1. Johdatus tilastolliseen hahmontunnistukseen

Hahmontunnistus (pattern recognition) tarkoittaa menetelmiä ja algoritmeja, joiden avulla datasta pystytään tunnistamaan säännönmukaisuuksia (hahmoja, rakenteita).

Hahmontunnistus on oleellinen osa älykästä järjestelmää, joka pyrkii havainnoimaan toimintaympäristöään ja toimimaan havaintojensa mukaisesti. Mittausdataa voidaan saada ympäristöstä moninaisilla sensoreilla mitaten esimerkiksi suureita: lämpötila, ilmankosteus, äänet (mikrofoni), näkymät (kamera), paikka, nopeus, kiihtyvyys, etäisyys (ultraääni), ilmanpaine, kosketus, virta, jännite.

Käyttökohde voi liittyä mihin tahansa sovellukseen, jossa on tarkoituksenmukaista tulkita automaattisesti mittasignaaleja ja päätellä siten jotain tarkasteltavasta kohteesta.

- automaattinen puheentunnistus, puhujan tunnistus äänestä, äänenvärianalyysi
- konenäkö (tuotteiden visuaalisen laadun tarkastus, robottinäkö)
- kuva-analyysi (sormenjälkitunnistus, henkilötunnistus kasvokuvasta)
- visuaalinen valvonta videokameralla (liikenneristeykset, taksiasemat, varastot)
- spektrimittaukset (kaukokartoitus, pitoisuusmittaukset, kasvien elinvoimaisuus)
- tilanneherkkä laskenta (laite/ohjelma mukautuu ympäristön ja käyttäjän tilaan)
- tietohakujärjestelmät (tietokantojen sisällön automaattinen analyysi)
- biosysteemien analyysi (EEG, EKG,..)
- optinen merkintunnistus (OCR)

1.1. Esimerkki: konenäköön perustuva kalan tunnistus kalanpakkaamossa

Liukuhihnan viereen asennettu digitaalinen kamerajärjestelmä luokittelee (classify) jokaisen kalan automaattisesti; pakkaajarobotti sijoittaa sitten kalat oikeisiin laatikoihin.
Tunnistus perustuu kalalajien muotojen mallintamiseen joukolla niitä luonnehtivia piirteitä (feature). Varsinainen luokittelu vaatii erityiset algoritmit.

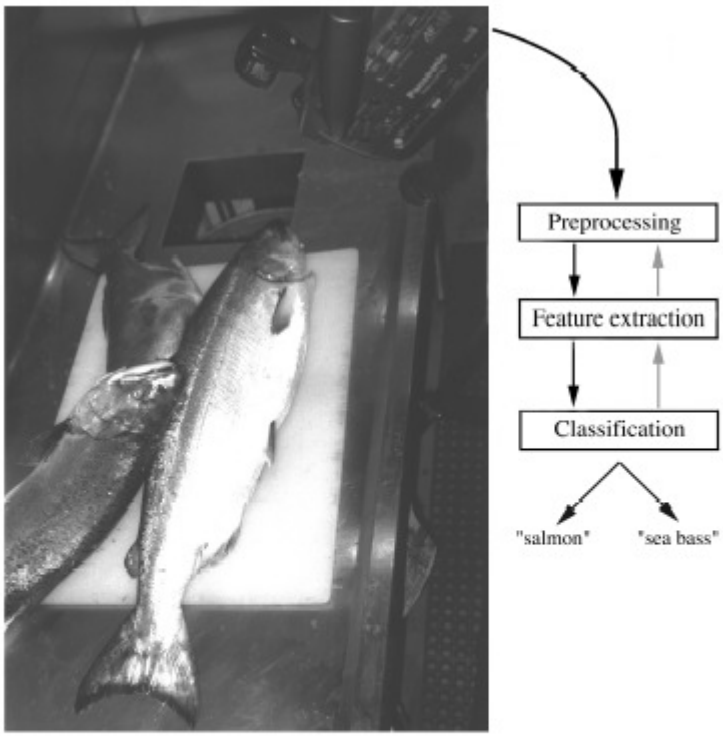


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Esikäsittely (preprocessing): valoisuusvaihtelun kompensointi, kalan erottaminen taustasta (segmentointi), kohinan poisto
- Piirteenlaskenta (feature extraction): kalan luonnehdintojen mittaaminen (pituus, kirkkaus)
- Luokittelu (classification): kalan luokittelu määriteltyihin luokkiin luonnehdintojen avulla

