Kapitel 5

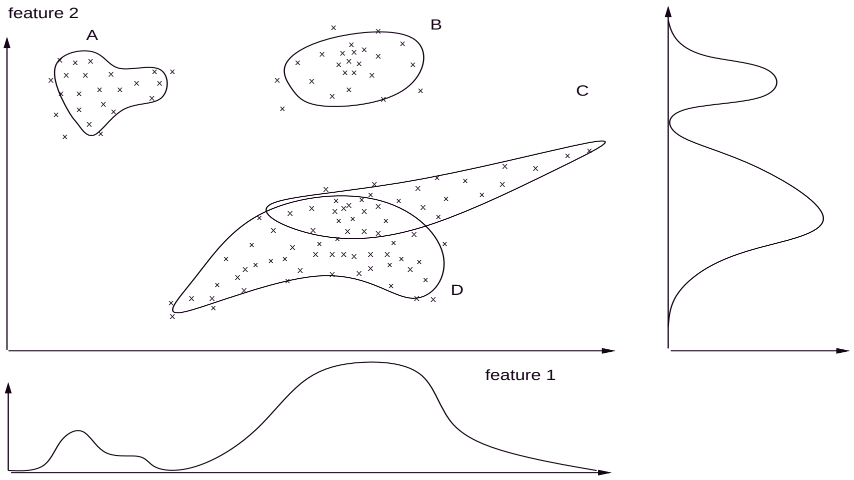
Genkendelse og klassifikation

Dette kapitel omhandler hvorledes observationer kan klassificeres (eller genkendes). Mange opgaver i billedbehandling afsluttes ofte med en genkendelsesfase, hvor identificerede "objekter" i et billede klassificeres efter en række kendte "objekt"-prototyper, eller efter et sæt af regler. "Objekterne" er sædvanligvis fremkommet som resultat af en eller flere faser involverende filtrering, segmentering e.lign. De indentificerede "objekter" er typisk beskrevet ved en egenskabsliste. Denne kan eksempelvis indeholde poster angivende form, størrelse, farve etc. Klassifikationens opgave er at knytte navne som f.eks. "hus", "vej" eller "ukendt" til "objekterne".

Klassifikationen foretages sædvanligvis ud fra en større mængde af formaliseret viden om de "objekter" som man er interesseret i at genkende. Den aktuelle viden om "objekternes" egenskaber kan eksempelvis være fremkommet ved manuel klassifikation, eller ved konsultation af en ekspert. De benyttede formalismer kan variere meget. I nogle systemer formuleres den aktuelle viden som et regelsæt af formen "hvis x så y". I de senere år er der fremkommet en mængde systemer, der ud fra ufuldkommen information er i stand til at foretage klassifikation. Disse systemer markedsføres ofte under modeordert *ekspertsystemer*. Vi skal ikke i disse noter gå dybere ind i de mange metoder, der benyttes i sådanne ekspertsystemer. Dette er omfangsmæssigt et kursus for sig. I andre systemer beregnes et sæt af *beslutningsfunktioner* på basis af et antal af stikprøver. Det er ofte nødvendigt at systemerne kan håndtere talstørrelser, som er behæfter med usikkerhed. Derfor er statistiske metoder til genkendelse af mønstre ofte nødvendige.

I afsnit 5.1 beskrives en metode til gruppering af ekstraherede features for et sæt af objekter i klynger (eller klasser). I afsnit 5.2 introduceres til metoder i "feature matching". I afsnit 5.3 beskrives klassifikation ved beslutningsfunktioner. I afsnit 5.4 introduceres nogle grundlæggende metoder til statistisk klassifikation. Endelig er klassifikation ved neurale net kort beskrevet i afsnit 5.4.

Formålet med en *klassifikation* er at afgøre hvilken *klasse* en given observation stammer fra. Det antages at alle observationer fra samme klasse deler en række fællestræk, der således karakteriserer klassen. Opgaven, at bestemme hensigtsmæssige mål for de fælles karakteristika betegnes *feature ekstraktion*. Problemet i en klassifikation er, at der ofte kan forekomme stor variation af hver feature værdi inden for hver klasse. Dette er illustreret i nedenstående figur, hvor fire klasser af observationer, hver karakteriseret ved to features er illustreret.



Figur 5.1: Observationer hørende til 4 klasser plottet som funktion af 2 features. Klasse A er velsepereret fra de øvrige. Klasse B og C overlapper som fuktion af feature 1, men ikke som funktion af feature 2. Observationerne hørende til klasse C og D er ikke vel sepererede hverken som funktion af feature 1, eller som funktion af feature 2.

I visse heldige situationer kan features vælges således at de sepererer klasserne fuldstændigt. Denne situation er imidlertid sjælden. Generelt vil der der findes observationer, der faktisk stammer fra klasse i spredt overalt i featurerummet, og derfor ikke umiddelbart kan skelnes fra observationer stammende fra klasse j. Klassifikationsfejlen er derfor sjældent lig 0. De væsenligste problemer i statistisk klassifikation er således:

- Bestemmelse af antallet af klasser, hver klasses position i featurerummet, samt variationen af (og korrelationen imellem) de enkelte featureværdier for hver klasse. Denne proces foretages ofte på et sæt af *træningsdata*, hvis korrekte klassifikation er kendt.

- Bestemmelse af en klassifikationsmetode, dvs. en procedure, hvormed en ny observation navngives efter en af de i træningsfasen bestemte klasser.

I visse situationer er antallet klasser kendt på forhånd. Dette er eksempelvis tilfældet i mange industrielle opgaver. Et eksempel er genkendelse af tegn i et OCR-system (eng. Optical Character Recognition). Antag at en tekst er scannet, at de enkelte linier i teksten er sepereret og at de enkelte tegn i hver linie er fundet. Ofte er disse repræsenteret ved et binært rasterbillede. De enkelte billedelementer (pixels) har værdi 0 hvis det tilsvarende område på papiret var sværtet, og værdi 1, hvis området var usværtet (hvidt). Hvis de tilladte tegn er dem, der findes på mit tastatur, vil der være ialt 99 klasser (29 små bogstaver, 29 store bogstaver, 10 tal, og 31 specialtegn). Da trykkeprocessen sjældent er ideel, papiret ikke er fuldstændig homogent, og scanneprocessen ikke er fejlfri, vil billederne af tegnene ikke svare fuldstændigt til prototyperne på disse. Opgaven er derfor at bestemme en procedure for bestemmelse af det tegn, der forelagt et billede, ligner bedst.

I andre situationer er hverken antallet af klasser eller fordelingen af observationerne inden for hver klasse kendt. Ved kortlægning af et dyrkningsmøster på basis af multispektrale flyfotos er opgaven at bestemme typerne af afgrøder, hvor store arealer der benyttes til hver sort mv. Da en given afgrødes multispektrale signatur varierer over tid siden såning, med markens egenskaber (herunder jordbundsforhold og hældning i forhold til solen) er det vanskeligt at definere prototyper. Ofte må antallet og karakteriseringen af klasserne bestemmes løbende ud fra de til rådighed værende observationer.

