TKK / Informaatiotekniikan laboratorio

Syyslukukausi, periodi II, 2007

Erkki Oja, professori, ja
Heikki Mannila, akatemiaprofessori:

T-61.2010 DATASTA TIETOON

JOHDANTO: Miksi tällainen kurssi?

- Uudessa tutkintorakenteessa on haluttu tarjota ohjelman yhteisissä opinnoissa (O-moduuli) kurssi, joka toimii johdatuksena *Informaatiotekniikan* opinnoille: perusmoduulille A1 (informaatiotekniikka I) sekä informaatiotekniikan pää- ja sivuainemoduuleille A2, A3 ja vastaaville *bioinformatiikan* ja *kieliteknologian* A3-moduuleille (kuvat)
- Muiden tutkinto-ohjelmien opiskelijoille (esim. TF, S, AS) kurssi sisältyy informaatiotekniikan sivuainemoduuliin B1, jonka jälkeen voi edetä A2-moduuliin
- Informaatiotekniikan laboratorio on hyvin *tutkimusintensiivinen*. TKK:lla toimivista valtakunnallisista tutkimuksen huippuyksiköistä 2 toimii info-labrassa (ainoat huippuyksiköt Tietotekniikan osastolla):
  1. "Adaptiivisen informatiikan tutkimusyksikkö" (johtajana Erkki

Oja);

2. “Datasta tietoon - tutkimusyksikkö” (johtajana TKK:lla Heikki Mannila)

- Haluamme antaa jo perusopiskelijoille tilaisuuden tutustua meillä tehtävään tutkimukseen
- Ne opiskelijat jotka *eivät valitse* informaatiotekniikkaa pää- tai sivuaineekseen saavat kuitenkin jonkinlaisen käsityksen siitä mitä ala pitää sisällään
- Ne opiskelijat jotka *valitsevat* informaatiotekniikan pää- tai sivuaineen saavat Datasta Tietoon - kurssissa katsauksen koko kenttään, joka sitten syvenee laboratorion muiden kurssien avulla ja esim. osallistumalla kesäteekkarina tutkimushankkeisiin
- Muut aiheeseen liittyvät kurssit (those with *English names* will be given totally in English):