Hyvien piirteiden valinta (feature selection) on kriittisen tärkeää luokittimen suorituskyvylle. Tarkastellaan kalalajien erottumista pituuden ja kirkkauden avulla. Apuna käytetään testijoukkoa kaloja, joista mitatut piirrejakaumat piirretään samaan kaavioon:

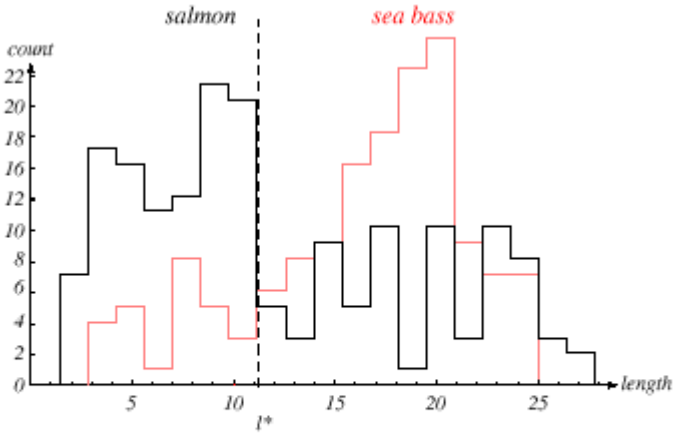


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

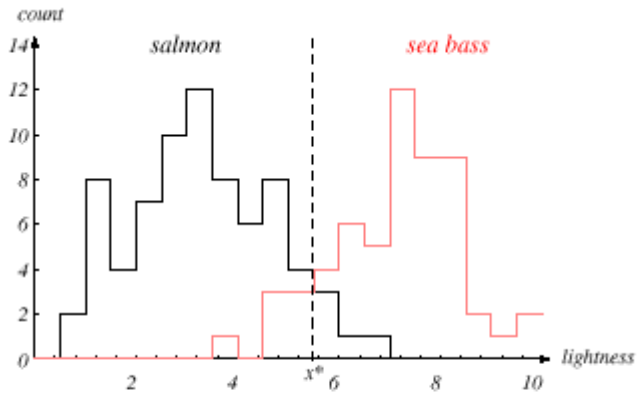


FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Kuten kuvista nähdään, luokat eivät erotu toisistaan täydellisesti parhaallakaan rajakohdan valinnalla, jota luokittelija pyrkii noudattamaan. Tästä seuraa luokitteluvirheitä, joten pakkauslaatikoihin sijoitetaan virheellisesti molempia kalalajeja.

Päätösteorian (decision theory) avulla on mahdollista johtaa optimaaliset rajakohdat. Luokitteluvirheiden suhteellisen osuuden minimoiminen on usein käytetty kriteeri optimoinnissa. Toinen kriteeri on virheellisen (oikeellisen) luokittelun kustannus/haitta (cost). Esimerkiksi asiakasta ei paljoa haittaa (pieni kustannus), jos särkien joukkoon päätyy muutama lohikala, mutta meteli nousee särkien joutuessa lohipakkaukseen (suuri kustannus)! Tällöin tulisikin minimoida kokonaiskustannus. Toinen hyvä esimerkki tästä on syövän diagnosointi terveyskeskuksessa: kustannus/haitta luokitella terve sairaaksi on pienempi kuin sairaan luokittelu terveeksi. Miksi? Sairaaksi arveltu terve lähetetään jatkotutkimuksiin, mistä aiheutuu hieman rahallisia kustannuksia, mutta todellinen tilanne lopulta paljastuu. Terveeksi arveltu sairas sensijaan passitetaan kotiin, mikä saattaa johtaa jopa hengen menetykseen. Siis kokonaiskustannuksen minimointi on tässä sovelluksessa järkevämpi optimointikriteeri kuin virheellisten luokitusten lukumäärä, koska erityyppiset virheet maksavat eri määrän!

