

Hahmontunnistuksen perusteet

T-61.231, 3ov, L

Syksy 2003

Luennot: **Timo Honkela ja
Kimmo Raivio**

Laskuharjoitukset: **Markus Koskela**

Harjoitustyö: **Matti Aksela**

1.	FOREIGN STUDENTS	1
2.	YLEISTÄ KURSSISTA	2
2.1	Kurssin suorittaminen	2
2.2	Ilmoittautuminen	2
2.3	Tiedotukset	2
2.4	Luennot	3
2.5	Laskuharjoitukset	3
2.6	Kirja	4
2.7	Luentomonisteet	5
2.8	Suhde vanhaan Tik-61.131-kurssiin	5
2.9	Tentti	6
2.10	Harjoitustehtävä	7
3.	JOHDANTO	8
3.1	Johdattelevia esimerkkejä	8
3.2	Mitä on hahmontunnistus?	15
3.3	Sovelluksia	16
3.4	Tunnistusjärjestelmä	17
3.5	Esikäsittely ja normalisointi	20

3.6	Piirrevalinta	25
3.7	Tunnistusmenetelmät	32
3.8	Arviointi	38

1. FOREIGN STUDENTS

Lectures and slides are in Finnish.

Exercises are held in Finnish but the problems and their solutions are published in English. Course assistants do speak English.

To pass this course you need to 1) pass the exam and 2) do the course assignment.

For more information, see

http://www.cis.hut.fi/Opinnot/T-61.231/index_en.shtml

2. YLEISTÄ KURSSISTA

2.1 Kurssin suorittaminen

Kurssin suorittaminen sisältää pakollisen harjoitustehtävän ja tentin.

2.2 Ilmoittautuminen

Ilmoittautukaa kurssille [www-topin](#) avulla.

2.3 Tiedotukset

Kurssista tiedotetaan webissä <http://www.cis.hut.fi/Opinnot/T-61.231>, ryhmässä <news://nntp.tky.hut.fi/opinnot.tik.informaatiotekniikka> sekä Informaatiotekniikan laboratorion ilmoitustaululla kolmannen kerroksen aulassa B-käytävän suulla.

2.4 Luennot

Luennot pidetään maanantaisin kello 10–12 salissa T2.

Luennoitsijat: prof. (ma) Timo Honkela ja tekn.tri Kimmo Raivio (timo.honkela@hut.fi, kimmo.raivio@hut.fi).

Luentokalvot ovat luennon jälkeen nähtävillä osoitteessa <http://www.cis.hut.fi/Opinnot/T-61.231/luentorunko.shtml>.

Luennoitsijan tapaamisaika luentojen jälkeen. Pitemmistä tapaamisista kannattaa sopia etukäteen sähköpostitse ja ne voidaan järjestää informaatiotekniikan laboratorion kirjastossa tai kahvihuoneessa.

2.5 Laskuharjoitukset

Laskuharjoitukset keskiviikkoisin kello 12–14 päärakennuksen K-salissa 24.9.20 alkaen.

Harjoitukset pitää DI Markus Koskela (markus.koskela@hut.fi).

Harjoitustehtävät ovat jo ennakoon nähtävillä osoitteessa
<http://www.cis.hut.fi/Opinnot/T-61.231/laskarit2003.shtml>.

2.6 Kirja

Kurssikirjaksi on kaksi vaihtoehtoa:

1) Sergios Theodoridis, Konstantinos Koutroumbas:

Pattern Recognition, Academic Press, 1998.

2) Robert Schalkoff:

Pattern Recognition – Statistical, Structural and Neural Approaches,

John Wiley & Sons, 1992. (Kurssin vanha oppikirja)

Kirjat ovat saatavilla eri kirjakauppojen kautta. Theodoridoksen kirjasta on juuri äskettäin ilmestynyt uusi painos, jonka eroja vuoden 1998 painokseen ei ole päästy vertaamaan. Opetus perustuu vuoden 1998 kirjaan mutta mitä ilmeisimmin kannattaa tilata vuoden 2003 versio. Tentissä ei kysytä asioita, joita ei ole esitetty uusimmassa painoksessa (tai Schalkoffin kirjassa tai

luennoilla annetussa lisämateriaalissa).

Kirjojen tutustumiskappaleet ovat nähtävillä Informaatiotekniikan laboratorion kirjastossa ja laboratorion sihteerin Tarja Pihamaan huoneessa B326 olevassa harmaassa peltisessä vetolaatikostossa. Tutustumiskappale pitää palauttaa!

