Oppgave 1

Innledning

a) Skissér et typisk mønstergjenkjenningssystem, og forklar hva de ulike delene i systemet gjør.

b) Forklar hva som menes med et *treningssett*, og gi eksempler på hvordan et slikt datasett kan brukes til å trene opp en klassifikator.

c) Forklar hva som menes med *diskriminantfunksjoner* og hvordan slike funksjoner brukes til å klassifisere objekter.

d) Forklar hva som menes med et *testsett*, og redegjør for hvordan og hvorfor man bør bruke et slikt datasett.

Oppgave 2

Beslutningsteori

a) Gjør rede for begrepene *klassebetinget sannsynlighetstetthet*, *a priori sannsynlighet* og *a posteriori sannsynlighet* og sett opp *Bayes regel* (Bayes formel) som knytter disse størrelsene sammen.

b) Forklar kort hva som menes med *handlinger* (actions) og *kostnader* (tap) knyttet til ulike handlinger. Redegjør for hvordan kostnader kan inngå i løsningen av et klassifiseringsproblem.

c) La $R(\alpha_i|\boldsymbol{x})$ være *betinget risk* for en gitt handling α_i og en gitt egenskapsvektor $\boldsymbol{x}$. Sett opp et uttrykk for denne størrelsen ved hjelp av kostnader og a posteriori sannsynligheter for klassene i problemet. La a være antall mulige handlinger og $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ være kostnaden (tapet) forbundet med handling α_i for $i = 1, \ldots, a$, når sann klasse for objektet representert ved $\boldsymbol{x}$ er ω_j. Forklar hvilket valg av handling som leder til minimum *total risk* (minimum forventet tap) og formulér den tilhørende beslutningsregelen.

d) I et éndimensjonalt (univariat) toklasseproblem med $a = c = 2$ (antall handlinger lik antall klasser) er fordelingsfunksjonene for egenskapen x gitt ved de univariate normalfordelingene $N(\mu_1, \sigma^2)$ for klasse ω_1 og $N(\mu_2, \sigma^2)$ for klasse ω_2. Vis at desisjonsgrensen (terskelen) x_0 som minimaliserer den totale risken er gitt ved

$$x_0 = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_1 - \mu_2} \ln \left[\frac{(\lambda_{12} - \lambda_{22}) P(\omega_2)}{(\lambda_{21} - \lambda_{11}) P(\omega_1)} \right].$$

e) Vis hvordan et spesielt valg av kostnader leder til *minimum-feilrate* klassifisering og formulér beslutningsregelen i dette tilfellet.

f) Hva blir desisjonsgrensen x_0 med dette spesielle valget av kostnader? Lag en skisse som illustrerer feilraten og viser plasseringen av desisjonsgrensen for et tilfelle med like a priori sannsynligheter.

Oppgave 3

Parametriske metoder

a) Beskriv *maksimum-likelihood* metoden for estimering av parametervektoren $\boldsymbol{\theta}$ i en antatt fordelingsfunksjon $p(\boldsymbol{x}|\boldsymbol{\theta})$ ved ledet læring, og utled et likningssystem for estimatet av $\boldsymbol{\theta}$ basert på et sett av treningssampler (egenskapsvektorer) $\boldsymbol{x}_k$, $k = 1, \ldots, n$ trukket fra den aktuelle fordelingsfunksjonen. Hvilken forutsetning må man gjøre om disse samplene?

b) Finn maksimum-likelihood estimatet av parameteren θ i den univariate fordelingen gitt ved

$$p(x|\theta) = \frac{1}{2}\theta^3 x^2 \mathrm{e}^{-\theta x},$$

der $x \geq 0$ og $\theta > 0$. La treningssettet være gitt ved $\mathscr{X} = \{x_1, ..., x_n\}$.

Oppgave 4

Lineære diskriminantfunksjoner

a) Sett opp en lineær diskriminantfunksjon $g(\boldsymbol{x})$ for et toklasseproblem, forklar størrelsene som inngår og gjør rede for hvordan diskriminantfunksjonen brukes til klassifisering av objekter. Anta at egenskapsrommet har dimensjon d.

b) Omskriv diskriminantfunksjonen til *utvidet* form, som produktet av en utvidet vektvektor $\boldsymbol{a}$ med en utvidet egenskapsvektor $\boldsymbol{y}$, og forklar hva som inngår i $\boldsymbol{a}$ og $\boldsymbol{y}$. Hvilken dimensjon har det utvidede egenskapsrommet?

c) Gjør rede for minste kvadraters metode til trening av den utvidede vektvektoren, og vis hvordan man kan komme frem til en minste kvadraters løsning ved hjelp av *Pseudoinvers*-metoden.

d) Anta et treningssett som består av de univariate samplene $\mathscr{X}_1 = \{1, 2, 3\}$ fra ω_1 og $\mathscr{X}_2 = \{5, 6, 7\}$ fra ω_2. Bruk pseudoinvers-metoden til å finne den utvidede vektvektoren for dette toklasseproblemet. La marginvektoren $\boldsymbol{b}$ ha kun enere som komponenter. Hva blir desisjonsgrensen (terskelen mellom klassene) i dette tilfellet?

e) Lag en skisse som viser treningssettet, desisjonsgrensen og vektvektoren i det utvidede egenskapsrommet.

Oppgave 5

Ikke-ledet læring

a) Hva er det som kjennetegner ikke-ledet læring (i motsetning til ledet læring), og hva menes med en blandingstetthet?

b) Sett opp blandingstettheten for et toklasseproblem uttrykt ved tetthetsfunksjonene og klassenes a priori sannsynligheter.

c) Her skal vi se på et *univariat* toklasseproblem. Bruk maksimum-likelihood metoden til å vise at likningssystemet for parametervektorene til de to klassene kan skrives som

$$\sum_{k=1}^{n} P(\omega_i|x_k,\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}_i} \ln p(x_k|\omega_i,\boldsymbol{\theta}_i) = 0, \quad i = 1,2,$$

når klassenes a priori sannsynligheter forutsettes kjent. Her er $P(\omega_i|x_k,\boldsymbol{\theta})$ a posteriori sannsynlighet for klasse ω_i i punktet x_k. Treningsettet er gitt ved $\mathscr{X} = \{x_1,...,x_n\}$.

d) Anta videre at klassene er normalfordelte med like a priori sannsynligheter og standardavvik lik én for begge klasser, men med ukjente forventningsverdier μ_1 og μ_2. Utled et likningssystem for disse forventningsverdiene og foreslå en løsningsmetode.