiiksi
- Toukokuun tai kesän seminaari

Viitteitä

- Seppo Virtanen, Arto Klami, Samuel Kaski: **Bayesian CCA via Group Sparsity**. ICML 2011.
- Seppo Virtanen, Arto Klami, Suleiman Khan, Samuel Kaski: **Bayesian Group Factor Analysis**. AISTATS'12.
- Christopher M. Bishop: **Machine Learning and Pattern Recognition**. Springer, 2007 (2nd Ed.).