DATA-ANALYYSI: UMARU IMPACT STUDY

Datan kuvaus

UMARU IMPACT Study on Massachusettsin yliopiston AIDS-tutkimuksen yksikön teettämä "selviytymisanalyysi" (engl. *survival analysis*), jossa huumevieroitushoidon tehokkuutta tutkittiin viiden vuoden ajan kahden satunnaisotoksen avulla:

1. Lyhyt hoitojakso: Hoitojakson pituus 3 tai 6 kk. Koehenkilöille järjestettiin terveysvalistusta. Heille opetettiin myös keinoja raittiina pysymiselle, kuten riskien tunnistamista ja riskitilanteista selviytymisen menetelmiä
2. Pitkä hoitojakso: Hoitojakson pituus 6-12 kk. Koehenkilöillä oli selkeät päivärutiinit, heillä oli hoitojakson aikana hyvin järjestäytynyt elämäntyyli ja he osallistuivat yhteiskunnalliseen toimintaan

Tällaisessa analyysissä yleensä selvitetään, miten eri muuttujien arvot vaikuttavat jonkin tapahtuman todennäköisyyteen. Alla on listattuna selvityksessä koehenkilöistä luotujen profiilien muuttujat:

1. Tunniste (järjestysluku)
2. Ikä tarkasteluhetkellä
3. Beckin masennustestin tulos hoitojakson alettua
4. Suonensisäisten huumeiden käyttöhistoria
5. Aiempien lääkehoitojen lukumäärä
6. Etnisyys
7. Satunnaisotannan joukko (hoitojakson pituus)
8. Hoitopaikka
9. Ei käyttänyt huumeita hoitojakson jälkeen 12 kuukauteen

```
Description of Variables in the UMARU IMPACT Study
Described in Section 1.6.4 page 26

Variable        Description                     Codes/Values            Name
1               Identification Code             1-575                   ID
2               Age at Enrollment               Years                   AGE
3               Beck Depression Score at        0.000-54.000            BECK
                Admission
4               IV Drug Use History             1 = Never,              IVHX
                at Admission                    2 = Previous 3 = Recent
5               Number of Prior Drug            0-40                    NDRUGTX
                Treatments
6               Subject's Race                  0 = White, 1 = Other    RACE
7               Treatment Randomization         0 = Short, 1 = Long     TREAT
                Assignment
8               Treatment Site                  0 = A, 1 = B            SITE
9               Remained Drug Free              1 = Remained Drug Free  DFREE
                for 12 Months                   0 = Otherwise
```

```
clear all

% Ladataan paikallinen datasetti:

load("polku/tiedostoon/uis.dat", '-ascii')

% tulosta tiedoston uis.dat sisältö:

uis
```

Dataa kuvailevassa taulukossa on ilmoitettuna yhdeksän muuttujaa. Lukumäärä täsmää dat-tiedoston sarakkeiden lukumäärän kanssa. Myös muuttujien arvot sopivat kuvauksiinsa hyvin:

	1	2	3	4	5	6	7	8	9
1	1.0000	39.0000	9.0000	3.0000	1.0000	0	1.0000	0	0
2	2.0000	33.0000	34.0000	2.0000	8.0000	0	1.0000	0	0
3	3.0000	33.0000	10.0000	3.0000	3.0000	0	1.0000	0	0
4	4.0000	32.0000	20.0000	3.0000	1.0000	0	0	0	0
5	5.0000	24.0000	5.0000	1.0000	5.0000	1.0000	1.0000	0	1.0000

Perustunnusluvut

Kopioidaan muuttujien arvojen keskiarvot, mediaanit ja moodit omiin muuttujiinsa:

```
keskiarvot = mean(uis);

mediaanit = median(uis);

moodit = mode(uis);
```

Poistetaan vektoreista epäolennaiset järjestyslukujen tunnusluvut:

```
keskiarvot(:,1) = [];
```

```
mediaanit(:,1) = [];
moodit(:,1) = [];
```

Taulukoidaan tunnusluvut sisältävät vektorit, otsikoidaan muuttujat ja otsikoidaan tunnusluvut:

```
otsikot = {'AGE','BECK','IVHX','NDRUGTX','RACE','TREAT','SITE','DFREE'};
tunnuslukutaulukko = array2table([keskiarvot;mediaanit;moodit],
'VariableNames', otsikotMuuttujat, 'RowNames', otsikotTunnusluvut)
```

tunnuslukutaulukko = 3×8 table

		AGE	BECK	IVHX	NDRUGTX	RACE	TREAT	SITE	DFREE
1	keskiarvot	32.3826	17.3674	2.0348	4.5426	0.2522	0.4974	0.3043	0.2557
2	mediaanit	32	17	2	3	0	0	0	0
3	moodit	33	23	3	1	0	0	0	0

Koehenkilöt olivat keskimäärin n. 32-vuotiaita, ja heidän ikänsä jakautuvat mediaanista päätellen melko tasaisesti tämän lukeman puolin ja toisin. 33-vuotiaita oli ryhmässä eniten.