2.7 Luentomonisteet

Laskuharjoitukset ratkaisuihin ja luentokalvot ilmestyvät opetusmonisteina. Älä turhaan tulosta WWW-sivuilla olevaa materiaalia!

Materiaalia voi myös lainata luennoitsijalta ja assarilta itse kopioitavaksi.

2.8 Suhde vanhaan Tik-61.131-kurssiin

Kurssi korvaa vanhan samannimisen kurssin Tik-61.131, jonka laajuus oli 2,5 ov ja joka ei sisältänyt pakollista harjoitustyötä.

2.9 Tentti

Tenttejä järjestetään kaksi: ensimmäinen tiistaina 12. joulukuuta ja toinen tammi-helmikuussa. Lisäksi vuoden 2004 syksyn alussa järjestään tentti.

Tentissä on 4 tehtävää à 6-12 pistettä, yhteensä 30 pistettä. Yksi tehtävistä on essee (6), yksi koostuu useista pienistä sanallisista tehtävistä (12) ja loput ovat laskutehtäviä (yht. 12).

Tentissä saa olla mukana matemaattinen kaavakokoelma ja tavallinen funktiolaskin. (Kumpaakaan tuskin tarvitsee.)

2.10 Harjoitustehtävä

Kurssin suoritukseen kuuluu pakollinen harjoitustehtävä, joka arvostellaan hyväksytty/hylätty-periaatteella. Harjoitustehtävä on saatava hyväksytysti läpi tammikuun 31. päivään mennessä.

Huom! Harjoitustyön tulee olla tehtynä ennen kevään ja ensi vuoden alkusyksyn tenttiin osallistumista.

Harjoitustehtävän aiheet ovat esillä osoitteessa

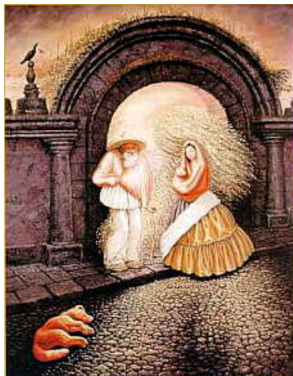
<http://www.cis.hut.fi/Opinnot/T-61.231/harjtyo2003.shtml>.

Harjoitustyöhön liittyvissä asioissa ottakaa yhteyttä kurssin assistenttiin (matti.aksela@hut.fi).

