Umetna inteligenca 2021-2022

Seminarska naloga 1

Aljaž Hribar

30 November 2021

Klasifikacijski problem

Ocenjevanje in konstrukcija atributov

najprej faktoriziramo vse atribute, ki niso zvezni

```
ucna$regija<-as.factor(ucna$regija)
testna$regija<-as.factor(testna$regija)
ucna$namembnost<-as.factor(ucna$namembnost)
testna$namembnost<-as.factor(testna$namembnost)
ucna$oblacnost<-as.factor(ucna$oblacnost)
testna$oblacnost<-as.factor(testna$oblacnost)
summary(ucna)
```

```
##     datum               regija          stavba
##  Length:24125       vzhodna:11315   Min.   :  1.00
##  Class :character   zahodna:12810   1st Qu.: 39.00
##  Mode  :character                   Median : 79.00
##                                     Mean   : 87.49
##                                     3rd Qu.:135.00
##                                     Max.   :193.00
##                 namembnost       povrsina       leto_izgradnje   temp_zraka
##  izobrazevalna       :13301   Min.   :  329.3   Min.   :1903   Min.   :-7.20
##  javno_storitvena    : 2979   1st Qu.: 4106.6   1st Qu.:1950   1st Qu.:10.00
##  kulturno_razvedrilna: 3263   Median : 6763.3   Median :1970   Median :20.00
##  poslovna            : 3057   Mean   :10958.1   Mean   :1970   Mean   :19.15
##  stanovanjska        : 1525   3rd Qu.:14409.3   3rd Qu.:2000   3rd Qu.:28.30
##                               Max.   :79000.4   Max.   :2017   Max.   :41.70
##   temp_rosisca     oblacnost    padavine          pritisk         smer_vetra
##  Min.   :-19.400   0:3090    Min.   :-1.0000   Min.   : 997.2   Min.   :  0.0
##  1st Qu.: -2.800   2:8390    1st Qu.: 0.0000   1st Qu.:1011.9   1st Qu.: 70.0
##  Median :  2.800   4:4514    Median : 0.0000   Median :1015.9   Median :140.0
##  Mean   :  3.816   6:5126    Mean   : 0.3113   Mean   :1017.1   Mean   :156.6
##  3rd Qu.: 11.100   8:2950    3rd Qu.: 0.0000   3rd Qu.:1021.8   3rd Qu.:250.0
##  Max.   : 25.000   9:  55    Max.   :56.0000   Max.   :1040.9   Max.   :360.0
##  hitrost_vetra        poraba
##  Min.   : 0.000   Min.   :   0.00
##  1st Qu.: 2.100   1st Qu.:  53.48
##  Median : 3.600   Median : 112.90
##  Mean   : 3.756   Mean   : 224.55
##  3rd Qu.: 5.100   3rd Qu.: 215.41
```

```
##  Max.   :12.400   Max.   :2756.54
```

opazimo da lahko je smer vetra podana koz zvezni podatek ampak bi nam bila bolj uporabna kot diskretni zato jo faktoriziramo

```
table(ucna$smer_vetra)
```

```
##
##  brezveterje          jug   jugo_vzhod   jugo_zahod        sever severo_vzhod
##         2482         4210         3325         1044         1740         2670
## severo_zahod        vzhod        zahod
##         2831         3799         2024
```

iz atributa "datum" lahko generiramo nov atribut "season", ki nam pove letni čas meritve in atribut "vikend" ki nam pove ali je na ta datum bil vikend ali delovni teden

```
table(ucna$season)
```

```
##
##   Fall Spring Summer Winter
##   6741   4235   5140   8009
```

```
table(ucna$vikend)
```

```
##
## FALSE  TRUE
## 17184  6941
```

prav tako lahko iz atributa "poraba" izvlečemo atributa "dosedanja_povpreča" in "dosedanja_skupna" ki nam povesta kolikšna je povprečna in skupna poraba stavbe do vključno trenutnega datuma meritve

```
summary(ucna$dosedanja_povprecna)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    3.616   60.733  118.774  228.265  204.453 2196.688
```

```
summary(ucna$dosedanja_skupna)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      3.7   4531.1  12638.2  28874.0  30728.0 424149.1
```

sedaj lahko izločimo atribut stavba saj je za klasifikacijo odvečen atribut, ki bi samo kvaril modele

```
ucna$stavba<-NULL
testna$stavba<-NULL
```