Koehenkilöiden masennustestien tulokset olivat keskimäärin 17 pisteen luokkaa, samoissa lukemissa mediaanipisteiden kanssa. Voitaneen olettaa, että pisteet jakautuvat melko tasaisesti molemmin puolin tätä lukemaa. Pistemäärän ylittäessä 20 puhutaan selvästi alentuneesta mielialasta.

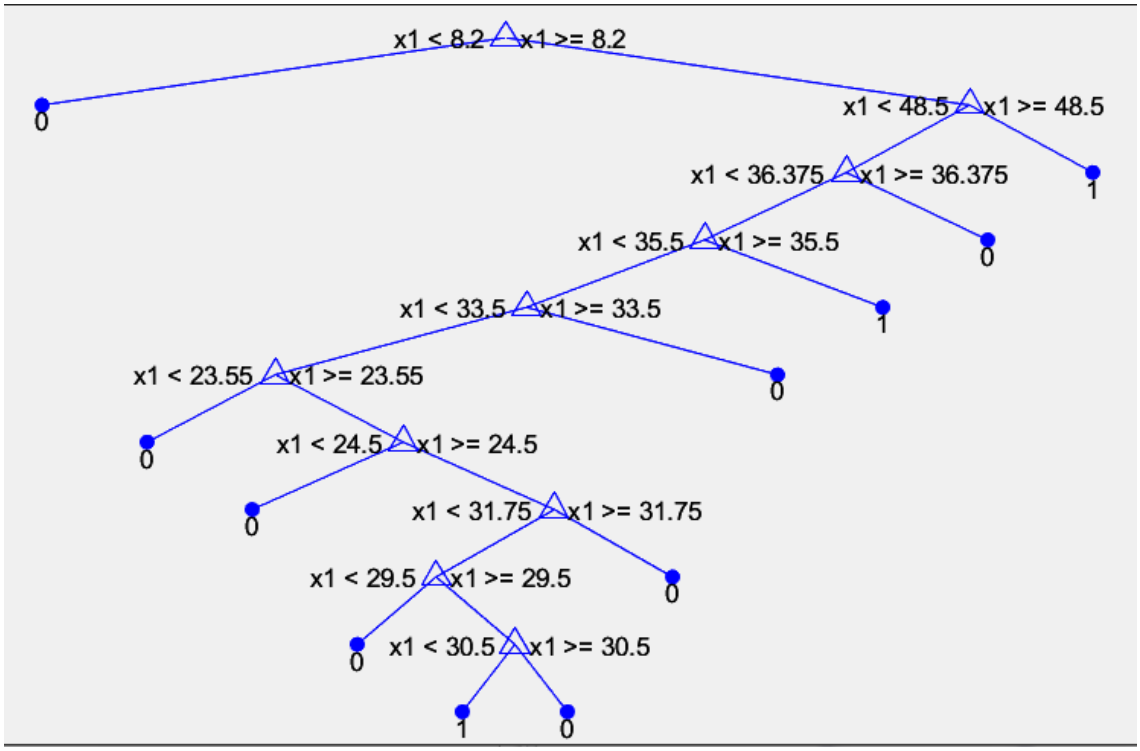
Koehenkilöitä datasarjassa on lyhyen aikajakson ryhmästä ja pitkän aikajakson ryhmästä lähes yhtä monta. Vain neljännes kaikista koehenkilöistä pysyi päihteettömänä 12 kk hoitojakson päättymisen jälkeen.

Luokittelutehtävä

Koehenkilön huumeidenkäytön jatkuminen on muuttujista kenties ilmeisimmin se, jota halutaan selittää muiden muuttujien avulla. Yksinkertaisimmillaan voitaisiin vaikkapa testata, millainen vaikutus koehenkilön saaman Beckin masennustestin tuloksella on vuoden "kuivilla pysymiseen".

```
BECK = uis(:,3);
DFREE = uis(:,9);
```

```
puu = fitctree(BECK, DFREE);
view(puu, 'Mode','graph')
```



Malli ennustaa pisterajaksi 30,5. Jos pisteet ovat alle tämän, on ennuste suotuisa:

```
predict(puu, 30)
predict(puu, 31)
```

```
ans = 1
ans = 0
```

Ennusteen mukaan hoitojakson alussa vakavastikin masentunut koehenkilö voi pysyä vuoden raittiina hoitojakson jälkeen.

