

Bolla Mariann - Csicsman József

Algoritmikus modellek és tanulóalgoritmusok a statisztikában

2011

Támogatás:

Készült a TÁMOP-4.1.2.A/1-11/1-2011-0064 számú, a „Természettudományos (matematika és fizika) képzés a műszaki és informatikai felsőoktatásban” című projekt keretében.



Készült:

a BME TTK Matematikai Intézet gondozásában

Szakmai felelős vezető:

Ferenczi Miklós

ISBN 978-963-279-893-6

Copyright: © 2013–2019, Bolla Mariann, Csicsman József, BME

A © terminusai: A szerző nevének feltüntetése mellett nem kereskedelmi céllal szabadon másolható, terjeszthető, megjelentethető és előadható, de nem módosítható.

Előszó

Jegyzetünk azoknak a hallgatóknak készült, akik matematikai statisztika és többváltozós statisztika tanulmányaik után szeretnék megismerni a modern statisztikai modelleket és módszereket is. A klasszikus statisztika fogalomrendszere és legtöbb tétele a XX. század első felében lett kidolgozva, elsősorban valószínűségszámítási alapokon. Ebben jelentős szerepet játszott az angolszász, orosz és indiai iskola. Érdekes, hogy olyan kulcsfontosságú eredmények, mint a Cramér–Rao egyenlőtlenség, Rao–Blackwellizálás, és a Wald-féle szekvenciális döntési eljárás a II. Világháború idején születtek meg, utóbbi töltények gazdaságos minőségellenőrzésére.

A XX. század közepére kifejlesztették a többváltozós statisztikai eljárásokat is, amelyek széleskörű alkalmazásának azonban csak a nagy teljesítményű számítógépek elterjedése nyitott utat a XX. század második felében (BMDP, SPSS programcsomagok), hiszen ezek a módszerek nagyméretű adatmátrixok és kovarianciamátrixok szinguláris- és spektrális felbontásán alapulnak. Nagyjából ezeket az ismereteket foglalja össze a BME matematikus képzés BSc és MSc statisztika anyagának gerincét képező Bolla–Krámlí, Statisztikai következtetések elmélete (Typotex, 2005 és 2012) könyv.

Az 1970-80-as években azonban már ez a tényanyag sem bizonyult elégségesnek. Valós életbeli (biológiai, pszichológiai, szociológiai) adatrendszerekkel foglalkozva azt találtuk, hogy a klasszikus módszerek nem alkalmazhatók mindig közvetlenül, illetve a problémák sokszor túlmutattak a tanult (elsősorban többváltozós normális eloszlású mintákra kifejlesztett) módszerek alkalmazhatósági körén (diszkrét, nem-paraméteres szituációk, időben is változó megfigyelések). L. Breiman, *Statistical modeling: the two cultures* (Statist. Sci. 16) 2001-es cikkében szintén rámutat arra, hogy gyakorlati problémákkal szembeesülve a klasszikus apparátus néha csődöt mond. Az ún. második kultúra egy algoritmikus szemléletet visz a klasszikusba, ami azonban nem a numerikus módszerek automatikus alkalmazását jelenti, hanem olyan elméleti algoritmusok kifejlesztését, melyek az információelmélet, a Hilbert-terek, sőt akár a gráfelmélet eszköztárát használják magas színvonalon. Ebbe az eszköztárba szeretnénk betekintést nyújtani.

Ilyen módon a tankönyv egy, a modern statisztikai módszerek iránt érdeklődő hallgatók számára a BME-n két évente tartott kurzus anyaga, de használható témalabor vagy diplomamunka készítéséhez is, illetve az elméleti részek kihagyásával a leírt algoritmusok nagyméretű adatrendszerek adatbányászataival foglalkozó szakemberek számára is hasznosak lehetnek. Az algoritmikus modellek köre egyre terjed, itt csak a legfontosabbakat foglaltuk össze, de utalunk egyéb, hasonló célú eljárásokra, illetve bőséges szakirodalmat közlünk a részletek iránt érdeklődőknek. Az elméleti részek tanulmányozása pedig az arra fogékony olvasók kezébe ötleteket és eszközöket adhat hasonló szituációk kezelésére.

Bolla Marianna,
Csicsman József

Budapest, 2013. július 5.

Tartalomjegyzék

1. Bevezetés	3
2. Az EM-algoritmus hiányos adatrendszerre	5
2.1. Egy konkrét példa	6
2.2. Elméleti megfontolások	8
2.3. Alkalmazások	13
2.3.1. EM-algoritmus normális eloszlások keverékfelbontására	13
2.3.2. EM-algoritmus polinomiális eloszlások keverékfelbontására	19
2.3.3. EM-algoritmus gráfok klaszterezésére	21
Irodalomjegyzék	24
3. Az ACE-algoritmus általánosított regresszióra	25
3.1. Elméleti megfontolások	26
3.2. ACE-algoritmus egymásba ágyazott ciklusokkal	31
3.3. ACE-algoritmus adatmárixra simításokkal	33
3.4. Az ACE-algoritmus outputja	37
3.5. Alkalmazások	37
Irodalomjegyzék	40
4. Reprodukáló magú Hilbert-terek	41
4.1. Elméleti háttér	41
4.2. Példák	44
4.3. Empirikus kernel	47
4.4. Szemléletes példák	48
Irodalomjegyzék	49
5. Spektrális klaszterezés	50
5.1. Gráfok és hipergráfok reprezentációja	51
5.1.1. Egyszerű és súlyozott gráfok	51

5.1.2.	Hipergráfok	53
5.1.3.	Normált Laplace mátrix	55
5.1.4.	Modularitás mátrix	56
5.1.5.	Nevezetes gráfok spektruma	57
5.2.	Minimális vágások, maximális modularitás	60
5.2.1.	Arányos és kiegyensúlyozott vágások	60
5.3.	Általánosított véletlen gráfok	67
5.3.1.	Felfújtt zajos mátrixok	68
5.3.2.	Reguláris partíciók	71
5.4.	Algoritmusok gráfok és hipergráfok klaszterezésére	72
5.4.1.	Súlyozott gráfok	72
5.4.2.	Hipergráfok kétszemponjú klaszterezése	74
5.5.	Irodalom jegyzék	75
Irodalomjegyzék		76
6.	Dinamikus faktoranalízis	83
6.1.	Előzmények és célkitűzések	83
6.2.	A modell	84
6.3.	A paraméterek becslése	86
6.4.	Szimmetrikus mátrixok kompromisszuma	90
6.5.	Alkalmazás	91
Irodalomjegyzék		96
7.	A varianciaanalízis általános modelljei	98
7.1.	Többváltozós varianciaanalízis (MANOVA)	98
7.2.	Nemparaméteres varianciaanalízis	99
Irodalomjegyzék		103