Machine learning in R

Trainee sessie 02

Longhow Lam

Contents

---

1 Predictive modeling technieken

In R kan je veel verschillende predictive modellen fitten. We behandelen een paar in deze sessie. Lineaire regressie met de functie `lm`, logistische regressie met de functie `glm`, decision trees met de functie `rpart` en ensemble van trees met `ranger` en `xgboost`. Ook `h2o` zullen we kort aanstippen. Ik zal deze functies apart behandelen maar we zullen later in de sessie zien met het package `mlr` hoe je op een meer uniforme manier meerdere modellen kan proberen op een data set.

1.1 lineare regressie

We beginnen met simpele lineaire regressie, bruikbaar voor voorspel modellen waar de Target variable continu (numeric) is. We nemen als voorbeeld huizen prijs data die ik gescraped heb van jaap.nl. We willen de prijs van een huis voorspellen basis van een aantal input variabelen/kenmerken.

```
jaap = readRDS("data/Jaap.RDs")

modelout  = lm( prijs ~ kamers               , data = jaap)
modelout2 = lm( prijs ~kamers + Oppervlakte , data = jaap)

modelout
modelout2
```

Modeling functies in R retourneren objecten met van alles er nog wat in. De functie `lm` levert een object af van de klasse lm.

```
class(modelout)
names(modelout)
modelout$coefficients

summary(modelout)
plot(modelout)
```

1.1.1 formula objects

Modellen in R kan je specificeren met zogenaamde de formula objects. Hieronder zie je een aantal voorbeelden.

```
names(jaap)
jaap = jaap %>% mutate(PC1Positie = stringr::str_sub(PC,1,1))
f1 <- prijs ~ Oppervlakte + kamers + PC1Positie
m1 = lm(f1, data = jaap)
m1
summary(m1)


## interactie termen
f2 <- prijs ~ Oppervlakte*kamers*PC1Positie
m2 = lm(f2, data = jaap)
summary(m2)


##  termen weglaten
f3 <- prijs ~ Oppervlakte*kamers*PC1Positie - Oppervlakte:kamers:PC1Positie
m3 = lm(f3, data = jaap)
summary(m3)

## een target en de rest van de variabelen als inputs
f4 = prijs ~ . -PC6 -PC
m4 = lm(f4, data = jaap)
summary(m4)
```

Als je verschillende model objecten hebt gemaakt kan je de functie `anova` gebruiken om ze met elkaar te vergelijken. Dit gebeurt met behulp van F statistics.

```
anova(m1,m2, m3)
```

1.2 Splitsen in train en test

Het is gebruikelijk om een data set random te splitsen in een train en test set. Op de train set wordt een predictive model getraind. En het model dat we getraind hebben testen we op de test set.

We gebruiken hier een copy van de titanic set omdat we de data iets wijzigen.

```
perc = 0.80

## maak een categorische kolom van survived
myTitan = titanic::titanic_train
myTitan %>% mutate(
  Survived = ifelse(Survived < 1, "N", "Y") %>% as.factor
)

table(myTitan$Survived)

N = dim(myTitan)[1]

train = sample(1:N, size = floor(perc*N))

TTrain = myTitan[train,]
TTest = myTitan[-train,]
```

1.3 logistic regression

Een logistic regression is een van de simpelste predictive modellen om mee te beginnen als je classificatie wilt doen. We gaan uit van een binaire Target (Y /N). We gebruiken de `TTrain` data set die we zojuist gemaakt hebben om een model te fitten.

```
out.glm = glm(Survived ~ Sex + Age + Pclass,  data = TTrain , family = binomial)
summary(out.glm)
```

1.4 decision tree

Een decision tree genereert op basis van een algoritme regels die je kan gebruiken om te classificeren. Het is een eenvoudig algoritme dat per variabele kijkt hoe deze te gebruiken om de data set in twee stukken te splitsen (kan ook meer, maar gebruikelijk is twee).

```
tree.out = rpart(Survived ~ Sex + Age +Pclass, data = TTrain)

plot(tree.out)
text(tree.out, use.n = TRUE)

fancyRpartPlot(tree.out)

### larger trees with complexity parameter
tree.out = rpart(Survived ~ Sex + Age +Pclass, data = titanic_train, control = list(cp=0.005))
fancyRpartPlot(tree.out)
```