Man skelner mellem *overvåget* klassifikation og *uovervåget* klassifikation. I overvåget klassifikation antages det at et træningssæt af observationer med korrekt klassifikation er givet. Hver klasse er således implicit karakteriseret ved et sæt af observationer. Disse kan benyttes til estimation af fordelingen af observationer som funktion af featurevektoren. I *uovervåget klassifikation* forudsættes intet kenskab til den korrekte klassifikation af observationerne. Antallet af klasser er ofte ukendt. Typisk er formålet med en uovervåget klassifikation netop at bestemme antallet af klasser, samt deres beliggenhed mv. i featurerummet. Denne proces kaldes også for *klyngeanalyse*.

Klyngeanalyse benyttes ofte i en indledende fase af en dataanalyse. Typisk eksperimenteres med forskellige features. For hvert valg af features benyttes klyngeanalysen sammen med en statistisk karakteristikation af de fundne klasser til vurdering af valget af features. Når valget af features er foretaget, og klasserne karakteriseret kan en overvåget (egentlig) klassifikation benyttes til klassifikation af nye observationer.

5.1 Klyngeanalyse

I uovervåget klassifikation ønskes de klynger af observationer, som data grupperer sig i, fundet. Hvis antallet af features (dimensionen af observationsvektoren) er lille (1-3), og antallet af klasser er lille, kan analysen ofte foretages interaktivt ved at plotte observationerne, og manualt foretage identifikationen af klynger. Hvis antallet af features er stort er denne metode sjældent en farbar vej frem. I stedet benyttes en algoritme til navngivning af observationerne efter den klynge hvori de grupperes. Der findes en lang række metoder til klyngeanalyse. Nedenstående gives kun et kort resumé af *k-means* algoritmen.

I *k-means* algoritmen antages at antallet k af klasser er kendt. Initielt vælges k (ofte meget) approksimativt bestemte klyngecentre. Derefter navngives alle observationer efter den klynge, der ligger tættest på observationen. Nye klyngecentre udregnes ved middelværdivektoren af observationerne inden for hver klynge. Derefter itereres navngivningen og genberegningen af klyngecentre sålænge blot ét klyngecenter flytter sig.

1. Lad iterationsnummeret $t = 1$. Vælg k klyngecentre $\mathbf{C}_i^t$, f.eks. ved featurevektorene for de første k af de ialt n observationer.
2. Navngiv hver observation $\mathbf{x}$ med klyngenavnet j hvis der for alle $i \neq j, 1 \leq i \leq k$ gælder at:
$$||\mathbf{x} - \mathbf{C}_j^t|| \ < \ ||\mathbf{x} - \mathbf{C}_i^t||$$
3. Beregn nye klyngecentre ved:
$$\mathbf{C}_i^{t+1} \ = \ \frac{1}{n_i^t} \sum_{\mathbf{x} \in K_i^t} \mathbf{x}$$
hvor K_i^t er klynge nummer i, og hvor n_i^t er antallet af observationer med klyngenavnet i.
4. Lad c være et konvergensmål, f.eks. bestemt ved:
$$c = \sum_{l=1}^{k} ||\mathbf{C}_i^{t+1} \ - \ \mathbf{C}_i^t||$$
Lad $t := t + 1$. Hvis $c = 0$ så stop, ellers fortsæt med punkt 2.

Ovenstående simple metode kan modificeres på en lang række punkter således at nye klyngecentre kan introduceres, således at klynger med meget få "medlemmer" kan nedlægges etc. En ofte refereret metode af denne type er *ISODATA* algoritmen [2, 10].

Væsentlig for *k-means* algoritmen er kravet om kendskab til antallet af klyngecentre, valget af initielle klyngecentre samt det afstandsmål hvormed observationer sammenlignes med klyngecentre. Det er ofte fordelagtigt at benytte en anden algoritme til indledende bestemmelse af antal klynger samt deres omtrentlige position. Ved hovedløs anvendelse af det euclidiske afstandsmål vil resultatet sjældent svare til det forventede. Dette skyldes at visse featureværdier varierer mere end andre (globalt set), og at visse featureværdier varierer forskelligt afhængig af den korrekte klassifikation. Vi skal ikke her diskutere de forskellige afstandsmål. Endelig skal det bemærkes at der intet bevis findes for konvergens af *k-means* algoritmen. I praksis fungerer metoden imidlertid ofte tilfredsstillende.

Som tidligere omtalt er resultatet af en klyngeanalyse afhængig af valget af features, af det valgte afstandsmål mv. Dette gør klyngeanalyse til lidt af en kunst. Det er sjældent muligt at vælge optimale features og afstandsmål i første fortsøg. Klyngeanalyse må derfor betragtes som et redskab i en interaktiv proces, med det formål at forstå strukturen i observationerne.

Givet en manuel klassifikation, eller en klassifikation på basis af en klyngeanalyse, er det som regel hensigtsmæssigt at karakterisere de fundne klasser ved en række statistiske mål. Positionen af hver klasse bestemmes ved feature middelværdivektoren for observationerne hørende til hver klasse. Yderligere kan hver klasse karakteriseres ved kovariansmatricen for observationerne. Hver klasse karakteriseres således ved $p + p * (p + 1)/2$ reelle tal. Disse størrelser er ikke kendt på forhånd, men må estimeres ud fra de foreliggende observationer.