Tarkastellaan seuraavaksi, millaisella todennäköisyydellä koehenkilön päihteetön jakso onnistuu lyhyellä hoitojaksolla sekä pitkällä hoitojaksolla. Tämä voidaan selvittää jakamalla kunkin ryhmän onnistumiset kaikilla ryhmän tapauksien lukumäärällä:

```
TREAT = uis(:,7)
lyhytN = 0;
```

```
pitkaN = 0;
lyhytP = 0;
pitkaP = 0;

for i = 1:size(TREAT)
   a = TREAT(i);
   b = DFREE(i);
   if (a == 0)
       if (b == 0)
           lyhytN = lyhytN + 1;
       else
           lyhytP = lyhytP + 1;
       end
   elseif (a == 1)
       if (b == 0)
           pitkaN = pitkaN + 1;
       else
           pitkaP = pitkaP + 1;
       end
   end
end

kuivillaLyhyt = lyhytP / (lyhytN + lyhytP)
kuivillaPitka = pitkaP / (pitkaN + pitkaP)
```

```
kuivillaLyhyt = 0.2145
kuivillaPitka = 0.2972
```

Tulosten mukaan noin viidennes lyhyen hoitojakson koehenkilöistä pysyi vuoden päihteettömänä hoitojakson jälkeen. Pitkän hoitojakson koehenkilöistä lähes kolmasosa pysyi vuoden päihteettömänä.

Entä sitten muiden muuttujien vaikutus päihteettömän jakson toteutumiseen? Otetaan testiin selittäviksi muuttujiksi kaikki koehenkilön henkilökohtaiset ominaisuudet ja historia: Ikä, BDI-tulos, suonensisäisten huumeiden käyttöhistoria, aiemmat päihdehoidot ja etnisyys. Rajataan samalla ulkoiset tekijät pois. Selitettäväksi muuttujaksi valitaan jälleen päihteettömyyden onnistuminen. Luodaan tietojen avulla päätöspuu ja ennustetaan joidenkin keksittyjen henkilöiden tulokset. Henkilöt voisivat olla

- Aki: 22-vuotias, BDI-tulos 36, on kokeillut suonensisäisiä huumeita, on ollut päihdehoidossa aiemmin 2 kertaa, etniseltä alkuperältään aasialainen
- Hector: 36-vuotias, BDI-tulos 14, on lähiaikoina käyttänyt suonensisäisiä huumeita, on ollut päihdehoidossa aiemmin 16 kertaa, etniseltä alkuperältään afroamerikkalainen
- Julia: 14 vuotias, BDI-tulos 26, on joskus kokeillut suonensisäisiä huumeita, ei aiempaa kokemusta päihdehoidosta, etniseltä alkuperältään eurooppalainen (kaukasialainen)

```
koehenkilot = [uis(:,2), uis(:,3), uis(:,4), uis(:,5), uis(:,6)];
aki = [22, 36, 2, 2, 1];
hector = [36, 14, 3, 16, 1];
julia = [14, 26, 2, 0, 0];
puu = fitctree(koehenkilot, uis(:,9));
predict(puu, aki)
predict(puu, hector)
predict(puu, julia)
```

```
ans = 0
ans = 1
ans = 0
```

Ennusteiden mukaan Aki ja Julia pysyvät vuoden raittiina hoitojaksojensa jälkeen, mutta Hector ei.

Luokittimen arviointi

Toteutetaan päätöspuu-luokittimen toimivuuden arviointi. Mallin testaamista varten varataan datasta neljäsosa. Loput datasta on mallin opettamista varten. Opetuksessa ei saa käyttää testidataa. Tarkoitus on, että malli oppii yleistämään oppimaansa. Malli opetetaan opetusdatalla, ja tämän jälkeen mallin kykyä ennustaa kunkin testidatan profiilin tulos verrataan saman profiilin todelliseen tulokseen. Jos ennuste toteutuu, kasvatetaan muuttujan *toteutuu* arvoa. Toteutumisten osuus koko testidatan profiilien lukumäärästä antaa onnistumiskertoimen.

```
opetusdata = uis(1:floor(size(uis, 1)*0.75), 2:9);
testidata = uis(floor(size(uis, 1)*0.75)+1:size(uis, 1), 2:9);
puu = fitctree(opetusdata(:,1:7), opetusdata(:,8));
toteutuu = 0;

for i = 1:size(testidata, 1)
    if (predict(puu, testidata(i,1:7)) == testidata(i,8))
        toteutuu = toteutuu + 1;
    end
end

kerroin = toteutuu / size(testidata, 1)
```

```
kerroin = 0.5764
```

Onnistuneiden ennusteiden osuus kaikista testeistä on noin 58 %. Tästä päätellen malli ei ole erityisen luotettava, vaan tarjoaa ns. "50-50" onnistumismahdollisuuden.