1.5 random forest met ranger

Een random forest is een zogenaamde ensemble model. Het is de combinatie van (veel) verschillende decision trees.

```
ranger.out = ranger( Survived ~ Sex + Age + Pclass, data = TTrain)

### er zijn missende waarden. We zouden ze kunnen verwijdern
TTrain = TTrain %>% filter(!is.na(Age))
ranger.out = ranger( Survived ~ Sex + Age +Pclass, data = TTrain, importance = "impurity")

ranger.out
```

1.6 xgboost

Extreme gradient boosting wordt de laatste tijd ook veel gebruikt in Kaggle competities. Zoals bij random forests is een xgboost model ook een ensemble van decision trees, maar de trees zijn nu niet onafhankelijk van elkaar. Eerst wordt een tree gefit, daarna een andere op basis van de eerste, etc.

Met de library `xgboost` kan je in R extreme gradient boosting modellen fitten. De aanroep is anders dan wat we tot nu toe gezien hebben. De `xgboost` functie moet een matrix met input variabelen worden meegegeven.

```
Titan_Inputmatrix = sparse.model.matrix( Survived ~ Sex + Age + Pclass, data = TTrain)

Titan_Inputmatrix

xgboost.out = xgboost(Titan_Inputmatrix, label = TTrain$Survived, nrounds = 10)

## er zijn diverse opties mee te geven aan xgboost
param = list(
  objective = 'binary:logistic',
  eval_metric = 'auc'
)

xgboost.out = xgboost(params=param, Titan_Inputmatrix, label = TTrain$Survived, nrounds = 10)
```

1.7 predictie en validatie

Met een test set kan je bepalen hoe goed een model is. Gebruik het model object van een modelfit om een test set te scoren en de scores met de ware uitkomsten te vergelijken.

1.7.1 predicties

Voor binaire classificaties is het handig om response kansen uit te rekenen. Voor logistische regressie met `glm` gebeurt dit niet automatisch

```
pred_GLM = predict(out.glm, data = TTest, type='response')
hist(pred_GLM)
```

Voorspelling van de decision tree, en random forest ranger.

```
pred_tree = predict(tree.out, data = TTest)
hist(pred_tree)
```

```
TTest = TTest %>% filter(!is.na(Age))
pred_ranger = predict(ranger.out, data = TTest)
hist(pred_ranger$predictions)
```

En voor xgboost moet je ook de test set als matrix veranderen

```
Titan_Testmatrix = sparse.model.matrix( Survived ~ Sex + Age +Pclass, data = TTest)
pred_xgboost = predict(xgboost.out, newdata = Titan_Testmatrix)
hist(pred_xgboost)
```

1.7.2 Variable importance in trees

Als je een tree of ensemble van trees hebt getrained kan je een idee krijgen welke variabelen in het model belangrijk zijn geweest in het trainings proces. Laten we voor de tree modellen die we hierboven getraind hebben de variable importance zien.

```
# enkele decision tree
tree.out$variable.importance

# ranger random forest
ranger.out$variable.importance

# xgboost
xgb.importance( colnames(Titan_Inputmatrix), model =xgboost.out)
```

1.7.3 roc curves and hit rates

```
testpr = predict(tree.out, TTest)
TTest$predictie = testpr

rocResultTEST = roc(Survived  ~ predictie, data = TTest , auc=TRUE, ci =TRUE)
plot(rocResultTEST)

## HITRATES
TTest %>%
  ggplot(aes(predictie, Survived))  +
  geom_smooth() +
  geom_abline(slope = 1,intercept = 0) +
  ggtitle("survived rate rate on test set") + scale_y_continuous(limits=c(0,1))
```

2 The h2o package

H2O is een schaalbaar machine learning platform die je vanuit R kan bedienen. Het bevat veel machine learning algoritmes, en voor grotere sets waar gewoon R moeite mee heeft kan h2o een uitkomst bieden. H2o heeft een eigen 'executie engine' geschreven in java. Bij het opstarten van h2o vanuit R wordt dan ook een apart h2o proces opgestart waar je data vanuit R naar toe moet uploaden om daar de algoritmes op los te laten.