Koska esimerkkimme piirteet eivät kyenneet erottelemaan kaloja toisistaan riittävän hyvin eikä parempiakaan piirteitä keksitä, niin koetetaan luokittelua käyttäen molempia piirteitä yhtä aikaa. Tämä voidaan toteuttaa eri tavoilla, nyt sovellamme tilastollista lähestymistapaa, jossa piirteet kootaan vektoriksi, piirrevektoriksi (feature vector). Saamme siis kalan pituudesta x_1 ja kirkkaudesta x_2 2-ulotteisen vektorin **x** (lihavoitu), joka sijaitsee 2-ulotteisessa piirreavaruudessa (feature space):

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Jokainen kala siis mallinnetaan mittaustiedoista tällaisella piirrevektorilla ja siitä tulee yksi piste piirreavaruuteen. Sanotaan myös, että kyseinen piirrevektori on kalan esitystapa (representation) tässä hahmontunnistusjärjestelmässä. Joukosta kaloja siis muodostuu pisteryhmä (cluster) kalojen yksilöllisten erojen aiheuttaessa hajontaa, ja toivottavasti eri kalalajit muodostavat erilliset ryhmät.

Luokittelijaa suunniteltaessa piirreavaruus tulisi nyt jakaa kahteen osaan siten, että kalalajit erottuisivat toisistaan mahdollisimman virheettömästi. Alla olevassa kuvassa jakaminen pyritään tekemään piirreavaruuden halkaisevalla suoralla. Kuten näkyy, luokat erottuvat toisistaan varsin hyvin kun suora on onnistuttu sijoittamaan oikeaan kohtaan. Erottavaa käyrää kutsutaan päätösrajaksi (decision boundary), tai moniulotteisessa tapauksessa päätöspinnaksi.

Päätössääntö (decision rule) on: luokittele kala meriahveneksi, jos piste sijoittuu päätöspinnan yläpuolelle, ja muulloin loheksi.
Jos piste osuu päätöspinnalle, on sama kummaksi luokitellaan, koska piirteemme eivät kykene tekemään eroa, joten olkoon vaikkapa lohi.

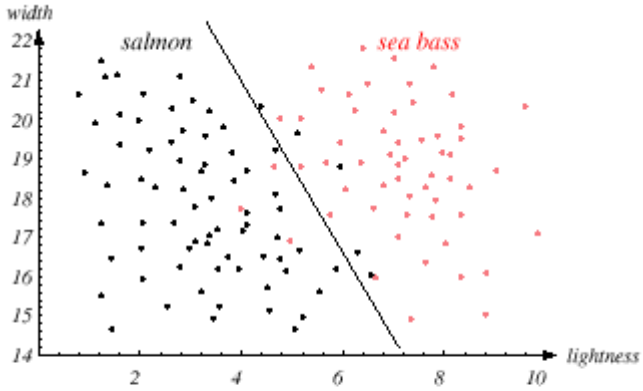


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Luokittelijan suorituskyky saattaisi olla vielä parempi, jos piirteitä lisättäisiin. Käytännössä täytyy kokeilla useita vaihtoehtoisia piirrevektoreita riittävän suorituskyvyn saavuttamiseksi. Kahden piirteen korreloidessa voimakkaasti keskenään, toinen voidaan usein jättää pois tarkastelusta. Toisia piirteitä voi myös olla kallista mitata tai raskasta laskea, jolloin on harkittava niiden käyttökelpoisuutta.

Kun piirrevektoria ei haluta pidentää, voidaan vielä kokeilla monimutkaisempia päätöspintoja lineaarisen sijasta. Allaoleva kuva näyttää voimakkaasti mutkittelevan päätöspinnan, joka onnistuu erottamaan luokat toisistaan täydellisesti. Onko tämä hyvä valinta? Todennäköisesti ei ole. On huomattava, että luokittelija opetetaan (train) pienehköllä opetusaineistolla, mikä johtaa harvasti miehitettyyn piirreavaruuteen. Tällöin rajan paikan tarkka määrääminen on epävarmaa. Lisäksi luokittelijan on kyettävä luokittelemaan mahdollisimman hyvin myös uudet tapaukset, jotka satunnaisvaihtelusta johtuen saattavat jakautua piirreavaruuteen hieman toisella tavalla. Luokittelijan täytyy pystyä yleistämään (generalize) opetusaineisto, eli tekemään mahdollisimman hyvä arvaus todellisista jakaumista.