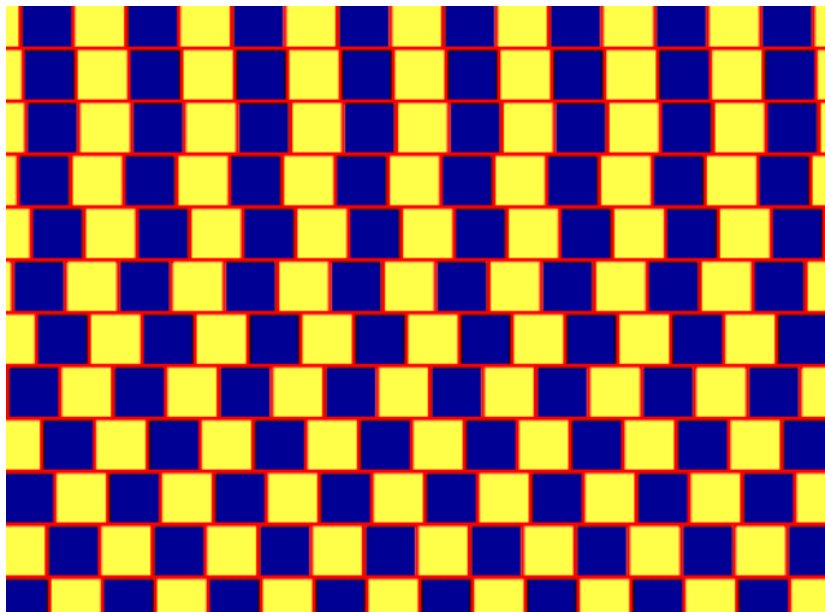
3. JOHDANTO

3.1 Johdattelevia esimerkkejä

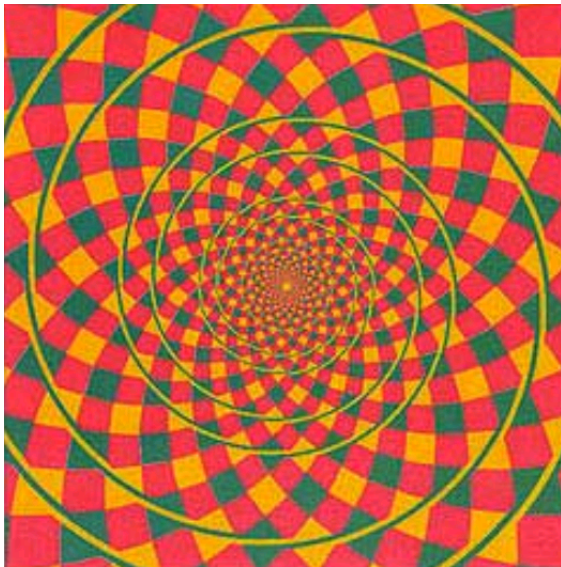
1. Kuinka monta ihmishahmoa näet kuvassa?



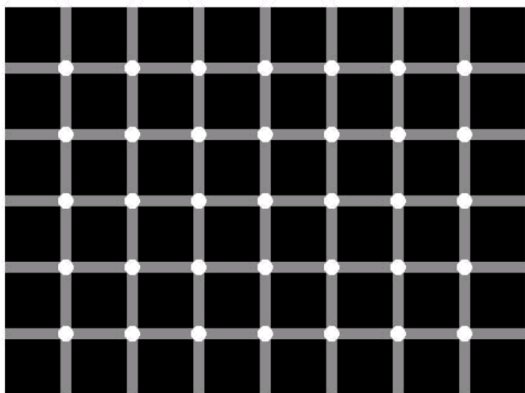
2. Ovatko viivat yhdensuuntaisia?



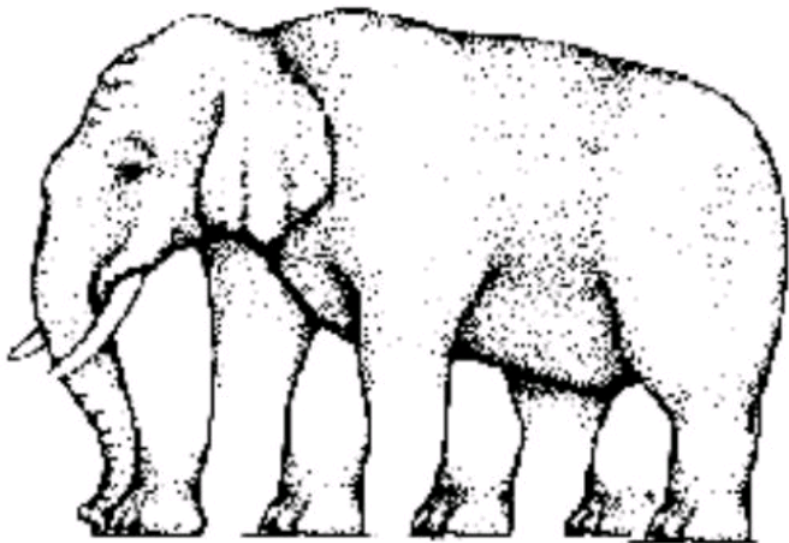
3. Mitä erikoista on spiraalissa?



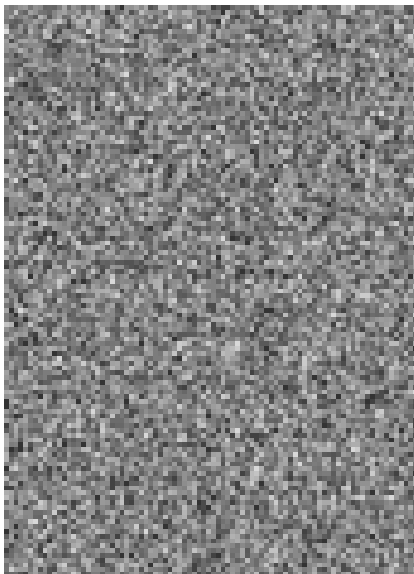
4. Onko risteyskohdissa mustia vai valkoisia pisteitä?