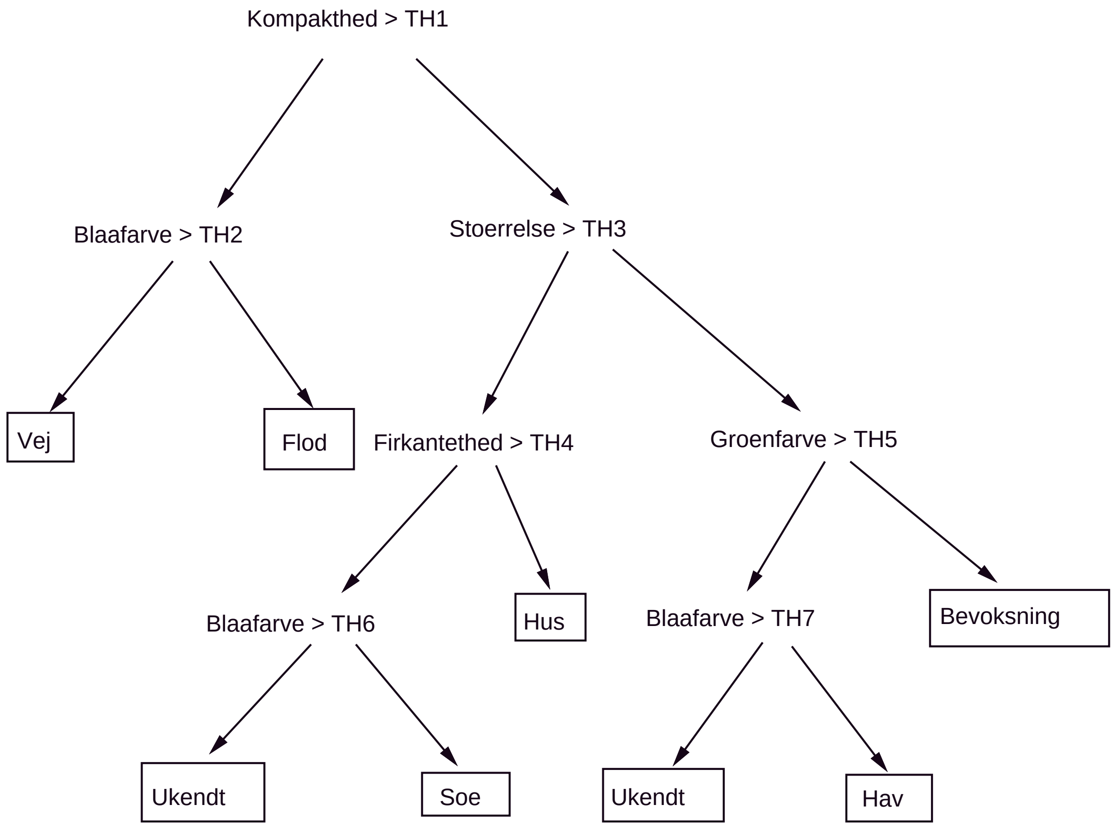
5.2 Feature matching

Der findes et meget stort antal metoder til navngivning af observationer til et sæt af fortolkninger. Vi skal senere diskutere nogle af de statistisk funderede metoder. Her skal nogle få andre metoder kort nævnes. Disse bygger ofte på anvendelse af heuristikker eller særlig a priori viden.

Antag, som eksempel, at vi ønsker at klassificere et billede af et tegn til et alfabet, at tegnet er givet ved et digitalt billede, og at samtlige tegn i alfabetet er givet ved en *prototype* af billedet. En oplagt metode er da, for hvert tegn i alfabetet, at krydskorrelere prototypebilledet for dette med billedet af det observerede tegn. Herved tilknyttes hver mulig fortolkning et godhedsmål (korrelationskoefficienten) for matchet. Observationen klassificeres til det tegn i alfabetet, der opnår størst godhedsmål. Denne metode kaldes *template matching*. Metoden er generelt anvendelig, men ofte langsom, fordi antallet af tegn i alfabetet kan være stort. Et andet problem kan være at godhedsmålet for et match kan være omtrentligt lige stort for et antal prototyper. Valget af vinder kan derfor bero på mængden af støj (forvrængning mv.) af observationen.

For at mindske de beregningsmæssige omkostninger ved klassifikationen er en mulighed at knytte et antal features til observationen. Disse kan nok være beregningskrævende, men hvis antallet af features er væsenligt mindre end antallet af prototyper er der en ofte en samlet gevinst. Opgaven er nu at sammenligne featurevektoren for observationen med de tilsvarede featurevektorer for hver prototype. Dette kræver fastlæggelse af en norm. Ved direkte brug af den euclidiske norm vil man ofte få uventede resultater, fordi alle features vægtes ens ved denne norm. I praksis varierer visse features mere end andre. Derfor er en normalisering ofte påkrævet. Hertil kommer at visse features ofte betyder mere end andre. Erfaring, evt. opnået gennem maskinel indlæring, kan være afgørende for et velfungerende system. Alternativt kan tænkes at de enkelte feature ekstraktions modeller resulterede i en binær værdi (accept eller reject). Det kan argumenteres at dette blot flytter problemet fra klassifikationsmodulet til ekstraktionsmodulerne. Klassifikationen blever her en simpel optælling af stemmer. Den fortolkning, som flest feature ekstraktionsmoduler stemmer på, vinder. Denne metode er ofte mere robust (pga. kvantificeringen til binære værdier). Ofte vil flere fortolkninger imidlertid få samme stemmetal. Yderligere semantisk information er derfor påkrævet for at kunne afgøre tvivlstilfældene.

Hvert feature ekstraktionsmodul med binær værdi kan betragtes som en test. I situationer, hvor tid er afgørende, er det ofte en god ide, at rangordne de enkelte test i et træ. Således udføres en test kun hvis de foregående test ikke har kunnet afgøre klassetilhørsforholdet. Et sådant træ kaldes et klassifikationstræ . Opbygningen af et klassifikationstræ er vanskelig. Dels er det ikke klart hvilke test der bør foretages før andre test. Dels er beregningstunge tests (der gerne skulle anvendes så sjældent som muligt) også ofte de mest sigende om klassetilhørsforholdet. Metoden er sædeles anvendelig, men kræver godt kendskab til de benyttede feature ekstraktionsmetoder.



Figur 5.2: Eksempel på et simpelt klassifikationstræ for et segment med feature attributterne: Størrelse, kompakthed, firkantethed, og farve.

5.3 Beslutningsfunktioner

Det grundlæggende problem i designet af en klassifikationsmetode er at opdele featurerummet i delrum, der hver svarer til en klasse. Forelagt en ny observation består klassifikationen i at bestemme, hvilket delrum featurevektoren tilhører. Opdelingen af featurerummet kan foretages på mange måder. Den kan tage udgangspunkt i en statisktisk karakterisering (middelværdi og kovariansmatrice) af hver klasse, på basis af en klyngeanalyse, eller direkte ud fra et sæt af observationer hvis korrekte klassifikation er kendt. Ofte ønskes at opdelingen er disjunkt og at hele featurerummet dækkes. Opdelingen specificeres som regel ved grænserne mellem nabo-klasser. Ofte antages, at disse grænser har en matematisk simpel form, f.eks. lineære, kvadratiske, eller stykvis lineære.

I det følgende vil vi betegne den i'te klasse med ω_i. En *beslutningsfunktion* $g_{ij}(\mathbf{x})$ er en reel funktion, defineret på featurerummet, med den egenskab at mønsterklasserne i og j befinder sig på hver side af hyperfladen $g_{ij}(\mathbf{x}) = 0$, dvs. således at:

$$\begin{aligned} g_{ij}(\mathbf{x}) > 0 \quad &\text{hvis} \quad \mathbf{x} \in \omega_i \\ g_{ij}(\mathbf{x}) < 0 \quad &\text{hvis} \quad \mathbf{x} \in \omega_j \end{aligned}$$

