

Antal Péter, Antos András, Horváth Gábor, Hullám Gábor, Kocsis Imre,
Marx Péter, Millinghoffer András, Pataricza András, Salánki Ágnes

Intelligens adatelemzés

Szerkesztette: Antal Péter

A jegyzetben az adatelemzés folyamata szerint ismertetünk „intelligens” megközelítéseket, amelyekben az intelligencia a felhasznált háttértudás, a számítási erőforrások vagy a modellek komplex volta miatt is jelenik meg.

Budapesti Műszaki és Gazdaságtudományi Egyetem



Semmelweis Egyetem



Typotex Kiadó 2014

Copyright: 2014–2019, Kocsis Imre, Horváth Gábor, Hullám Gábor, Millinghoffer András, Antal Péter, Marx Péter, Antos András, Salánki Ágnes, Pataricza András, Budapesti Műszaki és Gazdaságtudományi Egyetem, Semmelweis Egyetem

Creative Commons NonCommercial-NoDerivs 3.0 (CC BY-NC-ND 3.0) A szerző nevének feltüntetése mellett nem kereskedelmi céllal szabadon másolható, terjeszthető, megjeleníthető és előadható, de nem módosítható.

Lektorálta: Takács Gábor
Készült a Typotex Kiadó gondozásában
Felelős vezető: Votisky Zsuzsa

ISBN 978-963-279-171-5

Készült a TÁMOP-4.1.2/A/1-11/1-2011-0079 számú, „Konzorcium a biotechnológia és bioinformatika aktív tanulásáért” című projekt keretében.

Nemzeti Fejlesztési Ügynökség
www.ujszecsenyiterv.gov.hu
06 40 638 638



A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Tartalomjegyzék

1. Aktív tanulás	1
1.1. Bevezetés: becslés, döntés, felügyelt (passzív) tanulás	1
1.1.1. Bayes-döntés	2
1.1.2. Bayes-döntés közelítése	3
1.1.3. Bayes-becslés	4
1.1.4. Regressziós becslés; négyzetes középhiba minimalizálás	5
1.2. Aktív tanulás fogalma	5
1.3. Véges sok középérték aktív tanulása	6
1.3.1. Megvalósítási lehetőségek	8
1.3.2. GAFS algoritmus	9
2. Dimenzióredukció	11
2.1. Absztrakt	11
2.1.1. Kulcsszavak:	11
2.2. Bevezetés	11
2.3. A dimenzió átka	12
2.4. A dimenzióredukció alkalmazási területei	12
2.5. Főkomponens-analízis (PCA, KLT)	15
2.5.1. Főkomponens- és altér-meghatározó eljárások	18
2.6. Nemlineáris dimenzióredukciós eljárások	21
2.6.1. Kernel PCA	23
2.6.2. Nemlineáris altér algoritmusok	27
2.7. Irodalom	28
3. Ritka események detektálása	30
3.1. Anomáliák és ritka események	31
3.2. Detektálási megközelítések	34
3.2.1. Távolság alapú módszerek	34
3.2.2. Sűrűség alapú módszerek	42
4. Hiányos adatok	48
4.1. Bevezetés	48
4.2. A hiányzás típusai	49

4.3.	Hiányos adatok kezelése	51
4.3.1.	Teljes eset módszer	51
4.3.2.	Ad-hoc módszerek	51
4.3.3.	Súlyozás	52
4.3.4.	Pótlás	52
4.3.5.	Többszörös pótlás	56
5.	Vizuális analízis	58
5.1.	Felderítés, megerősítés és szemléltetés	59
5.2.	Egydimenziós diagramok	60
5.2.1.	A doboz-ábra	60
5.2.2.	Hisztogram	61
5.3.	Kétdimenziós diagramok	62
5.4.	n-dimenziós diagramok	62
5.4.1.	Mozaik és fluktuációs diagram	62
5.4.2.	A párhuzamos koordináta ábra	63
5.4.3.	Eszköztámogatás	66
5.5.	Interaktív statisztikai grafika	66
5.5.1.	Lekérdezések	67
5.5.2.	Helyi interakciók	67
5.5.3.	Kiválasztás és csatolt kijelölés	67
5.5.4.	Csatolt analízis	69
5.6.	Összefoglalás	69
6.	Monte-Carlo-módszerek, Bayesi modellátlagolás, Bayesi predikció	70
6.1.	Monte-Carlo-integrálás	71
6.1.1.	Közvetlen mintavételezés	71
6.1.2.	Elutasító mintavételezés	72
6.1.3.	Fontossági mintavételezés	73
6.2.	Markov-láncok	73
6.3.	Metropolis–Hastings-algoritmus	76
6.4.	A Metropolis–Hastings-algoritmus aletei	78
6.4.1.	Gibbs-mintavételezés	78
6.4.2.	Többláncos MCMC	79
6.4.3.	Reversible jump MCMC	80
6.5.	Konvergencia	80
6.5.1.	Konvergencia tényének vizsgálata	81
6.5.2.	Mintavételezés hatékonysága	82
6.6.	Alkalmazás Bayes-hálókbán	83
7.	Bootstrap-módszerek	86
7.1.	Ensemble-módszerek áttekintése	87
7.2.	A bootstrap alapjai	88

7.3. További aspektusok	89
7.3.1. Adatmodell	89
7.4. Konfidencia becslések	90
7.5. Permutációs teszt és hipotézistesztelés	91
8. Kernel technikák az intelligens adatelemzésben	93
8.1. Bevezetés	93
8.2. Kernelek konstrukciója	94
8.2.1. Leggyakrabban használt kernel függvények	94
8.2.2. Műveletek kernelekkel	94
8.3. Prior információ hozzáadása	95
8.3.1. Tanító halmaz bővítése	95
8.3.2. Prior információ kernelbe ágyazása	97
8.4. Kernelek gráfokra	99
8.4.1. Diffúziós kernelek	100
8.5. Adatbázisok	100
Jelölések	102
9. Valószínűségi Bayes-hálók tanulása	105
9.1. Paramétertanulás Rejtett Markov Modellekben	105
9.1.1. Paramétertanulás RMM-ekben ismert állapotszekvenciák esetében	106
9.1.2. E-M alapú paramétertanulás RMM-ekben ismeretlen állapotszekvenciák esetében	107
9.2. Naiv Bayes-hálók tanulása	109
9.2.1. A bayesi feltételes modellezés	109
9.2.2. Bayes-hálók tanulása feltételes modellként	111
9.2.3. Naiv Bayes-hálók teljesítménye osztályozásban és regresszióban	112
9.2.4. Naiv Bayes-hálók kiterjesztései	113
9.2.5. Teljes modellátlagolás NBN-ek felett	114
9.3. Egy információelméleti pontszám Bayes-háló tanulásához	115
10. Oksági Bayes-hálók tanulása	117
10.1. Bayesi következtetés és tanulás rögzített oksági struktúra esetén	117
10.2. A prekvenciális modellkiértékelés	120
10.2.1. Általános és valószínűségi előrejelző rendszerek vizsgálata	121
10.2.2. Bayes-hálók prekvenciális vizsgálata	122
10.3. Oksági struktúrák tanulása	123
10.3.1. Kényszer alapú struktúratanulás	123
10.3.2. Pontszámok oksági struktúrák tanulására	124
10.3.3. Az optimalizálás nehézsége struktúratanulásban	126