5. Kuinka monta jalkaa norsulla on?



6. Näetkö kuvassa numeron 3?



Vastauksia

1. Esimerkiksi 9 on varsin hyvä vastaus.
2. Kyllä.
3. Kyseessä ei ole spiraali vaan joukko ympyröitä.
4. Valkoisia. Näkökentän laidoilla olevat näyttävät tummilta.
5. Monta.
6. Kuva on pelkkää kohinaa, mutta ei ole tyystin mieletöntä sanoa, että siinä näkee (esimerkiksi) numeron kolme. Lisää tietoa asiaan liittyvästä Gosselinin ja Schynsin (ym.) tutkimuksesta löytyy esimerkiksi sivulta *[http : //www.perceptionweb.com/ecvp02/0163.html](http://www.perceptionweb.com/ecvp02/0163.html)*

3.2 Mitä on hahmontunnistus?

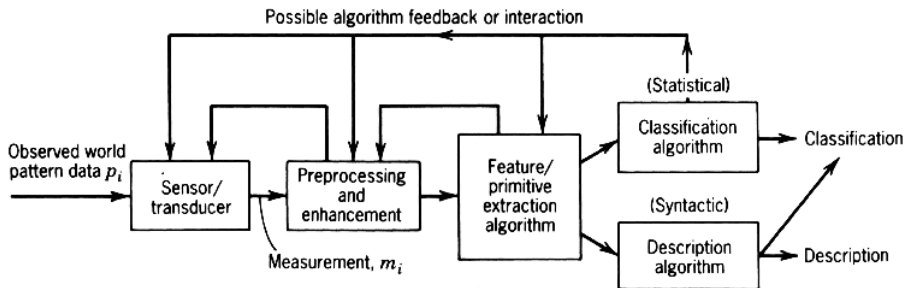
- Hahmontunnistus on tieteenala, jossa pyritään luokittelemaan tehtyjä havaintoja (hahmot) erilaisiin kategorioihin (luokat) tai löytämään hahmojen välisiä suhteita.
- Keskeisiä ongelmia ovat tiedon (mittaukset, havainnot) pelkistys (esikäsittely), tiedon kuvaus (piirreirrotus), ja tiedon luokittelu tai muu kuvaus (tunnistus)
- Hahmontunnistukseen liittyviä tieteenaloja: digitaalinen signaalinkäsittely, tekoäly, neuraalilaskenta, optimointioppi, estimointiteoria, sumea logiikka, strukturaalinen mallintaminen, formaalit kielet, kieliteknologia

3.3 Sovelluksia

- Digitaalisten kuvien käsittely, tietokonenäkö, esim. satelliittikuvien tulkinta
- Puheen tunnistus, painetun ja käsin kirjoitetun tekstin tunnistus
- Henkilöiden tunnistus puheen, sormenjälkien, nimikirjoituksen tai silmän iriksen perusteella
- Lääketieteellisten mittausten analysointi
- Tietokoneavustettu päätöksenteko
- Laadunvalvonta, lajittelu, esim. reikä paperirainassa, palautuspullojen käsittely.
- Postin ja pankkitositteiden automaattinen käsittely
- ... ja vaikka mitä muuta! - Esimerkkejä käydään läpi seuraavalla luenolla.

3.4 Tunnistusjärjestelmä

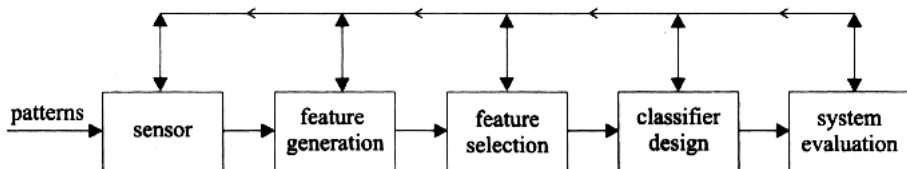
Tyypillinen tunnistusjärjestelmä:



- Edellisten lisäksi tarvitaan opetus- ja testidataa
- Esikäsittelyn tarkoituksena on korostaa oleellista mittausinformaatiota ja helpottaa piirreirrotusta

- Piirteet ja niiden väliset suhteet ovat hahmon representaatio, jonka perusteella hahmo voidaan luokitella tai kuvata
- Hahmon representaatio voi olla esim. vektori, matriisi, puu, graafi, merkkijono
- Luokitin voi perustua erilaisiin laskennallisiin tekniikoihin (tilastolliset ja geometriset menetelmät, neuroverkot, jne...)
- Kysymykset joihin etsitään vastauksia: “Mitkä hahmoista ovat keskenään samankaltaisia?”, “Kuinka monta ja millaisia ryhmiä hahmot muodostavat?”

Kaaviokuva suunnittelusta:



Tunnistusjärjestelmän suunnittelu on iteratiivinen prosessi, jossa keskeisiä kysymyksiä ovat:

- Mitä mitataan ja millä laitteilla?
- Millaisia piirteitä mittauksista voidaan laskea, mitkä piirteet kuvaavat hyvin hahmoja?
- Mitkä piirteistä ovat tehtävän kannalta oleellisia? Mitkä ovat parhaita? Kuinka monta tarvitaan?
- Millainen luokitteluteknikka soveltuu tehtävään parhaiten?
- Kuinka hyvin luokitin toimii opetus- ja testidatalla, kuinka yleisiä saadut tulokset ovat?