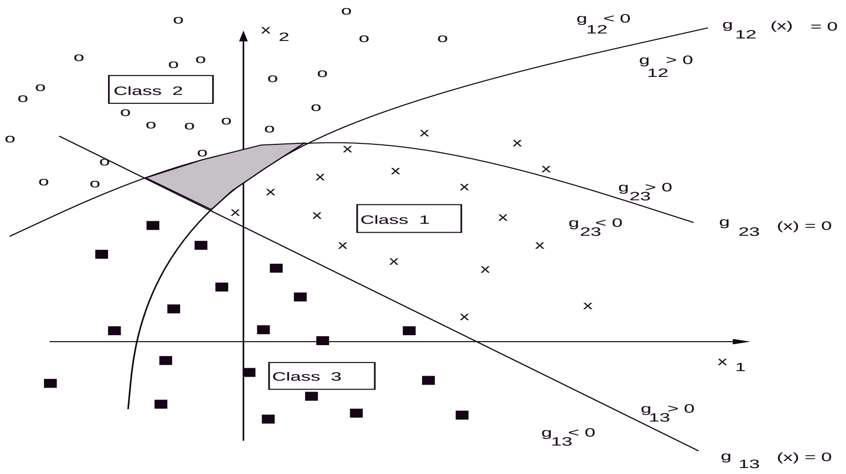
Der gælder at $g_{ij}(\mathbf{x}) = -g_{ji}(\mathbf{x})$. Hvis der ialt er L klasser vil der derfor være $L(L-1)/2$ beslutningsfunktioner. Lad $c(\mathbf{x})$ være en *klassifikationsfunktion*, bestemt ved:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad \mathbf{x} \in \omega_i$$

Denne funktion kan jvf. det ovenstående specificeres ved:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad g_{ij}(\mathbf{x}) > 0 \quad \forall j \neq i$$

På grænsen mellem to områder i featurerummet vil det ikke være muligt at bestemme en entydig klassifikation. En mulighed er, i dette tilfælde, at vælge klassetilhørsforholdet tilfældigt. Ligeledes vil der ofte være et område hvor det ikke vil være muligt at afgøre klassetilhørsforholdet. Nedenstående figur illustrerer brugen af beslutningsfunktioner for et 3-klasse problem.



Figur 5.3: Beslutningsfunktioner for et 3-klasse problem med 2 features. I de grå område er det ikke muligt at foretage nogen entydig klassifikation.

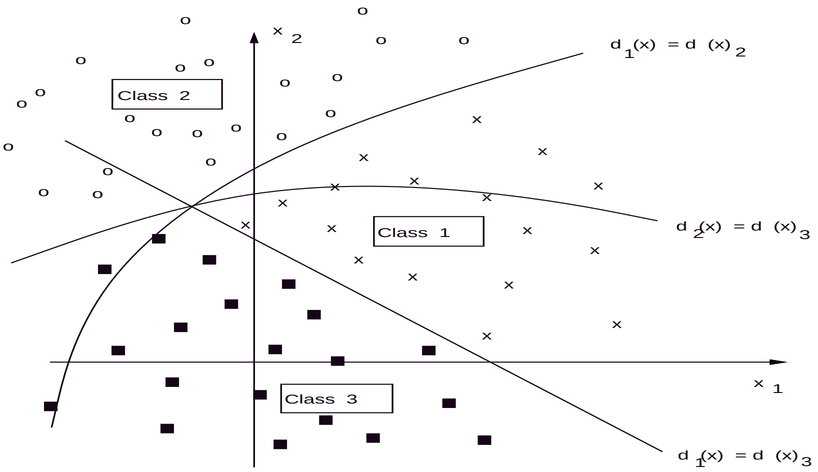
For at undgå problemerne med områder, der ikke har nogen entydig klassifikation, samt for at reducere antallet af forskellige beslutningsfunktioner antages ofte at disse kan skrives:

$$g_{ij}(\mathbf{x}) = d_i(\mathbf{x}) \; - \; d_j(\mathbf{x})$$

hvor d_i kaldes en *diskriminantfunktion*. Klassifikationsfunktionen bliver nu:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad d_i(\mathbf{x}) \; > \; d_j(\mathbf{x}) \quad \forall j \neq i \tag{5.1}$$

Der findes netop L diskriminantfunktioner. Grænsefunktionerne, bestemt ved $d_i(\mathbf{x}) = d_j(\mathbf{x})$ vil alle skære hinanden i ét fælles punkt. Dette er illustreret i nedenstående figur. Analyser baseret på konstruktion og brug af diskriminantfunktioner kaldes *diskriminantanalyse*.



Figur 5.4: Diskriminantfunktioner for et 3-klasse problem med 2 features.

De mest benyttede diskriminantfunktioner er lineære. Ved lineær diskriminantanalyse i et n-dimensionalt featurerum kan diskriminantfunktionen skrives:

$$\begin{aligned} d_i(\mathbf{x}) &= w_1^i x_1 + w_2^i + \cdots w_n^i x_n + w_{n+1} \\ &= \mathbf{w}_i^t \mathbf{x} + w_{n+1} \end{aligned}$$

Hyperplanerne H_{ij}, der udgør grænsefunktionerne, er bestemt ved $d_i(\mathbf{x}) = d_j(\mathbf{x})$. Lineære diskriminantfunktioner kan være gode hvis klasserne er vel separerede. Hvis der findes et sæt af lineære diskriminantfunktioner, der separerer klasserne fuldstændigt kaldes disse for *lineært separable klasser*. Vi skal senere, i afsnittet om neurale net, diskutere lineære diskriminantfunktioner i større detalje.

5.3.1 Minimal afstands klassifikatorer

Antag at vi kun kender klassernes position i featurerummet. Denne position kan betragres som en *prototype* på observationerne i klassen. Hvis hver prototype er en god approksimation til enhver observation fra klassen vil en klassifikationsmetode baseret på minimal afstand ofte være effektiv. Klassifikationen er bestemt ved beslutningsreglen (5.1) samt definitionen:

$$d_i(\mathbf{x}) = -\,dist(\mathbf{x}, \mathbf{x}_i) \tag{5.2}$$

hvor $\mathbf{x_i}$ er prototypen for klasse i, og $dist()$ er en afstandsfunktion. En ofte benyttet afstand er den euclidiske. I dette tilfælde bliver klassifikationsfunktionen:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad \sum_{k=1}^{n}(x_k - x_k^i)^2 < \sum_{k=1}^{n}(x_k - x_k^j)^2 \quad \forall j \neq i$$

hvor n er dimensionen af $\mathbf{x}$. Brugen af det euclidiske afstandsmål forudsætter som regel at hver af featureværdierne er skaleret således at de (totalt set) varierer lige meget. I modsat fald vil visse features ofte få dominerende indflydelse på afstandsmålet. Ved udregning ses at ovenstående klassifikationsfunktion svarer til den lineære diskriminantfunktion:

$$d_i(\mathbf{x}) \;=\; \sum_{k=1}^{n} x_k x_k^i \;-\; \frac{1}{2}\sum_{k=1}^{n}(x_k^i)^2 \tag{5.3}$$