z attrEval() funkcijo ocenimo atribute,

```
library(CORElearn)
```

```
## Warning: package 'CORElearn' was built under R version 4.0.5
```

```
sort(attrEval(namembnost ~ ., ucna, "Relief"), decreasing = TRUE)
```

```
## dosedanja_povprecna           povrsina             poraba    dosedanja_skupna
##         0.061888110         0.035969465         0.035938575         0.035583211
##      leto_izgradnje             regija            padavine              season
##         0.016055510         0.000000000        -0.000613520        -0.002901554
##             vikend          temp_zraka        temp_rosisca             pritisk
##        -0.005015544        -0.007857875        -0.011347804        -0.013097497
##           oblacnost          smer_vetra       hitrost_vetra               datum
##        -0.013471503        -0.015046632        -0.015718982        -0.057616580
```

```
sort(attrEval(namembnost ~ ., ucna, "ReliefFequalK"), decreasing = TRUE)
```

```
##      leto_izgradnje            povrsina              regija dosedanja_povprecna
##         0.242451468         0.166456481         0.139201471         0.136221147
##               poraba    dosedanja_skupna               datum          temp_zraka
##         0.115882015         0.105300619         0.102695830         0.082560008
##          smer_vetra              pritisk       temp_rosisca       hitrost_vetra
##         0.067548928         0.062092321         0.047547866         0.040365334
##           oblacnost              season              vikend            padavine
##         0.038386700         0.027151592         0.007127530         0.004231328
```

```
sort(attrEval(namembnost ~ ., ucna, "ReliefFexpRank"), decreasing = TRUE)
```

```
##      leto_izgradnje            povrsina              regija               datum
##         0.260707133         0.166830040         0.164866503         0.119054407
## dosedanja_povprecna              poraba          temp_zraka    dosedanja_skupna
##         0.115823728         0.095175641         0.094856759         0.085446344
##          smer_vetra              pritisk       temp_rosisca           oblacnost
##         0.079984195         0.074110398         0.056642911         0.048480491
##       hitrost_vetra              season              vikend            padavine
##         0.046881204         0.033750883         0.008384743         0.004979457
```

iz avaluacije atributov opazimo da imajo atributi ki opisujejo vremenske razmere in datum meritev le teh zelo majheno povezavo z namembnostjo stavbe zato jih lahko izločimo

```
ucna$temp_zraka<-NULL
ucna$pritisk<-NULL
ucna$temp_rosisca<-NULL
ucna$padavine<-NULL
ucna$hitrost_vetra<-NULL
ucna$smer_vetra<-NULL
ucna$oblacnost<-NULL
ucna$datum<-NULL
ucna$season<-NULL
ucna$vikend<-NULL
testna$temp_zraka<-NULL
testna$pritisk<-NULL
testna$temp_rosisca<-NULL
testna$padavine<-NULL
testna$hitrost_vetra<-NULL
testna$smer_vetra<-NULL
testna$oblacnost<-NULL
testna$datum<-NULL
testna$season<-NULL
testna$vikend<-NULL
ucna$dosedanja_skupna<-NULL
testna$dosedanja_skupna<-NULL
```