3.5 Esikäsittely ja normalisointi

- Onnistunut esikäsittely helpottaa piirreirrotusta ja parantaa luokittelumenetelmän toimivuutta: parempia piirteitä, nopeampi oppiminen, yleisempiä tuloksia
- Esikäsittely- ja normalisointimenetelmien valinta riippuu siitä, millaisia piirteitä halutaan laskea ja mitä luokittelutekniikoita käytetään
- Menetelmien valinta riippuu erittäin paljon sovelluksesta, erilaisia menetelmiä löytyy mm digitaalisen signaalin- ja kuvankäsittelyn piiristä
- Tyypillisiä menetelmiä: ns. outlier ja puuttuvien mittausten käsittely (hylkäys, heuristinen tai tilastollinen päättely), varianssien ja odotus-tusarvojen normalisointi, pääkomponenttianalyysi, arvoalueen skaalaus (esim. lineaarinen, softmax)

- Softmax-skaalaus:

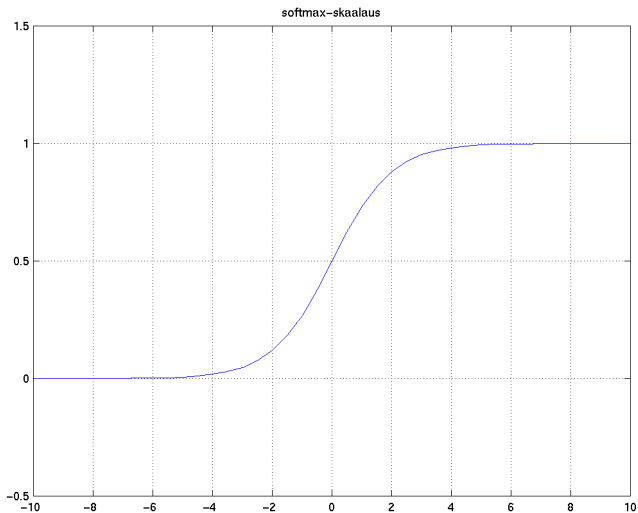
$$\bar{x}_k = \frac{1}{N} \sum_{i=1}^N x_{ik}, k = 1, 2, \dots, l$$

$$\sigma_k^2 = \frac{1}{N-1} \sum_{i=1}^N (x_{ik} - \bar{x}_k)^2$$

$$y = \frac{x_{ik} - \bar{x}_k}{r\sigma_k}$$

$$\hat{x}_{ik} = \frac{1}{1 + \exp(-y)}$$

(k .:n piirteen käsittely, piirteitä yhteensä l kappaletta) Skaalaus rajoittaa arvot välille $[0, 1]$. Parametri r säätelee aluetta, jossa skaalaus on lineaarinen.



- Pääkomponenttialyysi käsitellään tarkemmin laskareissa

- Erilaisia hahmojen vääristymiä:

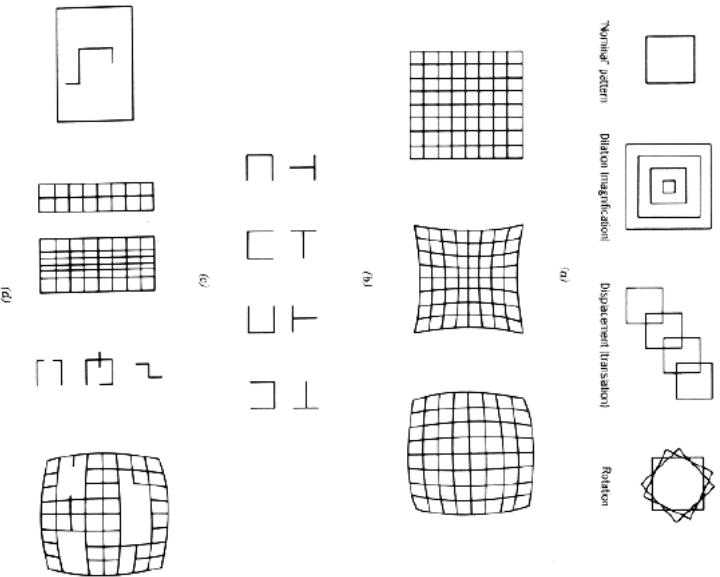


Figure 4. Examples of pattern distortions (visual patterns).

