Opgaver 1

Nedenstående opgaver løses ved første selv-studie gang i Data Mining kurset. Vi gennemgår opgaverne i lektion 2.

Hvis I finder fejl eller uklarheder så skriv til mig (`tvede@math.aau.dk`) - det er set før.

PCA

1. Lav en PCA analyse af `data(crabs, package = "MASS")`, hvor I benytter korrelations matricen i stedet for at log-transformere data inden rotationen.

   Plot parvist de første tre PC variable i mod hinanden hvor punkterne identificerer krabbernes køn og art.

MDS

2. I datasættet `aims_freq.csv` på moodle, findes allelfrekvenser for en række forskellige humane populationer genotypet på 128 AIMs (Ancestry Informative Markers). AIMs er genetiske Single Nucleotide Polymorphisms (SNP) markører, dvs. kun en enkelt DNA base (`A`, `C`, `T` eller `G`) varierer fra individ til individ. En SNP navngives typisk som `rs123456`, hvor `rs` er kort for "Reference SNP cluster ID" og `123456` kan linkes til en specifik position på genomet i en database (`dbSNP`). Typisk kan en SNP kun antage to tilstande, fx. `A` og `T`. AIMs kan fx benyttes til at estimere den geografiske oprindelse af en DNA profil.

   Lav en MDS analyse af allelfrekvenserne og visualiser resultatet i et to dimensionalt plot. Sammenlign placeringen af populationerne med et geografisk bestemt kort:

   ```
   library(ggplot2)
   ggplot() + borders("world", colour="gray50", fill="gray50") +
     geom_text(data = aims_freq, aes(x = long, y = lat, label = pop))
   ```

   Hint: Brug funktionen `dist` på søjlerne der begynder med `rs`.

C4.5

3. Vi kigger igen på golf datasættet fra lærebogen.

   (a) Kig på træningsdata fra "golf.csv" (på moodle) og analyser med C5.0. Kontroller at resultatet er som i bogen. Bemærk at datasættet indeholder både kategorisk og numerisk udgave af temperatur og luftfugtighed (hhv. 'temperature' og 'temp' samt 'humidity' og 'humid'). Resultatet skulle gerne være det samme om I benytter den ene type frem for den anden.

   (b) Undersøg hvordan træet bliver ved at bruge halvdelen af data som træningssæt og resten som testsæt. Dette kan gøres manuelt eller vha. `C5.0Control`-funktionen. Hvad er trænings- og testfejlen?

   Bemærk at der benyttes tilfældig opdeling af data - foretag analysen flere gange for at se variationen.

4. Konstruer en ny respons variabel så den indeholder både information om `sp` og `sex` i `crabs` data, fx. `sp_sex`.

(a) Brug `C5.0` i R til klassificere `sp_sex` hvor I bruger de numeriske variable i `crabs`-datasættet.

(b) Gentag (a), men benyt i stedet de PCA roterede data fra 1. som attributer.

(c) Plot de to klassifikationstræer. Er der forskelle på træernes kompleksitet og nøjagtighed?

5. Datasættet `credit.csv` på moodle indeholder historiske data fra 1000 tyske bankkunder. Af interesse er at lære et klassifikationstræ som kan identificere kunder som har risiko for at misligholde deres lån (`default`).

   (a) Opdel tilfældigt data i 90% til træningsdata og 10% testdata. Efter se at fordelingen af `default` er nogenlunde end i de to datasæt (ellers gentag opdelingen).

   (b) Fit et klassifikationstræ på træningsdata og predikter på testdata. Sammenlign prediktioner med sande værdier - hvad er fejlraten på hhv. trænings- og testdata?

   (c) Senere i kurset kommer vi til at snakke om boosting. Benyt argumentet `trials = 10` til at foretage 10 boosting iterationer. Mindsker dette prediktionsfejlen op trænings- og testdata?

   (d) Det vurderes at det er fem gange så omkostningstungt for banken at misklassificere en dårlig bankkunde som god, end det er at klassificere en god kunde som dårlig. Brug `cost` argumentet til at angive dette.

6. Entropien er for diskrete fordelinger givet ved $H(x) = -\sum_{i=1}^{c} p_i \log_2 p_i$. Dette generaliseres naturligt for kontinuerte variable til $H(x) = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx = -\mathbb{E}(\ln(p(x)))$. Antag $x \sim N(\mu, \sigma^2)$ og udled $H(x)$.

7. Implementer `entropi`, `gain` og `gainRatio` i R (bemærk at `entropi.R` på Moodle indeholder et eksempel på hvorledes dette løses vha. `table`-funktionen. Bruges `lapply` og `split` på passende vis kan dette gøres tilsvarende for `gain`).