SGN-2500 Johdatus hahmontunnistukseen, Kevät 2010
Tentti 8.4.2010 / Jari Niemi

Vastaa kaikkiin viiteen tehtävään 1-5. Jokaisen maksimipistemäärä on 6 pistettä. Ei kirjallisuutta. Tarvittavat kaavat annetaan tehtävien yhteydessä. Funktiolaskin sallittu. Ohjelmoitavaa/graafista laskinta ei saa käyttää. Tehtävät 1-5:

1. YLE:n Taloustutkimus Oy:llä teettämien tuoreiden tutkimusten [1] mukaan:

   - Vihreitä kannattaa 10.4% suomalaisista,
   - Vihreiden kannattajista 91% vastustaa ydinvoiman lisärakentamista ja
   - suomalaisista 51% vastustaa ydinvoiman lisärakentamista.

   Jos (satunnaisesti valittu) suomalainen vastustaa ydinvoiman lisärakentamista, niin millä todennäköisyydellä hän lisäksi kannattaa Vihreitä? (6 p.)
   Ohje: Käytä Bayesin kaavaa $P(F|E) = \frac{P(F)P(E|F)}{P(E)}$.
   Näytä kaikki laskujen vaiheet ja selitä ne ja saamasi vastaus lyhyesti.

2. Ajatellaan kolmen luokan k:n lähimmän naapurin luokittajaa ja allaolevaa 2-dimensioista opetusdataa luokista ω_1, ω_2 ja ω_3:

ω_1:	$[4,4]^T$	$[14,7]^T$	$[18,6]^T$	$[7,4]^T$	$[14,9]^T$	$[16,2]^T$	$[2,16]^T$
ω_2:	$[-7,8]^T$	$[3,11]^T$	$[-2,-1]^T$	$[4,3]^T$	$[3,4]^T$	$[4,-3]^T$	$[0,6]^T$
ω_3:	$[3,2]^T$	$[8,1]^T$	$[3,3]^T$	$[6,3]^T$	$[5,10]^T$	$[3,9]^T$	$[1,7]^T$

   Luokita piste $[4,4]^T$ käyttäen
   (a) lähimmän naapurin sääntöä ja (3 p.)
   (b) seitsemän lähimmän naapurin sääntöä. (3 p.)
   Etäisyydet käyttäen euklidista metriikkaa.
   Avuksi:

$$P(\omega|\mathbf{x}) = \frac{P(\omega)p(\mathbf{x}|\omega)}{p(\mathbf{x})}, \quad P_n(\omega_i|\mathbf{x}) = \frac{k_i}{k}, \quad d(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}.$$

   Näytä kaikki laskujen vaiheet ja selitä ne ja saamasi luokitustulokset lyhyesti.

[1] Lähteet:
http://yle.fi/uutiset/talous_ja_politiikka/2010/04/suurten_puolueiden_kisa_kiristyy_1577576.html
ja
http://yle.fi/uutiset/kotimaa/2010/03/suomalaiset_ovat_yha_nihkeita_uusille_ydinreaktoreille_1566582.html.

3. Olkoot $\mathbf{x_1} = [4,5]^T$, $\mathbf{x_2} = [1,4]^T$, $\mathbf{x_3} = [0,1]^T$ ja $\mathbf{x_4} = [5,0]^T$. Tarkastellaan seuraavia osituksia:

   - $\mathcal{D}_1 = \{\mathbf{x_1}, \mathbf{x_2}\}$, $\mathcal{D}_2 = \{\mathbf{x_3}, \mathbf{x_4}\}$,
   - $\mathcal{D}_1 = \{\mathbf{x_1}, \mathbf{x_4}\}$, $\mathcal{D}_2 = \{\mathbf{x_2}, \mathbf{x_3}\}$,
   - $\mathcal{D}_1 = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}\}$, $\mathcal{D}_2 = \{\mathbf{x_4}\}$ sekä
   - $\mathcal{D}_1 = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}\}$, $\mathcal{D}_2 = \emptyset$ ($\emptyset$ = tyhjä joukko).

   Mitä näistä osituksista k-means -kriteeri suosii? (6 p.)
   Avuksi: $J(\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c\}) = \sum_{D_i \neq \emptyset} \sum_{\mathbf{x} \in \mathcal{D}_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2$.
   Näytä kaikki laskujen vaiheet ja selitä ne ja saamasi vastaus lyhyesti.

4. Suunnittele yksidimensioinen kahden luokan luokitustehtävä, jolla on tasajakautuneet prioritodennäköisyydet ja jatkuvat tasajakautuneet luokkatiheysfunktiot niin, että kyseisen tehtävän Bayesin (minimiluokitus)virhe on 0.1. (6 p.)
   Avuksi:

$$P(\omega|x) = \frac{P(\omega)p(x|\omega)}{p(x)}, \quad p(x) = \frac{1}{b-a}, \text{ kun } x \in [a,b] \text{ ja 0 muulloin,}$$

$$p(x) = \frac{1}{n}, \text{ kun } x \in \{x_1, \ldots, x_n\} \text{ ja 0 muulloin,} \quad P(error) = \int P(error|x)p(x)dx,$$

   missä $P(error|x)$ on todennäköisyys sille, että piirre x luokitetaan väärin.
   Selitä ja perustele vastauksesi yksityiskohtaisesti.

5. Kirjoita essee (300-1000 sanaa) otsikolla *Ohjattu ja ohjaamaton luokitus.* (6 p.)
   Ohje: Käsittele ainakin seuraavat asiat: ohjatun ja ohjaamattoman luokituksen määritelmät, kummankin hyödyt ja haitat toisiinsa verrattuina sekä kummallekin esimerkkejä erityyppisistä estimointi- ja luokitusmenetelmistä (käy läpi lyhyesti: suurin uskottavuus, Parzen, k lähintä naapuria, k-means, sekoitemallit ja näiden virhelähteet).