- (a) Geometric distortions of visual (image) patterns-square.
- (b) Geometric distortions of grid pattern.
- (c) Geometric distortions of character patterns
- (d) More extreme pattern distortions (missing parts and extra parts) using the patterns of parts (a) to (c).

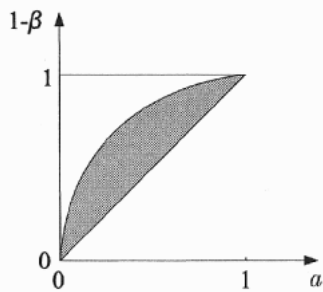
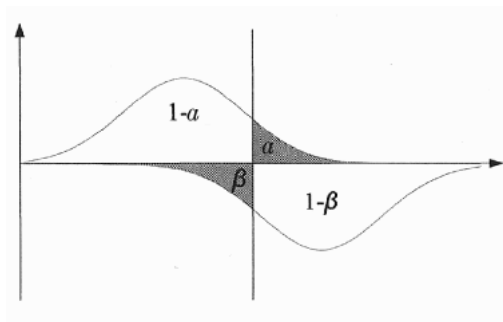
ks. myös *[http : //www.dai.ed.ac.uk/HIPR2/geomops.htm](http://www.dai.ed.ac.uk/HIPR2/geomops.htm)*

3.6 Piirrevalinta

- Piirteet valitaan siten, että ne kuvaavat hyvin tarkasteltavia hahmoja ja saavat *samankaltaisia arvoja luokkien sisällä ja erilaisia arvoja luokkien välillä*
- Piirteet voivat olla symbolisia tai numeerisia, jatkuva-, diskreetti- tai binääriarvoisia
- Piirteet pyritään valitsemaan siten että ne ovat invariantteja erilaisille vääristymille ja mittausolosuhteiden muutoksille (esim. skaalaus, rotaatio, translaatio, valaistus)
- Piirteiden valinta riippuu sovellutuskohteesta!
- Piirteiden valinnalla on ratkaiseva vaikutus luokittelun onnistumiseen
- Hyvien piirteiden valintana sopivia menetelmiä: tilastollinen testaus, "receiver operating characteristic (ROC) curve", erilaiset luokkien separoituvuusmitat

ROC-käyrä:

- Kertoo kuinka päällekkäisiä luokkien jakaumat ovat yhden piirteen suhteen
- Oletus: hahmo kuuluu luokkaan ω_1 , jos piirre $x < \theta$, tai, luokkaan ω_2 , jos $x \geq \theta$, kokeillaan θ :lle erilaisia arvoja
- α (β) on tn, että luokitellaan hahmo virheellisesti luokkaan ω_2 (ω_1)
- Kun jakaumat ovat täysin päällekkäiset, $\alpha = 1 - \beta$
- Mitä vähemmän päällekkäisyyttä, sitä suurempi alue ROC-käyrän ja edellä mainitun suoran välillä



“Ambiguity function”:

- Useamman kuin kahden luokan separoituvuusmitta laskettuna useamman kuin yhden piirteen suhteen
- Jaetaan piirteen arvoalue K :ksi intervalleiksi Δ_j , $P(\Delta_j)$ on tn, että kyseiseltä intervallilta on havaintoja, $P(\omega_i|\Delta_j)$ on luokan ehdollinen tn kyseisessä intervallissa, luokkia M kappaletta

$$A = - \sum_{i=1}^M \sum_{j=1}^K P(\Delta_j) P(\omega_i|\Delta_j) \log_M(P(\omega_i|\Delta_j))$$

- Kun luokkien jakaumat ovat täysin päällekkäiset, $A = 1$, kun jakaumat ovat täysin erilliset, $A = 0$
- Tarvittavat tn:t on helppo estimoida datasta
- Todistus laskareissa

Piirrejoukon valinta

Luokkien separoituvuutta eri piirrevalinnoilla voidaan siis mitata, mutta kuinka valita paras l :n piirteen osajoukko m :stä piirteestä?