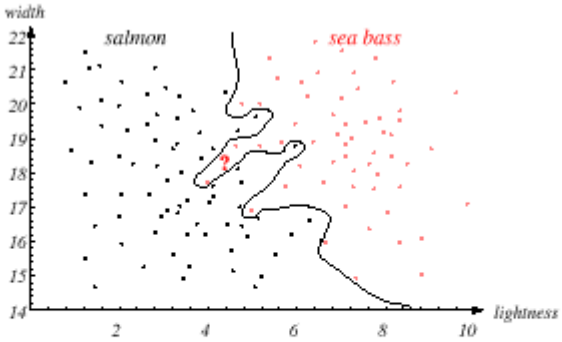


FIGURE 1.5. Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked **?** is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea bass. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Alla olevassa kuvassa on kokeiltu vähemmän monimutkaista epälineaarista päätöspintaa, joka näyttäisi noudattavan paremmin jakaumien välistä rajaa ja siten antavan hieman paremman perustan luokittelijan toteutukselle kuin lineaarinen päätöspinta.

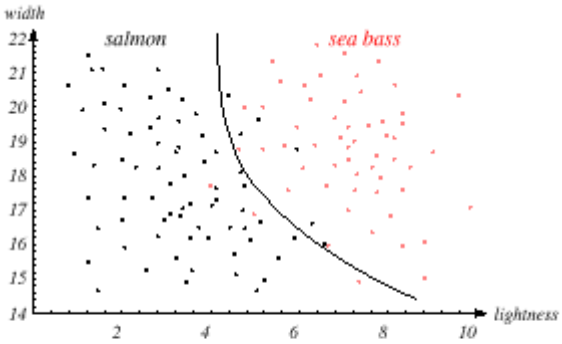


FIGURE 1.6. The decision boundary shown might represent the optimal tradeoff between performance on the training set and simplicity of classifier, thereby giving the highest accuracy on new patterns. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Piirteiden valinta, jakaumien arviointi ja hyvän päätössäännön löytäminen opetusaineiston avulla on tilastollisessa hahmontunnistuksessa yksi tärkeimpiä tehtäviä ja tutkimusongelmia. Yleiskäyttöistä hahmontunnistusjärjestelmää ei ole olemassa ja sitä ei ehkä pystytä edes suunnittelemaan. Käytännössä luokittelija on suunniteltava sovelluskohtaisesti, mikä edellyttää runsaasti kokeiluja, tietoa vaihtoehtoisista ratkaisuista, idearikkautta ja kokemusta.

Käytännössä piirteiden lukumäärä pyritään pitämään mahdollisimman pienenä, jolloin päätöspintojen väliin jäävät päätösalueet (decision regions) ovat todennäköisesti yksinkertaisia ja luokittelija voidaan opettaa pienehköllä aineistolla riittävän tarkaksi. On myös hyvä, jos piirteet ovat epäherkkiä (robust) mittauskohinalle ja muille virheille.

Luokittelijan suunnittelussa on pyrittävä käyttämään mahdollisimman paljon tietoa sovellusongelmasta. Tämä on erityisen tärkeää opetusaineiston ollessa pieni. Erityisen käyttökelpoista on tietous hahmojen muodostumismekanismista:

- automaattinen puheentunnistus: malli puheäänen ominaisuudet muodostavasta ääntöväylästä helpottaa piirteiden keksimistä (harmoonisia sisältävä spektri)
- tuolin tunnistaminen konenäöllä: tuolilla on tietty funktionaalinen aspekti, joka helpottaa luonnehtimista (stabiili keinotekoinen luomus jonka päällä ihminen voi istua, ja jossa on selkätuki).

1.2. Hahmontunnistusjärjestelmän rakenne

Allaolevassa kuvassa on esitetty tarkennettu kaavio hahmontunnistusjärjestelmän rakenteesta:

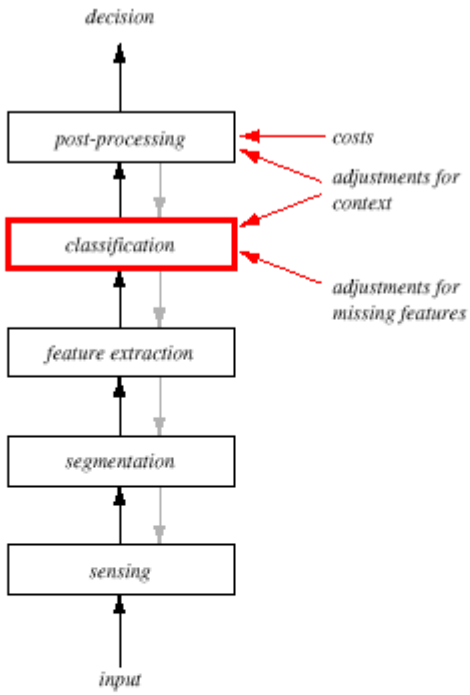


FIGURE 1.7. Many pattern recognition systems can be partitioned into components such as the ones shown here. A sensor converts images or sounds or other physical inputs into signal data. The segmentor isolates sensed objects from the background or from other objects. A feature extractor measures object properties that are useful for classification. The classifier uses these features to assign the sensed object to a category. Finally, a post processor can take account of other considerations, such as the effects of context and the costs of errors, to decide on the appropriate action. Although this description stresses a one-way or "bottom-up" flow of data, some systems employ feedback from higher levels back down to lower levels (gray arrows). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Sensorit