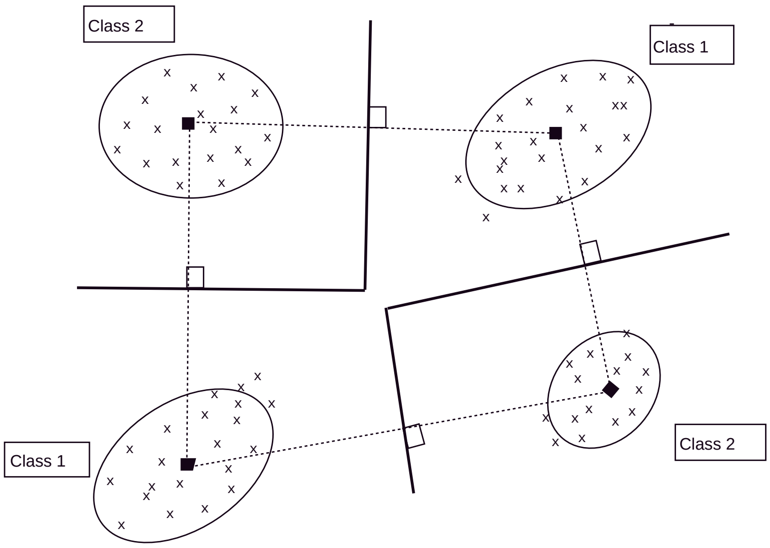
Bemærk at sidste led kan udregnes på forhånd for alle prototyper, og ikke behøver at blive evalueret ved klassifikation af hver ny observation $\mathbf{x}$. Hvis klasserne ikke er lineært separable, men visse klasser kan opdeles i underklasser således at problemet herved bliver lineært separabelt kan diskriminantfunktionen formuleres ved:

$$d_i(\mathbf{x}) \;=\; -\min_{j\in[1..N_i]} dist(\mathbf{x}, \mathbf{x}_{ij}) \tag{5.4}$$

hvor N_i er antallet af underklasser af klasse i. Benyttes det euclidiske afstandsmål bliver diskriminantfunktionen:

$$d_i(\mathbf{x}) \;=\; \max_{j\in[1..N_i]} \left\{(\mathbf{x}^t\mathbf{x}_{ij} \;-\; \frac{1}{2}(\mathbf{x}_{ij})^t)\mathbf{x}_{ij}\right\} \tag{5.5}$$

Som illusteret nedenstående vil denne formulering give anledning til et sæt af stykvis lineære diskriminantfunktioner.



Figur 5.5: Stykvis lineære diskriminantfunktioner for et 2-klasse problem hvor hver klasse er beskrevet ved 2 prototyper.

5.4 Diskriminantanalyse

I mange situationer varierer de ekstraherede features meget i forhold til afstandene mellem prototyperne. I disse tilfælde er det ofte umuligt at bestemme grænsefunktioner der separere klasserne fuldstændigt. I stedet er det nødvendigt at bestemme klassifikationsregler, der er statistisk bestemt. I dette afsnit beskrives en statistisk optimal klassifikation. Denne har meget bred anvendelse, og betragtes som en reference, i forhold til hvilken nye metoder kan sammenlignes.

Beslutningsprocessen i et system til mønstergenkendelse kan betragtes som et statistisk spil , hvor systemet spiller mod "naturen". Denne antages at kende klasserne hvorfra data kan stamme. Hver runde af spillet foregår ved at naturen vælger klasse ω_i, samt en observation $\mathbf{x}$ fra denne klasse. Systemet skal, præsenteret for $\mathbf{x}$ "gætte" klassen som observationen stammer fra. Hvert korrekt gæt belønnes med en vis størrelse, og hvert forket gæt straffes med en anden størrelse. Spillet antages at være en såkaldt "0-sum spil", hvor den ene spillers tab er lig den anden spillers gevinst. Opgaven for systemet vil typisk være at finde en (optimal) beslutningsproces, der minimerer det gennemsnitlige tab. En alternativ strategi, som vi ikke skal omtale yderligere, er at minimere det maksimale tab.

Antag at systemet klassificerer $\mathbf{x} \in \omega_j$ når $\mathbf{x}$ faktisk stammede fra ω_i. *Genvinst- eller tabs-funktionen* L_{ij} angiver da udgiften (eller gevinsten) for systemet. Som konvension vil vi betegne tab med positive værdier, og gevinster med ikke-positive værdier. Antag at antallet af klasser er M. L_{ij} kan da opskrives som en matrice med dimension $M \times M$, hvis antallet af faktiske klasser svarer til det antal, som systemet opererer med. Som nævnt vælger "naturen" først en klasse ω_i. Dette valg foretages ud fra den a priori sandsynlighed $p(\omega_i)$ for klassen, dvs. ud fra hyppigheden af forekomster af elementer fra denne klasse. Derefter vælges en observation $\mathbf{x}$ fra ω_i. Det antages at "naturen" ikke er en intelligent modspiller (der forsøger at maksimere egen gevinst), men vælger blindt på basis af de a priori sandsynligheder. Derimod tillades at systemet "udspionere naturen", ved at konstruere eksperimenter, der skal kortlægge antallet af klasser, samt variabiliteten af $\mathbf{x}$ for hver klasse. Denne proces, der tænkes at gå forud for spillet, har vi diskuteret i afsnittet om klyngeanalyse.

Antag at sandsynligheden for at $\mathbf{x}$ stammer fra klasse ω_i er bestemt ved $p(\omega_i|\mathbf{x})$. Det forventede tab (for systemet) ved klassifikation af $\mathbf{x}$ til ω_j er da givet ved:

$$r_j(\mathbf{x}) \;=\; \sum_{i=1}^{M} L_{ij}\, p(\omega_i|\mathbf{x}) \tag{5.6}$$

Udtrykket kaldes også *det betingede gennemsnitlige tab.* Systemet kan vælge at klassificere $\mathbf{x}$ til enhver af de M klasser. Metoden er derfor at beregne det gennemsnitlige tab $r_1(\mathbf{x}), r_2(\mathbf{x}), \ldots, r_M(\mathbf{x})$ for alle M muligheder, og at vælge den klassifikation der resulterer i det mindste tab. Dette svarer til klassifikationsfunktionen:

$$c(\mathbf{x}) \;=\; j \quad \text{hvis} \quad r_j(\mathbf{x}) < r_i(\mathbf{x}) \quad \forall i \neq j, i \in [1..M] \tag{5.7}$$

Den klassifikator, der minimerer det totale forventede tab kaldes en *Bayes klassifikator* . Denne er, statistisk set, derfor optimal. Fra Bayes regel :

$$p(\omega_i|\mathbf{x}) \;=\; \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(\mathbf{x})}$$

har vi at:

$$r_j(\mathbf{x}) \;=\; \sum_{i=1}^{M} L_{ij}\; p(\mathbf{x}|\omega_i)p(\omega_i) \tag{5.8}$$

idet konstanten $p(\mathbf{x})$ er en fælles faktor for alle funktionerne $r_j(\mathbf{x})$, og derfor ingen indflydelse har på valget af klassifikationen. Funktionen $p(\mathbf{x}|\omega_i)$ kaldes for *likelihood funktionen* for klasse ω_i. Lad os betragte situationen $M = 2$. Vi har da:

$$\begin{aligned} r_1(\mathbf{x}) &= L_{11}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{21}p(\mathbf{x}|\omega_2)p(\omega_2) \\ r_2(\mathbf{x}) &= L_{12}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{22}p(\mathbf{x}|\omega_2)p(\omega_2) \end{aligned}$$

