

T-61.2010 Datasta tietoon, syksy 2011

professori Erkki Oja

professori Heikki Mannila

31.10.2011

Sisältö

1	Datasta tietoon -johdantoluento	3
1.1	Miksi tällainen kurssi?	3
1.2	Kurssin suorittamiseen liittyviä asioita	5
1.3	Kurssin sisältö luvuittain	9
2	Datasta tietoon: mitä dataa? mitä tietoa?	11
2.1	Data-analyysin ongelma	11
2.2	Mallit ja oppiminen	11
2.3	Esimerkkejä	14
2.4	Case study: WEBSOM	15
3	Data vektoreina	16
3.1	Datamatriisi	16
3.2	Piirreirrotus: ääni- ja kuvasignaalit	19
3.3	Dimensionaalisuuden kirous	27
4	Vektoridatan tiivistäminen ja dekorrelointi	30
4.1	Datamatriisi	30
4.2	Pääkomponenttianalyysi (PCA)	31
4.3	PCA-esimerkkejä: ominaiskasvot	34
4.4	DSS-menetelmä	37
5	Estimointiteorian perusteita	42
5.1	Perusjakaumat 1-ulotteisina	42
5.2	Yleistys vektoridatalle, d :n muuttujan normaalijakauma	44

5.3	Suurimman uskottavuuden periaate	47
5.4	Bayes-estimointi	50
5.5	Regressiosovitus	52
5.6	Esimerkki regressiosta: neuroverkko	55
6	Hahmontunnistuksen perusteita	63
6.1	Luokittelu	63
6.2	Lähimmän naapurin luokitin (kNN)	66
6.3	Bayes-optimaalinen luokitin	67
6.4	Ryhmittelyanalyysi	71
6.5	Hierarkkinen ryhmittely	73
6.6	c-means ryhmittelyalgoritmi (k-means, KM)	75
7	Itseorganisoiva kartta	82
7.1	Itseorganisoiva kartta (SOM)	82
7.2	Itseorganisoivan kartan yhteys biologiaan	86
7.3	Itseorganisoivan kartan suppenevuus 1-ulotteisessa tapauksessa	86
7.4	Käytännön valintoja	88
7.5	Mihin SOM:ia käytetään?	89
8	Diskreettejä menetelmiä laajojen 0-1 datajoukkojen analyysiin	98
8.1	Suuret 0-1 datajoukot	98
8.2	Usein esiintyvien muuttujakombinaatioiden etsintä: kattavat joukot	105
8.3	Tasoinen algoritmi kattavien joukkojen etsintään	108
8.4	Riippumattomuus kattavissa joukoissa	111
8.5	Kattavien joukkojen merkitsevyyden tutkiminen	113
9	Relevanttien sivujen etsintä verkosta: satunnaiskulut verkossa	119
9.1	Webin lyhyt historia	119
9.2	Etsintä verkosta	121
9.3	Keskukset ja auktoriteetit	122
9.4	PageRank	126