

Kvalitatiivsete andmete analüüs
4. praktikum
Seose tugevus. Entroopiakordaja ja AUC

Loeme sisse kasutatava näidisandmestiku – TÜ tudengite küsitlemisel saadud andmed:

```
load(url("http://www-1.ms.ut.ee/mart/kvalitatiivsed/tudengl.RData"))
head(tudengl)
```

Tunneme huvi selle vastu, kui paljud tudengitest suitsetavad (mida kõik need käsud teevad?):

```
attach(tudengl)
table(suits)
prop.table(table(suits))
prop.table(table(suits))*100
round(prop.table(table(suits))*100,1)
barplot(prop.table(table(suits))*100, ylab="Protsent (%)",
        col=c("green", "pink", "red"))
```

Entroopiakordaja

Arvatavasti märkad, et mitu vaadeldud tunnust on statistiliselt sõltuvad suitsetamisest. Aga milliste tunnuste puhul on sõltuvus tugevam, milliste puhul nõrgem? Seda uurime entroopiakordajat (*uncertainty coefficient*, Theili U) kasutades.

Kuidas leida entroopiakordaja väärtust?

Leiame esmalt entroopiakordaja käsitsi – siis on ehk lootust ka mõista, mida ta mõõdab.

Arvutame esmalt tunnuse *suits* entroopia (mõõdetuna bittides):

```
Psuits=prop.table(table(suits))
H=sum(Psuits*log(1/Psuits, base=2)); H
```

Kui suure osa sellest entroopiast suudab näiteks ära kirjeldada tunnus sugu?

```
# Leiame tinglikud entroopiad mõlema soo jaoks:
Psuits_N =prop.table(table(sugu, suits),1)[1,]
Psuits_M =prop.table(table(sugu, suits),1)[2,]
H_N=sum(Psuits_N*log(1/Psuits_N, base=2))
H_M=sum(Psuits_M*log(1/Psuits_M, base=2))
Psugu=prop.table(table(sugu))
# Tinglik entroopia kokku:
H_tinglik=H_N*Psugu[1]+H_M*Psugu[2]
# entroopiakordaja (uncertainty coefficient):
(H-H_tinglik)/H
```

Võrdle:

```
library(DescTools)
UncertCoef(table(suits, sugu), direction = c("row"))
```

Ülesanne 1

Leia ise kui suure osa tunnuses suits olevast informatsioonist suudab ära kirjeldada tunnus olu. Võrdle saadud tulemust käsu UncertCoef abil saadud väärtusega!

Ülesanne 2

Vaata, milline oleks entroopiakordaja siis, kui prognoosiksime õlletarbimist selle järgi, kas tudeng suitsetab või mitte!

Tundlikkus, spetsiifilisus, ROC-kõver ja AUC

Loengus tutvusime ka teise seose tugevust iseloomustava näitajaga – ROC-kõvera aluse pindalaga (AUC). ROC-kõverast ja AUC-st räägitakse enamasti siis, kui uuritaval (prognoositaval) tunnusel on kaks võimalikku väärtust. Teeme enda elu õppimise ajaks veidi lihtsamaks ja muudame tunnus suits selliselt, et sellel oleks kõigest kaks väärtust:

```
table(suits)
suits2=factor(suits, labels=c("ei suitseta", "suitsetab", "suitsetab"))
table(suits, suits2)
```

Vaatame, kuidas suitsetajaks olemise tõenäosus muutub erineva õllelembusega tudengite seas:

```
prop.table(table(olu, suits2),1)
prop.table(table(olu, suits2),1)[,2]
pvek=prop.table(table(olu, suits2),1)[,2]
pvek
```

Näeme, et palju õlut joovate tudengite seas on suitsetajaid 85,7%, seevastu õlut mittejoovate tudengite seas kõigest 13,9%.

Teeme vektori, mis näitab igale tudengile tema õlletarbimise põhjal ennustatud tõenäosust olla suitsetaja:

```
table(olu, as.numeric(olu))
Psuits=pvek[as.numeric(olu)]
Psuits
cbind(tudengl, Psuits)[1:10,]
```

Alternatiivne variant, kuidas sellist vektorit teha:

```
mudel=glm( (suits2=="suitsetab") ~ olu3, family=binomial())
Psuits_alt=predict(mudel, newdata=tudengl, type="response")

# Veendume, et tulemused on samad
cbind(tudengl, Psuits, Psuits_alt)[1:10,]
```

Oletame, et lahterdaksime iga sellise tudengi suitsetajaks, kelle puhul õllejoomisharjumuste pealt ennustatud tõenäosus olla suitsetaja oleks suurem kui 0,5:

```
prognoosSuitsetaja=Psuits>0.5  
table(prognoosSuitsetaja)
```

Kui hea on selline prognoos?

Prognoosi (klassifitseerimise) headust kirjeldatakse enamasti tundlikkuse ja spetsiifilisuse abil:

Tundlikkus = $P(\text{prognoos õige} \mid \text{tegelikult suitsetaja})$
Spetsiifilisus = $P(\text{prognoos õige} \mid \text{tegelikult mitesuitsetaja})$

Leiame ka meie prognoosi jaoks tundlikkuse ja spetsiifilisuse:

```
table(suits2, prognoosSuitsetaja )  
prop.table(table(suits2, prognoosSuitsetaja),1)
```

Mittesuitsetajatest 97,8%
prognoosimie korrektselt
mittesuitsetajaks – meie testi
spetsiifilisus on 97,8%

```
> prop.table(table(suits2, prognoosSuitsetaja),1)  
           prognoosSuitsetaja  
suits2      FALSE      TRUE  
ei suitseta 0.97821782 0.02178218  
suitsetab   0.83225806 0.16774194
```

Suitsetajatest kõigest 16,8%
prognoosime olevat
suitsetaja – testi tundlikkus
on kõigest 16,8%

Kuna testi tundlikkus on niru, siis võime proovida seda parandada langetades piirmäära, kust alates me kedagi suitsetajaks peame:

```
prop.table(table(suits2, Psuits>0.2),1)
```

Saamegi testi tundlikkuse juba veidi paremaks – leiame üles 76% suitsetajatest (mis juhtub spetsiifilisusega?)

Kõik mõeldavad spetsiifilisused ja tundlikkused, mida võime saavutada:

```
# install.packages("pROC")  
library("pROC")  
roc_k6ver=roc(suits2=="suitsetab", Psuits)  
plot(roc_k6ver)  
auc(roc_k6ver)  
ci.auc(roc_k6ver) # Usalduspiirid - kui täpne on hinnang  
  
data.frame( roc_k6ver$thresholds, roc_k6ver$sensitivities,  
            roc_k6ver$specificities)
```

Ülesanne 3

Proгноosi tudengi sugu tema õlletarbimise järgi. Milline tuleb ROC-kõver, milline aga ROC-kõvera alune pindala?

Ülesanne 4

Kas tudengi sugu saab täpsemalt prognoosida õlletarbimise või suitsetamise järgi?