Observationen $\mathbf{x}$ tilordnes klassen ω_1 hvis $r_1(\mathbf{x}) < r_2(\mathbf{x})$, dvs. hvis:

$$L_{11}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{21}p(\mathbf{x}|\omega_2)p(\omega_2) \quad < \quad L_{12}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{22}p(\mathbf{x}|\omega_2)p(\omega_2)$$

eller, ækvivalent, hvis:

$$(L_{21} - L_{22})p(\mathbf{x}|\omega_2)p(\omega_2) \quad < \quad (L_{12} - L_{11})p(\mathbf{x}|\omega_1)p(\omega_1) \tag{5.9}$$

Vi kan generelt antage at $L_{ij} > L_{ii}$, fordi tabet ved en fejlklassifikation må være større end ved en korrekt klassifikation. Alle størrelserne i ovenstående udtryk er derfor positive. Vi kan videre skrive betingelsen:

$$l_{12} \;=\; \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \quad > \quad \frac{p(\omega_2)}{p(\omega_1)}\,\frac{L_{21} - L_{22}}{L_{12} - L_{11}} \;=\; \theta_{12} \tag{5.10}$$

Venstresiden af dette udtryk, dvs. forholdet mellem de to likelihood funktioner, betegnes *likelihood ratioen* og betegnes l_{12}. Højresiden af udtrykket kaldes *grænseværdien* (eng. threshold value) og betregnes θ_{12}. Bayes beslutningsregel for $M = 2$ bliver derfor:

1. Vælg klasse ω_1 hvis $l_{12} > \theta_{12}$.
2. Vælg klasse ω_2 hvis $l_{12} < \theta_{12}$.
3. Vælg klasse tilfældigt hvis $l_{12} = \theta_{12}$.

I det generelle tilfælde ($M > 2$) defineres tilsvarende:

$$l_{ij} \;=\; \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)} \tag{5.11}$$

$$\theta_{ij} \;=\; \frac{p(\omega_j)}{p(\omega_i)}\,\frac{L_{ji} - L_{jj}}{L_{ij} - L_{ii}} \tag{5.12}$$

Klassifikationsfunktionen bliver

$$c(\mathbf{x}) \;=\; i \quad \text{hvis} \quad l_{ij} > \theta_{ij} \quad \forall j \neq i, \;\; j \in [1..M] \tag{5.13}$$

Bayes klassifikator, som beskrevet ovenfor, er nok den mest benyttede klassifikator i mønstergenkendelse. Der findes imidlertid en række andre metoder, der er baseret på andre statistiske beslutningskriterier. Ved *minimax* kriteriet ønskes det gennemsnitlige tab under de værst tænkelige omstændigheder minimeret. Et andet kriterie kaldes *Neyman-Pearson* kriteriet. Vi skal ikke her diskutere disse kriterier nærmere (se evt. [10]), blot notere, at for begge alterative metoder er likelihood ratioen l_{ij} den samme, hvorimod grænseværdien θ_{ij} er forskellig.

Eksempel

Antag at et binært signal sendes over en støjfyldt kanal. De enkelte dataelementer er altså enten 0 eller 1. Antag videre at den støj der tilføres signalet er additiv og kan modelleres ved en normalfordeling med middelværdi 0 og spredning σ. Lad ω_1 være hypotesen at et 0 var sendt, og ω_2 være hypotesen at et 1 var sendt. Lad yderligere $p(0)$ hhv. $p(1)$ være de priori sandsynligheder for afsendelse af et 0 hhv. et 1. Vi har da at:

$$\begin{aligned} p(x|\omega_1) &= \frac{1}{\sqrt{2\pi}\sigma} \; exp(-\frac{x^2}{2\sigma^2}) \\ p(x|\omega_2) &= \frac{1}{\sqrt{2\pi}\sigma} \; exp(-\frac{(x-1)^2}{2\sigma^2}) \end{aligned}$$

Likelihood ratioen bliver derfor:

$$l_{12}(x) \; = \; exp(\frac{1-2x}{2\sigma^2})$$

Ved Bayes klassifikation vil x blive fortolket som et 0 hvis $l_{12} > \theta_{12}$, dvs. hvis:

$$x \; < \; \frac{1}{2} - \sigma^2 ln(\theta_{12}) \; = \; \frac{1}{2} - \sigma^2 ln(\frac{L_{21}p(1)}{L_{12}p(0)})$$

Hvis kanalen er støjfri, dvs. hvis $\sigma = 0$, vil x blive fortolket som et 0 hvis $x < 0.5$. Dette svarer til intuitionen. Tilsvarende gælder hvis $\theta_{12} = 1$. I andre tilfælde er den optimale grænseværdi ikke i overensstemmelse med intuitionen.

Eksempel slut

I mange situationer er bestemmelse af tabs-funktionen essentiel for den optimale løsning. Ved klassifikation af flyvemaskiner som fjendtlige eller venlige kan klassifikation af en fjendtlig flyver som venlig være katastrofal, hvorimod klassifikation af en venlig flyver som fjende kan have mindre dramatiske konsekvenser. Et andet eksempel er kvalitetskontrol i en industriel proces. Her kan det koste dyrt at klassificere et dårligt produkt som godt, medens den omvendte fejlklassifikation sjældent er vigtig. Den sidste type af fejl kan ofte korrigeres ved efterfølgende manuelt check af de frasorterede produkter. Konsekvenserne af de forskellige typer af fejl er således generelt forskellige. I resten af dete kapitel skal vi imidlertid antage at alle fejl er lige slemme. Denne antagelse kan modelleres ved:

$$L_{ij} = 1 - \delta_{ij} = \begin{cases} 1 & \text{hvis} \quad i \neq j \\ 0 & \text{hvis} \quad i = j \end{cases} \tag{5.14}$$

Ved denne definition er tabet for en korrekt klassifikation lig 0. Tabet for enhver fejlagtig klassifikation er 1. Dette betyder at *fejlraten* (det betingede gennemsnitlige tab) bliver:

$$\begin{aligned} r_j(\mathbf{x}) &= \sum_{i=1}^{M} (1 - \delta_{ij})\, p(\mathbf{x}|\omega_i) p(\omega_i) && (5.15) \\ &= p(\mathbf{x}) - p(\mathbf{x}|\omega_j) p(\omega_j) && (5.16) \end{aligned}$$

Bayes klassifikationen, der i dette tilfælde vil minimere sandsynligheden for fejl i klassifikationen, bliver derfor:

$$c(\mathbf{x}) = i \quad \text{hvis} \quad p(\mathbf{x}|\omega_i)p(\omega_i) > p(\mathbf{x}|\omega_j)p(\omega_j) \quad \forall j \neq i, \;\; 1 \leq j \leq M \tag{5.17}$$

Som det huskes fra afsnittet om beslutningsfunktioner udtrykker ovestående regel blot at beslutningsfunktionerne er bestemt ved:

$$d_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)p(\omega_i) \tag{5.18}$$

