HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos
PL 68 (Gustaf Hällströmin katu 2b)
00014 HELSINGIN YLIOPISTO

582631 Johdatus koneoppimiseen
Laskuharjoitus 3
ke 12.11.2008 klo 10-12
sali C221

---

1. Toteuta k lähintä naapuria -luokitin. Voit olettaa, että syötteiden piirteet ovat numeerisia, ja että etäisyysfunktiona käytetään euklidista etäisyyttä. Generoi jotain yksinkertaista dataa, ja testaa, että menetelmäsi näyttää toimivan oikein. (Vihje: Katso ensin seuraavasta tehtävästä, millä tavalla talletettuun dataan luokitinta on tarkoitus soveltaa.)

2. Tässä tehtävässä sovelletaan edellisen tehtävän k lähintä naapuria -luokitinta dataan, joka löytyy tästä hakemistosta. Opetusdatan x- ja y-komponentit löytyvät tiedostoista `train-x.txt` (yksi syöte per rivi) ja `train-y.txt` (yksi luokka per rivi). Testidata löytyy vastaavasti talletettuna tiedostoista `test-x.txt` (yksi syöte per rivi) ja `test-y.txt`.

   (a) Lataa opetus- ja testidata tiedostoista.

   (b) Luokittele testidata edellisen tehtävän k lähintä naapuria luokittimella. Voit halutessasi kokeilla eri k:n arvoja, aloita vaikka arvosta $k = 3$.

   (c) Kuinka pieneen $0/1$-tappioon pääset testidatalla? Entä opetusdatalla?

   (d) Lisätehtävä: Yritä selvittää, mitä luokittelemasi data on.

   (Jos k lähintä naapuria -luokittelijasi on liian hidas koko datan käsittelyyn, voit valita opetus- ja testidataksi vain osan tiedostoista löytyvästä datasta.)

3. (a) Toteuta k lähintä naapuria -regressio (samoin oletuksin kuin tehtävässä 1). Sovella menetelmää edellisistä harjoituksista tuttuun viiksi-dataan. Kokeile menetelmää dataan eri k:n arvoilla (esim. $k = 1, 5, 20$). Piirrä näin saatujen ennustajien ennusteista kuvaajat.

   (b) Lisätehtävä: Minkä k:n arvon valitsisit? Miksi?

4. Toteuta Naive Bayes -luokittelija. Yksinkertaisuuden vuoksi voit olettaa, että $\mathcal{X} = \{0, 1\}^d$ ja $\mathcal{Y} = \{1, \ldots, l\}$ jollakin $l \in \mathbb{N}$.

5. Tässä tehtävässä luokitellaan dataa, joka löytyy täältä. Data on tallennettu tiedostoihin samoin kuin tehtävässä 2.

   (a) Opi Naive Bayes -luokittelija opetusdatasta.

   (b) Mikä on Naive Bayes -luokittelijan $0/1$-tappio testidatalla? Entä opetusvirhe?

   (c) Kokeile dataan myös tehtävässä 1 toteuttamaasi k lähintä naapuria -luokittelijaa (valitsemallasi k). Kumpi menetelmä vaikuttaa paremmalta tämän ennustusongelman ratkaisemisessa?