- Vaihtoehtoja on runsaasti, kun l :kään ei tunneta:

$$\sum_{i=1}^m \binom{m}{i}$$

eli kaikkia vaihtoehtoja ei yleensä voida evaluoida

- Eräs piirteiden välisiin korrelaatioihin perustuva tekniikka:
 - Laske piirteiden väliset korrelaatiot:

$$\rho_{ij} = \frac{\sum_{n=1}^N x_{ni}x_{nj}}{\sqrt{\sum_{n=1}^N x_{ni}^2 \sum_{n=1}^N x_{nj}^2}}$$

- Valitse luokkien separoituvuusmitta C
- Laske C kaikille piirteille erikseen, järjestä paremmuusjärjestykseen, valitse parhaan C :n tuottama piirre x_{i_1}
- Valitse seuraava piirre x_{i_2} siten että

$$i_2 = \arg \max_j \{\alpha_1 C(j) - \alpha_2 |\rho_{i_1 j}|\}, j \neq i_1$$

α_1 ja α_2 ovat 'hihavakiota'

- Valitse x_{i_k} , $k = 1, \dots, l$ siten että

$$i_k = \arg \max_j \left\{ \alpha_1 C(j) - \frac{\alpha_2}{k-1} \sum_{r=1}^{k-1} |\rho_{i_r, j}| \right\}$$
$$j \neq i_r, r = 1, \dots, k-1$$

- Variaatio: käytetään kahta separoituvuusmittaa. Silloin

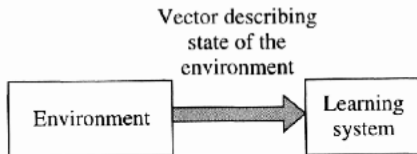
$$i_k = \arg \max_j \left\{ \alpha_1 C_1(j) + \alpha_2 C_2(j) - \frac{\alpha_3}{k-1} \sum_{r=1}^{k-1} |\rho_{i_r, j}| \right\}$$
$$j \neq i_r, r = 1, \dots, k-1$$

3.7 Tunnistusmenetelmät

Tunnistusmenetelmät voidaan jakaa erilaisiin ryhmiin. Yksi tärkeä jakoperuste on oppimisperiaate:

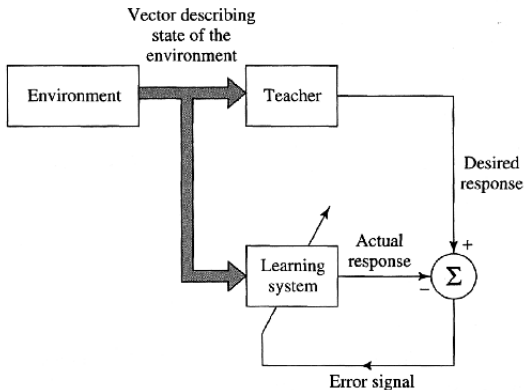
Ohjaamaton oppiminen

- Hahmojen *luokkia ei tiedetä* etukäteen
- Tavoitteena on muodostaa hahmoista ryhmiä, joiden sisällä hahmot ovat samankaltaisia ja joiden välillä on selkeitä eroja (klusterointi)
- Optimoitava funktio on klusteroinnin onnistumista kuvaava mitta
- Aina ei tiedetä edes ryhmien lukumäärää



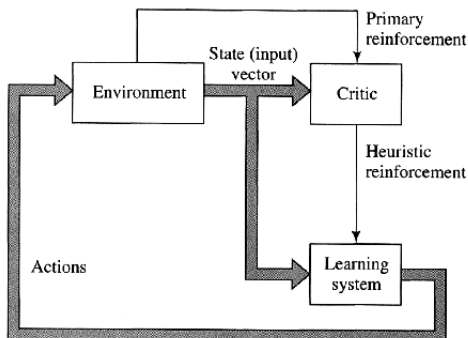
Ohjattu oppiminen

- Hahmojen *luokat tunnetaan* etukäteen
- Tavoitteena on muodostaa kuvaus piirreavaruudesta luokka-avaruuteen
- Optimoitava funktio perustuu kuvauksessa tapahtuviin virheisiin



Vahvistusoppiminen

- “Reinforcement learning”
- Ohjatun ja ohjaamattoman oppimisen välimuoto: luokkia ei tunneta, mutta onnistumisesta saadaan positiivista ja epäonnistumisesta negatiivista palautetta



Jaottelu voi perustua myös käytettyihin laskennallisiin tekniikoihin

Geometriset menetelmät

- Piirreavaruus jaetaan joko lineaarisesti tai epälineaarisesti osiin, jokainen osa vastaa tiettyä luokkaa
- Esim. hypertasot, Fisherin diskriminantti, SVM

Neuroverkkomenetelmät

- Ns. “black box”-menetelmä, joka tekee epälineaarisen kuvauksen piirreavaruudesta luokka-avaruuteen
- Järjestelmän tulkinta hankalaa
- Esim. MLP, RBF, SOM