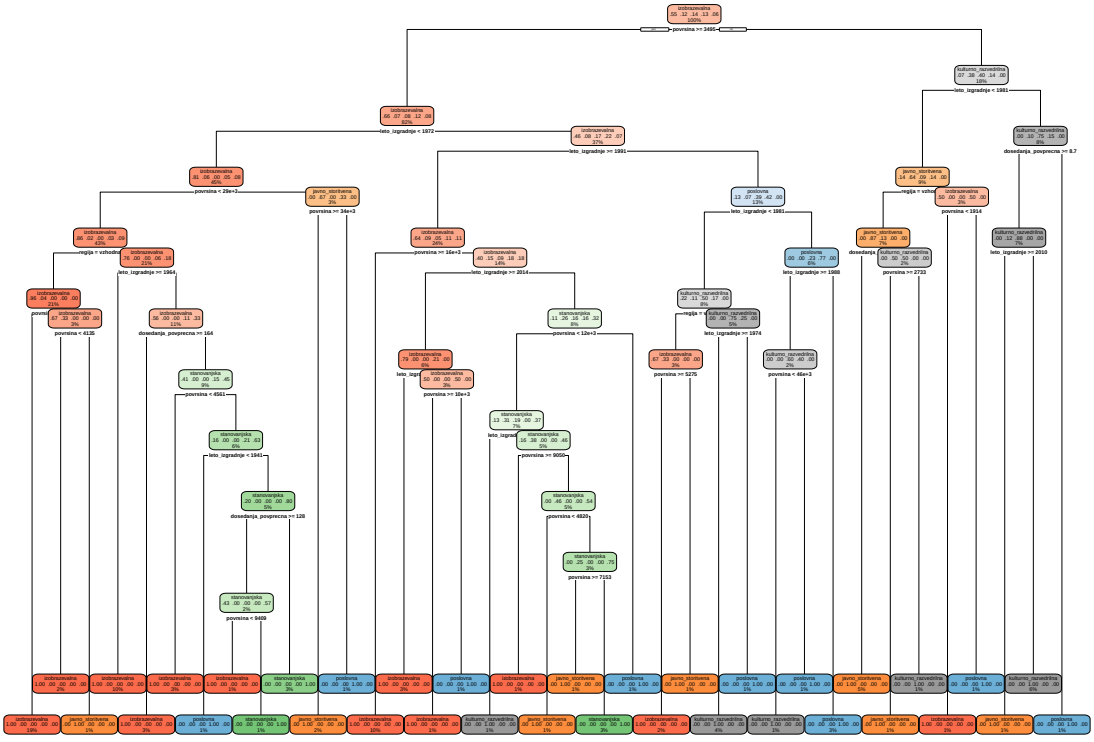
gradnja modelov

odločitveno drevo

najprej sem zgradil model z vsemi atributi

```
rpart.plot(dt)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



```
CA(observed,predicted)
```

```
## [1] 0.5093227
```

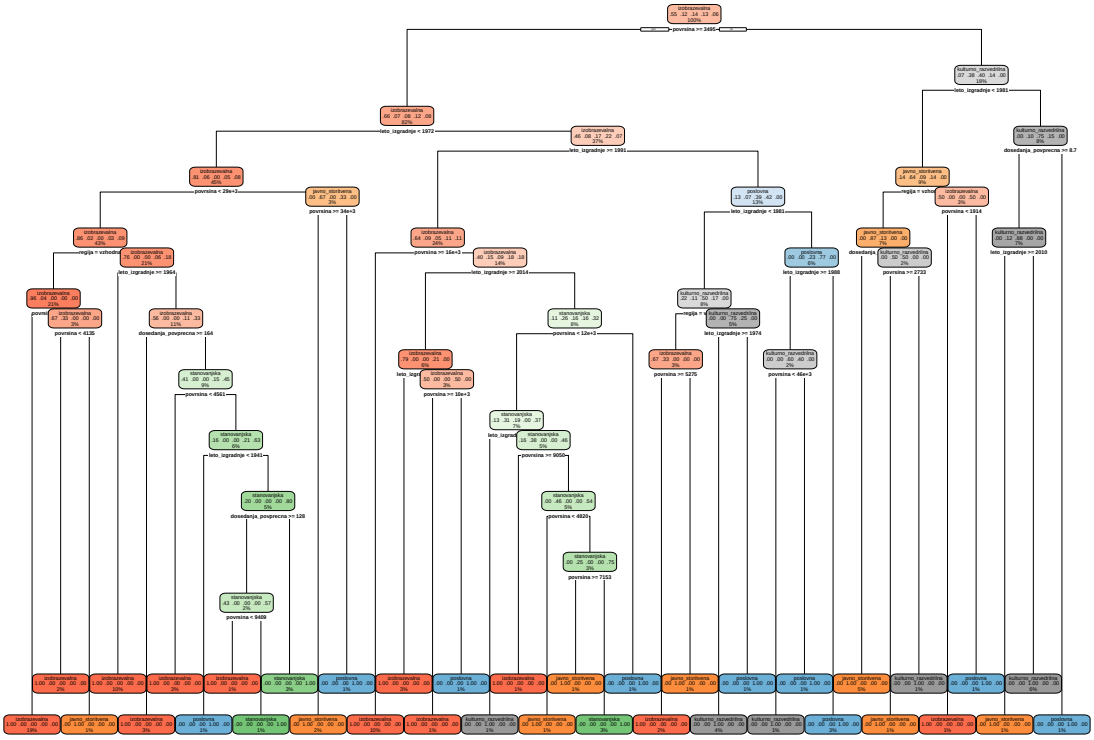
```
predMat <- predict(dt, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.9812315
```

nato pa še poiskusil maksimalizirati točnost z fukcijo wrapper(), ki je vrnila da so najboljšo točnost imeli atributi povrsina in leto_izgradnje z pričakovano napako 0.007543904

```
rpart.plot(dt)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



```
CA(observed,predicted)
```

```
## [1] 0.5005853
```

