Oppgaver, STK4040, uke 48

27. november 2005

Oppgave 1

Filene `datasett_48_tren.txt` og `datasett_48_test.txt` inneholder data fra gjenkjenning av håndskrevne sifre. De første fem kolonnene inneholder ulike egenskaper som er avledet av sifrene (0, 3 og 8). Den siste kolonnen er sifferet. `datasett_48_tren.txt` inneholder 200 observasjoner av hvert siffer, og skal brukes til trening. `datasett_48_tren.txt` inneholder ca. 200 observasjoner av hvert siffer, og skal brukes som testsett.

a) Eksplorativ dataanalyse

Se på treningsdataene vha. f.eks. PCA og ulike plott. Har de noen spesielle egenskaper/strukturer?

b) LDA

Konstruer en LDA på treningsdataene. Test den på testdatasettet og se på feilraten og forvirringsmatrisen («confusion matrix»).

Ekstraoppgave: Plott testdataene i koordinatsystemet utspent av diskriminantfunksjonene. Sammenlign med et scoreplott fra PCA. (Hvorfor) er det forskjellig?

c) QDA

Konstruer en QDA på treningsdataene og test den på testdatasettet. Gjør den det bedre eller annerledes på noen måte? Hvorfor? Var det (u)ventet?