- kriteereinä kaistanleveys, resoluutio, herkkyys, vääristymät, SNR

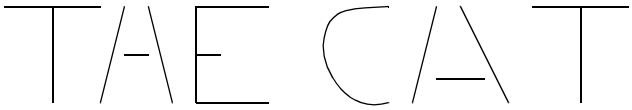
Segmentointi

- kuinka kohde erotetaan taustasta ja muista kohteista
- ongelmana mm. päällekkäisyydet, kohteen näkyminen vain osittain kuvassa, kohteen koostuminen useasta osasta, konenäössä valaisuolosuhteet
- hahmontunnistuksen syvällisimpiä tutkimusongelmia

Piirteen laskenta
- erottelukykyisen luonnehdinnan laskenta kohteesta (distinguishing features)
- samaan luokkaan kuuluvat kohteet tulisi piirteiden valossa näyttää samankaltaisilta, ja eri luokkiin kuuluvat erilaisilta
- piirteiden tulisi olla epäherkkiä (invariant) sellaisille tekijöille, jotka eivät lisää erottelukykyä: kohteiden paikka (translation) ja kiertymä (rotation) kuvassa, jossakin tapauksissa myös kohteiden koko (scale)
- kohteiden mahdollinen päällekkäisyys tai osittainen näkymättömyys signaalissa tuottaa vaikeuksia piirteen laskennalle
- hyvillä piirteillä helpotetaan luokittelijan tehtävää
- esimerkkejä vaikeuksista:
    - puhenopeus vaikuttaa piirrearvoihin automaattisessa puheentunnistuksessa, samoin ympäristön hälyäänet
    - 3D-kappaleiden tunnistusta konenäöllä vaikeuttaa kappaleen kiertyminen
    - käsinkirjoitettujen merkkien optisessa tunnistuksessa (OCR) merkkien muodon vaihtelevuus

Luokittelu
- hahmon luokittelu etukäteen opetettuun kategoriaan/luokkaan piirrevektorin avulla
- täydellistä onnistumista ei yleensä saavuteta luokkajakaumien osittaisen päällekkäisyyden takia, joten tilastollinen luokittelija pyrkii sijoittamaan tunnistettavan hahmon todennäköisimpään luokkaan
    - mittauskohina ja luokkaan kuuluvien hahmojen luontainen vaihtelu synnyttävät jakauman
- kuinka luokittelija toimii, jos jotain hahmon piirrettä ei pystytäkään laskemaan, esimerkiksi kalojen päällekkäisyyden takia?

Jälkikäsittely
- liityntä järjestelmään, joka päättää toimenpiteistä, kuten kalan sijoittaminen pakkauslaatikkoon
- voidaan myös yhdistää useiden luokittelijoiden ehdotuksia
    - huuliltaluvun yhdistäminen puheen tunnistukseen
- tilannetiedon (context) hyödyntäminen:



1.3. Hahmontunnistusjärjestelmän suunnittelu

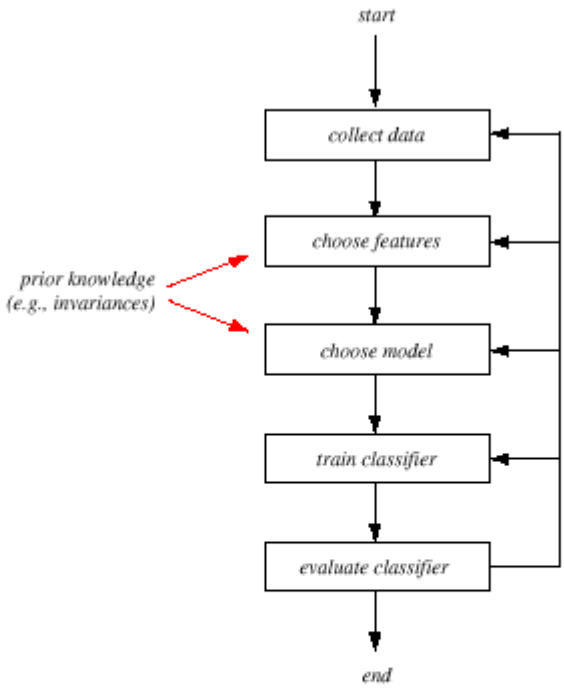


FIGURE 1.8. The design of a pattern recognition system involves a design cycle similar to the one shown here. Data must be collected, both to train and to test the system. The characteristics of the data impact both the choice of appropriate discriminating features and the choice of models for the different categories. The training process uses some or all of the data to determine the system parameters. The results of evaluation may call for repetition of various steps in this process in order to obtain satisfactory results. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Datan keruu
- “tyypillisten” tapausten keräämistä jokaisesta luokasta: aineiston tulee edustaa hyvin niitä tapauksia, joita luokitellaan varsinaisessa tuotantokäytössä
- kuinka paljon aineistoa tulee kerätä luotettavan luokittelijan toteuttamiseksi?

Piirteiden valinta
- vaihtoehtoja lukemattomia, tärkeänä kriteerinä luokkien erottelukyky
- etukäteistietoa kannattaa hyödyntää tehokkaasti
- kehitetty algoritmeja parhaiden piirteiden löytämiseen joukosta ehdokkaita; jokainen piirrevektori tulee testata erikseen
- kannattaa pitäytyä mahdollisimman pienessä lukumäärässä piirteitä
- yleensä ei kannata sisällyttää voimakkaasti keskenään korreloivia piirteitä

Mallin valinta
- kuinka piirteiden sisältämä informaatio käsitellään luokittelijassa, eli millainen rakenne luokittelijalla on? Kuvastaa luokittelijan kompleksisuusastetta
  - esimerkiksi parametriset ja ei-parametriset luokittelijat

Luokittelijan opetus
- koottu aineisto jaetaan opetus- ja testiaineistoon
- tilastollisen luokittelijan tapauksessa kyse on usein parametrien kiinnittämisestä
- varottava ylioppimista (overtraining, overfitting), jolloin yleistyskyky heikkenee

Luokittelijan testaus
- arviointimenetelmiä:
  - virheellisten luokitusten suhteellinen osuus
  - virheellisten luokitusten kokonaiskustannus
  - luokkien sekaannusmatriisit (riveillä oikeat luokat ja sarakkeilla luokittelijan tunnistustulos, soluissa prosentit)

	KISSA	KOIRA	LINTU
KISSA	95	4	1
KOIRA	1	98	1
LINTU	2	0	98

Aineiston jakaminen opetus- ja testiaineistoon voi tapahtua eri tavoin, usein käytetään seuraavia:
- hold-out: jaetaan kahteen osaan, joista toisella opetetaan ja toisella testataan. opetusaineistoa olisi hyvä olla yli puolet
- leave-k-out: poimitaan aineistosta k testinäytettä pois ja opetetaan lopulla sekä kirjataan kuinka moni testinäytteistä luokiteltiin oikein ja väärin; palautetaan k näytettä takaisin, poimitaan toiset k näytettä testiaineistoksi, opetetaan lopuilla sekä testataan ja merkitään tulos muistiin; näin jatketaan kunnes kaikki näytteet ovat kerran olleet testinäytteinä; lopuksi lasketaan keskimääräinen tulos

Oppimisen muunnelmia:

- ohjattu oppiminen (supervised learning)
  - jokaiseen datanäytteeseen liitetään ennen opetusta luokkatiedon kertova leima (label)
  - luokittelija opettelee leimatuilla aineistoilla päätösalueet tai päätöspinnat
  - tyypillinen luokittelijan opetusmenetelmä
- ohjaamaton oppiminen (unsupervised learning)
  - näytteiden leimoja ei kerrota opetusvaiheessa
  - soveltuu erityisesti tilastolliseen ryhmittelyanalyysiin (clustering), jossa pyritään löytämään datan ryhmittymisrakenne piirreavaruudessa
- vahvistava ohjaaminen (reinforcement learning)
  - luokittelijalle kerrotaan vain luokittelun onnistumisesta tai epäonnistumisesta eli vahvistetaan onnistuiko vai ei, mutta ei mihin monista luokista hahmo todellisuudessa kuului