T-61.152 Informaatiotekniikan seminaari

T-61.3010 Digitaalinen signaalinkäsittely ja suodatus

T-61.3020 Hahmontunnistuksen perusteet

T-61.3030 Neuraalilaskennan perusteet

T-61.3040 Signaalien tilastollinen mallinnus

T-61.3050 *Machine learning: basic principles*

T-61.5010 Informaation visualisointi

T-61.5020 Luonnollisen kielen tilastollinen mallinnus

T-61.5030 Neuraalilaskennan jatkokurssi

T-61.5050 Suurkapasiteettimittausten bioinformatiikka

T-61.5060 Tiedon louhinnan algoritmiset menetelmät

T-61.5070 Tietokonenäkö

T-61.5080 *Signal processing in neuroinformatics*

T-61.5090 *Image analysis in neuroinformatics*

T-61.5100 Digitaalinen kuvankäsittely

T-61.5110 Biologisten verkkojen mallintaminen

T-61.5120 Laskennallinen genomiikka

T-61.5130 *Machine learning and neural networks*

T-61.5140 *Machine learning: advanced probabilistic methods*

T-61.5900 Informaatiotekniikan erikoistyö

T-61.6010 Informaatiotekniikan erikoiskurssi I

T-61.6020 Informaatiotekniikan erikoiskurssi II

T-61.6030 Informaatiotekniikan erikoiskurssi III

T-61.6040 Informaatiotekniikan erikoiskurssi IV

T-61.6050 Informaatiotekniikan erikoiskurssi V

T-61.6060 Informaatiotekniikan erikoiskurssi VI

T-61.6070 Bioinformatiikan erikoiskurssi I

T-61.6080 Bioinformatiikan erikoiskurssi II

T-61.6090 Kieliteknologian erikoiskurssi

Käytäntöä

- Kurssin kotisivu `http://www.cis.hut.fi/Opinnot/T-61.2010/`; täydentyy kurssin edetessä
- Ilmoittaudu TOPI:n kautta `http://wwwtopi.hut.fi`
- Luennot keskiviikkoisin ja torstaisin klo 14-16, sali T1
- Laskuharjoitukset (2 h / viikko) Keskiviikkoisin klo 12-14 ja perjantaisin klo 14-16 salissa T1 (vaihtoehtoiset ryhmät). Laskuharjoitusten assistenttina toimii DI Ulpu Remes.
- Harjoitustehtävät suomeksi ja englanniksi löytyvät kurssin kotisivulta. Myös ratkaisut tulevat kotisivulle.
- Matlab-harjoitustyö (tehtävissä pienryhmissä tai yksin; arvioitu työmäärä 2 h/ viikko) on kiinteä osa kurssia ja sen voi tehdä itsenäisesti tai ohjatusti atk-keskuksen luokassa (DI Ville

Viitaniemi). Harjoitustyöaiheita on kaksi vaihtoehtoista. Tarkemmat ohjeet Matlab-työstä hieman myöhemmin.

- Ensimmäinen tentti ke 19.12.07 klo 9-12 (sali M), toinen ma 14.1.08 klo 13-16 (salit DEL), sitten keväällä 2008 ja syksyllä 2008. Tenttivaatimukset ja luentomateriaali kotisivuilla.

Kurssin sisältö

1. Johdanto
   - Miksi tällainen kurssi?
   - Käytäntöä
   - Kurssin sisältö
   - Kurssimateriaali
2. Datasta tietoon: mitä dataa, mitä tietoa?
   - Data-analyysin ongelma
   - Mallit ja oppiminen
   - Esimerkkejä
   - Case study: WEBSOM
3. Data vektorina
   - Vektorit, matriisit, etäisyysmitat

- Datan piirreirrotus ja vektorointi
- Dimensionaalisuuden kirous
- Esimerkki piirrreirrotuksesta: PicSOM

4. Vektoridatan tiivistäminen ja dekorrelointi
   - Pääkomponenttianalyysi
   - PCA:n laskeminen on-line-algoritmilla
   - Esimerkkejä
   - DSS-menetelmä: halutunlaisten aikakomponenttien etsiminen
5. Estimointiteorian perusteita
   - Perusjakaumat 1-ulotteisina
   - Yleistys vektoridatalle, d:n muuttujan normaalijakauma
   - Suurimman uskottavuuden periaate
   - Bayes-estimointi

- Regressiosovitus
- Esimerkki regressiosta: neuroverkko
- Esimerkkejä

6. Hahmontunnistuksen perusteita
   - Johdanto
   - Hahmoalueet, erotinfunktio
   - Lähimmän naapurin luokitin
   - Bayes-optimaalinen luokitin
   - Ryhmittelyanalyysi
7. Itseorganisoiva kartta
   - Perusidea
   - Yhteys biologiaan
   - Suppenevuus 1-ulotteisessa tapauksessa

- Käytännön valintoja
- Mihin SOM:ia käytetään?
- Esimerkkejä

8. Hahmojen etsintä diskreetistä datasta
   - Miten muodostetaan hyviä paikallisia kuvauksia datan osista
9. Web-etsintämenetelmien algoritmit
   - Perusongelmat
   - Merkkijonoetsintä
   - Linkkirakenteen ottaminen huomioon
10. Sekvenssien rakenteen etsintä
    - Segmentointiongelma, kustannusfunktiot
    - Dynaamisen ohjelmoinnin periaate
    - Esimerkkejä biologiasta

Kurssimateriaali:

- Luentokalvot ja harjoitustehtävät ratkaisuineen
- Muu materiaali ilmoitetaan myöhemmin. Se tulee olemaan saatavilla salasanan takana, koska siihen sisältyy copyrightin alaisia sivuja, joita ei missään nimessä saa jakaa edelleen.