I ovenstående udtryk for beslutningsfunktionerne kræves der eksplicit kendskab til både den betingede sandsynlighed $p(\mathbf{x}|\omega_i)$ og til den a priori sandsynlighed $p(\omega_i)$. I udledningen udnyttede vi Bayes regel. Tager vi alternativt udgangspunkt i (5.6) direkte, fremkommer en helt ækvivalent formulering:

$$d_i(\mathbf{x}) = p(\omega_i|\mathbf{x}) \tag{5.19}$$

Valget mellem hvilken af de to formuleringer (5.18) eller (5.19), der er mest hensigtsmæssig, afhænger af hvilke fordelinger der, i en given anvendelse, er kendte. I mange situationer foretages en indledende dataanalyse (herunder klyngeanalyse) på et sæt af træningsdata. Herved kan de a priori sandsynligheder estimeres ved frekvensen af observationer inden for hver klasse. Yderligere er det ofte muligt at modellere de betingede sandsynligheder $p(\mathbf{x}|\omega_i)$ ved en kendt parametrisk fordelingsfunktion (feks. en normalfordeling). Vi skal diskutere dette tilfælde nærmere i nedenstående afsnit. Sandsynlighederne $p(\omega_i|\mathbf{x})$ er umiddelbart sværere at estimere, men kan være fordelagtige ved brug i metoder hvor diskriminantfunktionerne indlæres inkrementalt. Vi skal ikke her diskutere dette nærmere, men referere til f.eks. [26] for nærmere diskussion.

5.4.1 Diskriminantanalyse for normalt fordelte mønstre

I dette afsnit beskrives Bayes klassifikatoren, bestemt ved diskriminantfunktionen (5.18) for normale fordelingsfunktioner $p(\mathbf{x}|\omega_i)$. Vi skal kalde denne type af klassifikationer for *diskriminantanalyse* . Antagelsen om normalitet af fordelingen $p(\mathbf{x}|\omega_i)$ er dels begrundet i analytisk medgørlighed, dels fordi denne fordeling er en god approksimation til mange faktisk forekomne empiriske fordelinger. Fordelingen er givet ved:

$$p(\mathbf{x}|\omega_i) \;=\; \frac{1}{(2\pi)^{n/2}|\mathbf{C}_i|^{1/2}} \quad exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)] \tag{5.20}$$

hvor $\mathbf{m}_i$ er middelværdivektoren for observationerne hørende til klasse i og $\mathbf{C}_i$ er den tilsvarende kovariansmatrice. Diskriminantfunktionen er givet ved: $d_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)p(\omega_i)$. På grund af den eksponentielle form af normalfordelingen er det hensigstsmæssigt at arbejde med logaritmen af denne. Logaritmefunktionen er monotont voksende og ænder derfor ikke på sammenligningerne af diskriminantfunktionerne. Vi får således:

$$\begin{aligned} d_i(\mathbf{x}) &= ln\;[p(\mathbf{x}|\omega_i)p(\omega_i)] \qquad (5.21)\\ &= ln\;p(\mathbf{x}|\omega_i) \;+\; ln\;p(\omega_i)\\ &= ln\;p(\omega_i) \;-\; \frac{n}{2}ln\;2\pi \;-\; \frac{1}{2}ln\;|\mathbf{C}_i| \;-\; \frac{1}{2}[\mathbf{x}-\mathbf{m}_i]^t\mathbf{C}_i^{-1}[\mathbf{x}-\mathbf{m}_i] \end{aligned}$$

I dette udtryk kan vi ignorere konstanten $\frac{n}{2}ln2\pi$, fordi dette udtryk er uafhængigt af i. Diskriminantfunktionerne bliver herved:

$$d_i(\mathbf{x}) \;=\; ln\;p(\omega_i) \;-\; \frac{1}{2}ln|\mathbf{C}_i| \;-\; \frac{1}{2}[\mathbf{x}-\mathbf{m}_i]^t\mathbf{C}_i^{-1}[\mathbf{x}-\mathbf{m}_i] \qquad i=1,\ldots,M \tag{5.22}$$

Ovenstående udtryk angiver Bayes diskriminantfunktionerne for normalt fordelte mønstre. Væsentlig ved udledningen af udtrykket var, ud over antagelsen om normalitet, antagelsen at tabs-funktionen var givet ved (5.14). Det ses af udtrykket, at diskriminantfunktionerne er anden ordens polynomier i $\mathbf{x}$. Det bedste en Bayes klassifikator kan gøre er således at skille klasserne ved denne type funktioner. Dette er imidlertid den optimale form, hvis klasserne virkelig er normalfordelte. Det er desuden det teoretisk optimale valg for en lang række af andre fordelingsfunktioner (se f.eks. [26]). Nedenstående skitseres en række specialtilfælde:

Antag først at kovariansmatricerne er ens for alle klasser, dvs. $\mathbf{C}_i = \mathbf{C}_j = \mathbf{C}$. Ved at ignorere alle led, der er uafhængige af i, fra diskriminantfunktionen fås:

$$d_i(\mathbf{x}) \;=\; ln\;p(\omega_i) \;+\; \mathbf{x}^t\mathbf{C}^{-1}\mathbf{m}_i \;-\; \frac{1}{2}\,\mathbf{m}_i^t\mathbf{C}^{-1}\mathbf{m}_i \qquad i=1,\ldots,M \tag{5.23}$$

Det ses at diskriminantfunktionerne i dette tilfælde bliver lineære. Hvis der yderligere gælder at kovariansmatricen er lig identiteten, $\mathbf{C} = \mathbf{I}$, samt at de a priori sandsynligheder for at observationerne stammer fra de M klasser er ens, dvs. $p(\omega_i) = 1/M$, da reduceres udtrykket yderligere til:

$$d_i(\mathbf{x}) \;=\; \mathbf{x}^t\mathbf{m}_i \;-\; \frac{1}{2}\,\mathbf{m}_i^t\mathbf{m}_i \qquad i=1,\ldots,M \tag{5.24}$$

Dette udtryk svarer til diskriminantfunktionerne for minimal afstands klassifikation, under brug af det euclidiske afstandsmål, hvor hver prototype er bestemt ved middelværdivektoren $\mathbf{m}_i$.