Tilastolliset menetelmät

- Piirteitä käsitellään tilastollisina muuttujina
- Tunnistus perustuu riskin odotusarvon minimointiin
- Menetelmissä tarvittavien tnjakaumien estimointi voi olla hankalaa

Syntaktiset menetelmät

- Hahmot esitetään esim. formaalin kielen tuottamina merkkijonoina
- Tunnistus perustuu jäsentämiseen
- Jos jäsenitys ei onnistu, on vaikea erotella hyviä ja huonoja vastauksia

Rakenteelliset menetelmät

- Hahmot esitetään esim. graafien avulla
- Tunnistus perustuu graafien vertailuun
- Suurien graafien vertailu laskennallisesti raskasta

Malleihin perustuvat menetelmät

- Luokat esitetään tyypillisten, mallihahmojen (prototyypit) avulla
- Tuntematonta hahmoa verrataan malleihin ja tunnistus tehdään samankaltasimpien mallien perusteella
- Vertailu laskennallisesti raskasta ja muistinkulutus korkea, jos malleja on runsaasti

3.8 Evaluointi

Tunnistusjärjestelmää evaluoidaan, jotta tiedetään kuinka se suoriutuu tositalanteessa. Lisäksi evaluointi on läheisesti kytketty kaikkiin muihin järjestelmän suunnitteluvaiheisiin.

Virhetn:n estimointi:

- tutkitaan kuinka monta virhettä syntyy testidatalla luokkaa kohden
- Kun oletetaan, että hahmot ovat toisistaan riippumattomia, luokan a *priori* tn on $P(\omega_i)$ ja luokasta ω_i on N_i näytettä, tn että k_i näytettä tunnistetaan virheellisesti on

$$P(k_i \text{ virhettä}) = \binom{N_i}{k_i} P_i^{k_i} (1 - P_i)^{N_i - k_i}$$

- Yksittäisen luokan ω_i todellista virhetn:ttä P_i ei tunneta

- Yksittäiselle luokalle suurimman uskottavuuden (ML) estimaatti virhetn:lle on

$$\hat{P}_i = \frac{k_i}{N_i}$$

- Kun luokkia on M kappaletta, ML- estimaatti kokonaisvirhetn:lle on

$$\hat{P} = \sum_{i=1}^M P(\omega_i) \frac{k_i}{N_i}$$

- ML-estimaatti on harhaton ja sen varianssi on

$$\sigma_{\hat{P}}^2 = \sum_{i=1}^M P^2(\omega_i) \frac{P_i(1 - P_i)}{N_i}$$

eli estimaatti on sitä tarkempi, mitä enemmän luokista on näytteitä

- Todistuksesta tarkemmin laskareissa

- Jos halutaan, että estimaatti poikkeaa korkeintaan ϵ :n verran todellisesta virhetn:stä tietyllä riskillä a , mikä on testidatan vähimmäismäärä N ?

$$\text{prob}\{P \geq \hat{P} + \epsilon(N, a)\} \leq a$$

- Kun oletetaan, että $\epsilon(N, a) = \beta P$, voidaan edellisestä yhtälöstä approksimoida ylä- ja alarajat N :lle.
- Tyypillisesti $a = 0.05$ ja $\beta = 0.2$. Silloin

$$N \approx \frac{100}{P}$$

esim. $P = 0.01$, $N = 10\,000$

- Huomaa, tulos ei riipu luokkien lukumäärästä
- Jos testidatan näytteet eivät ole riippumattomia, testidataa tarvitaan enemmän

Opetus- ja testidata

Mitä enemmän näytteitä käytetään opetuksessa, sitä paremmin tunnistusjärjestelmä pystyy yleistämään, pienellä testijoukolla vaarana “ylioppiminen”

Mitä enemmän näytteitä käytetään evaluointiin, sitä luotettavammat estimaatit virhetn:ille

Rajoitettu määrä dataa, kuinka tehdään jako opetus- ja testidataksi?

Erilaisia lähestymistapoja:

- Käytetään samaa dataa sekä opetukseen että testaukseen ('resubstitution'). Tulokset ylioptimistisia (seuraava kuva)
- Erilliset opetus- ja testidatat. Tulokset realistisempia, sopiva jako löytyy esim. kokeilemalla

- “Leave-One-Out” . Opetus tehdään kaikilla paitsi yhdellä näytteellä, jota käytetään testaukseen. Toistetaan kaikilla mahdollisilla osajoukoilla. Laskennallisesti raskasta!
- “Leave-k-Out” . Variaatio edellisestä. Ei käydä välttämättä läpi kaikkia osajoukkoja