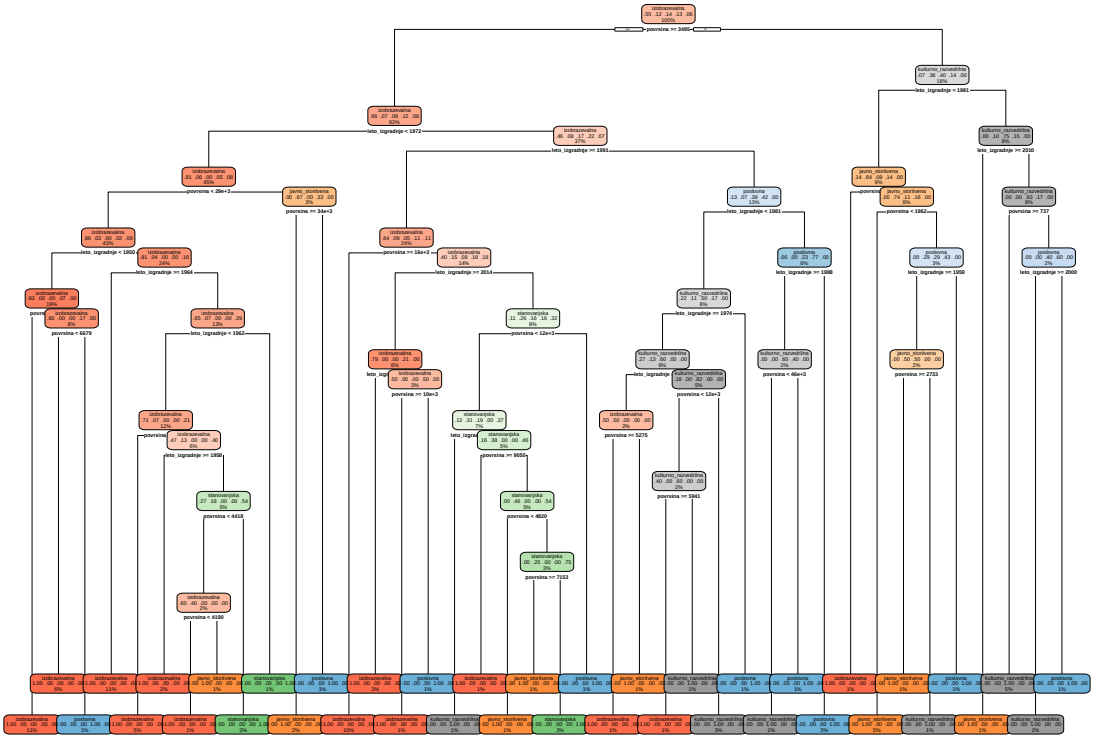
```
predMat <- predict(dt, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.9812315
```

ko sem pognal wraper() za minimizacijo "brier score" sem dobil podobne rezultate saj mi je funcija vrnila "best model: estimated error = 0.01390878 , selected feature subset = namembnost ~ povrsina + leto_izgradnje"

```
rpart.plot(dt)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



```
CA(observed,predicted)
```

```
## [1] 0.5005853
```

```
predMat <- predict(dt, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.9988294
```

naivni bayes

tukaj sem uporabil isti postoke kot pri gradnji drevesa in dobil naslednje rezultate:

z vsemi atributi

```
CA(observed, predicted)
```

```
## [1] 0.4540552
```

```
predMat <- predict(nb, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.7545656
```

wrapper z minimizacijo napake

```
CA(observed, predicted)
```

```
## [1] 0.4909699
```

```
predMat <- predict(nb, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.7259014
```

wrapper z minimizacijo brier

```
CA(observed, predicted)
```

```
## [1] 0.4909699
```

```
predMat <- predict(nb, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.7259014
```

k-najbližjih sosedov

```
CA(observed, predicted)
```

```
## [1] 0.4678512
```

```
predMat <- predict(knn, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 1.000356
```

naključno gozd

```
CA(observed, predicted)
```

```
## [1] 0.5141304
```

```
predMat <- predict(rf, testna, type = "prob")
brier.score(obsMat, predMat)
```

```
## [1] 0.7388644
```

SVM

```
brier.score(obsMat, predMat)
```

```
## [1] 0.6977981
```

```
CA(observed, predicted)
```

```
## [1] 0.5141304
```

Umetne nevronske mreže

najprej je bilo potrebno normalizirati zvezne atribute v učni in testni množici nato pa sem dobil te rezultate

```
CA(observed, predicted)
```

```
## [1] 0.5119147
```

```
predMat <- predict(nn, testna_scaled, type = "raw")
brier.score(obsMat, predMat)
```

```
## [1] 0.8076981
```

kobinirani modeli

za kombinirane modele sem se odločil uporabiti modele nevrenoske mreže, naibni bayes in SVM

najprej sem poskusil z glasovanjem in dobil

```
CA(observed, predicted)
```

```
## [1] 0.5484532
```

nato z uteženim glasovanje

```
CA(observed, predicted)
```

```
## [1] 0.4796405
```

nazadnje pa še z boostingom

```
CA(observed, predicted)
```

```
## [1] 0.5048913
```