Eksempel

Det er vanskeligt at give et overskueligt eksempel på ovenfor beskrevne statistiske tilgang til klassifikationsproblemet, fordi dette (for at være realistisk) kræver at antallet af stikprøver er stor. Ikke desto mindre gives her et sådant eksempel. Betragt nedenstående grupper af 12 observationer, hver beskrevet ved en featurevektor af dimension 3, og hver manuelt klassificeret til én af 2 klasser.

observation	klasse	x_1	x_2	x_3
1	ω_1	3	0	0
2	ω_1	4	3	1
3	ω_1	5	1	2
4	ω_1	6	1	0
5	ω_1	7	0	0
6	ω_1	5	1	3
7	ω_2	-1	15	0
8	ω_2	0	9	1
9	ω_2	1	3	3
10	ω_2	1	10	5
11	ω_2	-1	1	7
12	ω_2	0	4	8

På basis af ovenstående observationer antages, at de a priori sandsynligheder for de to klasser er ens, dvs. $p(\omega_1) = p(\omega_2) = \frac{1}{2}$. Vi kan estimere middelværdivektorene og kovariansmatricerne til:

$$\mathbf{m_1} = (5,1,1) \qquad \text{og} \qquad \mathbf{m_2} = (0,7,4)$$

$$\mathbf{C}_1 = \frac{1}{5}\begin{pmatrix} 160 & 28 & 29 \\ 28 & 12 & 8 \\ 29 & 8 & 14 \end{pmatrix} \quad \text{og} \quad \mathbf{C}_2 = \frac{1}{5}\begin{pmatrix} 4 & -3 & 1 \\ -3 & 432 & 107 \\ 1 & 107 & 148 \end{pmatrix}$$

Disse matricer har de inverse:

$$\mathbf{C}_1^{-1} \approx \begin{pmatrix} 0.0607 & -0.0934 & -0.0724 \\ -0.0934 & 0.8168 & -0.2732 \\ -0.0724 & -0.2732 & 0.6632 \end{pmatrix} \quad \text{og} \quad \mathbf{C}_2^{-1} \approx \begin{pmatrix} 1.2645 & 0.0133 & -0.0181 \\ 0.0133 & 0.0142 & -0.0104 \\ -0.0181 & -0.0104 & 0.0414 \end{pmatrix}$$

og der gælder at $|\mathbf{C}_1| = 1712.8$ og $|\mathbf{C}_2| = 41508.4$. Vi har derfor at:

$$ln\ \omega_1 - \frac{1}{2} ln|\mathbf{C}_1| \approx -4.4161$$
$$ln\ \omega_2 - \frac{1}{2} ln|\mathbf{C}_2| \approx -6.0100$$

Lad $\mathbf{x} = (x_1, x_2, x_3)$ være en ny observation. Bayes diskriminantfunktioner er da givet ved:

$$d_1(\mathbf{x}) = -4.4161 - 0.5 \begin{bmatrix} x_1 - 5.0 \\ x_2 - 1.0 \\ x_3 - 1.0 \end{bmatrix}^t \begin{pmatrix} 0.0607 & -0.0934 & -0.0724 \\ -0.0934 & 0.8168 & -0.2732 \\ -0.0724 & -0.2732 & 0.6632 \end{pmatrix} \begin{bmatrix} x_1 - 5.0 \\ x_2 - 1.0 \\ x_3 - 1.0 \end{bmatrix}$$

og

$$d_2(\mathbf{x}) \;=\; -6.0100 \;-\; 0.5 \begin{bmatrix} x_1 - 0.0 \\ x_2 - 7.0 \\ x_3 - 4.0 \end{bmatrix}^t \begin{pmatrix} 1.2645 & 0.0133 & -0.0181 \\ 0.0133 & 0.0142 & -0.0104 \\ -0.0181 & -0.0104 & 0.0414 \end{pmatrix} \begin{bmatrix} x_1 - 0.0 \\ x_2 - 7.0 \\ x_3 - 4.0 \end{bmatrix}$$

Antag nu konkret at $\mathbf{x} = (x_1, x_2, x_3) = (2, 5, 4)$. Ved indsættelse og udregning fås at:

$$\begin{aligned} d_1(\mathbf{x}) &= -12.05 \\ d_2(\mathbf{x}) &= -8.54 \end{aligned}$$

Bayes klassifikationsfunktion vil derfor anbefale, at klassificere den ny observation til klasse ω_2, da $d_2(\mathbf{x}) > d_1(\mathbf{x})$.

Eksempel slut

5.5 Neurale netværk

Dette afsnit skrives ikke i år.

5.6 Opgaver

1. Vis at ingen observationer i det grå område i figur 6.2 kan klassificeres ved brug af de i figuren viste beslutningsfunktioner.

2. Implementér en klyngeanalyse af én-dimernsionelle data baseret på *k-means*-algoritmen. Lad antallet k af prototyper være en parameter til programmet. Diskutér, implementér, og afprøv nogle forskellige måder hvorpå de initielle prototyper udvælges.

3. Lad A = (5,7), B = (-2,5), og C = (1,-1) være tre prototyper af klasser for mønstre karakteriseret ved to features. Opstil diskriminantfunktionerne ved brug af mindste afstands klassifikatorer og det euclidiske afstandsmål. Tegn funktionerne.

4. Med udgangspunkt i eksemplet om transmition af signaler skal likelihoodfunktionen l_{12} udregnes under antagelse at støjen er eksponentialfordelt med parameter λ. Tegn funktionen og sammenlig med den i eksemplet udledte funktion. Antag at $\theta_{12} = 1$. Hvad bliver grænsen for klassifikation mellem de to prototyper.

5. Antag at der er givet 3 prototyper (i 2 features), hver karakteriseret ved deres middeværdivektor of deres estimerede kovariansmatrice. Lad disse være givet ved:

$$\mathbf{m}_1 = (6, 7) \quad \mathbf{m}_2 = (3, 5) \quad \mathbf{m}_3 = (7, 3)$$

$$\mathbf{C}_1 = \begin{pmatrix} 0.6 & 0.0 \\ 0.0 & 0.6 \end{pmatrix} \quad \mathbf{C}_2 = \begin{pmatrix} 1.0 & -1.0 \\ -1.0 & 1.6 \end{pmatrix} \quad \mathbf{C}_3 = \begin{pmatrix} 0.8 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$$

Antag at alle tre mønstre er normalfordelte. Opstil diskriminantfunktionerne for Bayes klassifikationen. Afgør hvilke klasser som stikprøverne (3,2), (6,5), og (10,6) vil blive klassificeret til.

6. Antag at der er givet 3 prototyper (i 2 features), hver karakteriseret ved deres middeværdivektor of deres estimerede kovariansmatrice. Lad disse være givet ved:

$$\mathbf{m}_1 = (3,2) \quad \mathbf{m}_2 = (9,2) \quad \mathbf{m}_3 = (6,7)$$

$$\mathbf{C}_1 = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \quad \mathbf{C}_2 = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \quad \mathbf{C}_3 = \begin{pmatrix} 2.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$

   Antag at alle tre mønstre er normalfordelte. Hvilken klasse vil stikprøven (6,3) blive klassificeret til:
   1) Klasse 1
   2) Klasse 2
   3) Klasse 3
   4) Ingen af de tre klasser
   5) Alle tre klasser
   6) En tilfældig af klasserne
   7) Det ved jeg ikke

7. Otgag et billede (i billedlaboratoriet) af en række søm, skruer og møtrikker e.lign. For hvert objekt vælges en prototype ved et delbillede, der kun viser dette objekt. Benyt template matching ved korrelation til at markere alle forekomster i det store billede af de identificerede og klassificerede objekter.

8. Benyt en simpel segmenteringsalgoritme, og karakteriser hvert fundet segment ved segmentets størrelse, form mv. Vælg et hensigtsmæssigt billede, f.eks. som i ovenstående opgave. Opstil et klassifikationstræ til klassifikation af hvert segment og klassificer alle segmenter. Marker alle fundne forekomster af klassificerede objekter i billedet.