Als je h2o opstart is er ook een eigen GUI, daar kan je naar toe localhost:54321 (standard 54321 port).

```
library(h2o)

# initialiseer h2o via R

#h2o.init(nthreads=-1, port=54323, startH2O = FALSE)
h2o.init()

### upload een R data set naar h2o: titanic train en test voorbeeldje

TTrain = TTrain %>% mutate_if(is.character, as.factor)
TTest = TTest %>% mutate_if(is.character, as.factor)
TTrain$Survived = as.factor(TTrain$Survived)
TTest$Survived = as.factor(TTest$Survived)

ttrain.h2o = as.h2o(TTrain)
ttest.h2o = as.h2o(TTest)
### Je kan ook direct text files inlezen in h2o met
#h2o.importFile(path="C:\een file.txt", sep=",")

### welke files zijn er in h2o
h2o.ls()
```

Er zijn diverse modellen die je kan trainen, we zullen hier een aantal laten zien, neural netwerks, boosting en random forests.

```
## model op titanic
NNmodel = h2o.deeplearning(
  x = c(3,5:6),
  y = "Survived",
  training_frame  = ttrain.h2o,
  validation_frame = ttest.h2o,
  hidden = 5,
  epochs = 250,
  variable_importances = TRUE
)

show(NNmodel)
h2o.varimp(NNmodel)
```

```
GBMmodel = h2o.gbm(
  x = c(3,5:6),
  y = "Survived",
  training_frame  = ttrain.h2o,
  validation_frame = ttest.h2o
  )
GBMmodel

RFmodel = h2o.randomForest(
  x = c(3,5:6),
  y = "Survived",
  training_frame  = ttrain.h2o,
  validation_frame = ttest.h2o
  )
RFmodel
```

```
h2o.varimp_plot(RFmodel)
```

Grid search in h2o. Je kan makkelijk modellen fine-tunen, in een grid kan je verschillende waarden van hyperparameters proberen.

```
RFmodelGrid = h2o.grid(
  "randomForest",
  x = c(3,5:6),
  y = "Survived",
  training_frame  = ttrain.h2o,
  validation_frame = ttest.h2o,
  hyper_params = list(
    ntrees =c(50,100),
    mtries = c(1,2,3)
  )
)

#overzicht van het grid, gesorteerd op logloss
RFmodelGrid
```

Geef resources terug door h2o af te sluiten als je het niet meer nodig hebt.

```
h2o.shutdown(prompt = FALSE)
```

3 The mlr package

Met het `mlr` package kan je makkelijk verschillende modellen trainen en testen op een meer uniforme manier. In R hebben alle machine learning technieken net weer verschillende aanroepen en de uitkomst is vaak een object met net steeds weer andere componenten. Dit zagen we bijvoorbeeld in de bovenstaande code voor ranger en xgboost.

Met het `mlr` package kan je dit uniform stroomlijnen. Het maken van een predictive model (welk model dan ook) bestaat altijd uit een aantal stappen. Bijvoorbeeld:

- specificeren van de target,
- specificeren van inputs,
- specificeren van variabelen die je niet wilt gebruiken,
- splitsen data,
- het model / algoritme

Deze zijn te beschrijven en uit te voeren in mlr.

We gebruiken de `titanic` data set als test in `mlr` en doorlopen een aantal stappen on een aantal modellen te benchmarken. De modellen die we willen benchmarken zijn:

- neuraal netwerk,
- gradient boosting,
- random forest
- xgboost
- decision tree,
- logistic regression via glmnet

3.1 specificeren van technieken en hun opties

In mlr kan je een aantal algemene parameters weergeven.

```
## parameters die geen beschrijving hebben willen we ook kunnen opgeven
configureMlr(on.par.without.desc = "warn")

## we kijken naar maximaal dertig variabelen
n.importance = 30

## voorspel type, we willen kansen uitrekenen
ptype = "prob"

## aantal crossvalidation splitsingen
N_CV_iter = 10
```

Naast algemene parameters, heeft elk model bepaalde parameters die je kan zetten. Dit hoeft niet, dan worden default waarden gekozen.

```
parameters_rf = list(
  num.trees  = 500
)

parameters_rpart = list(
  cp=0.0001
)

parameters_glmnet = list(
  alpha  = 1
)

parameters_NN = list(
  hidden = c(15,15)
)

parameters_xgboost = list(
  nrounds  = 5,
  max.depth = 7
)
```

Maak nu een Lijst van modelen (ook wel learners genomed) die je wilt trainen op je data.

```
RF_Learner = makeLearner(
  "classif.ranger",
  predict.type = ptype,
  par.vals = parameters_rf
)

xgboost_Learner = makeLearner(
  "classif.xgboost",
  predict.type = ptype,
  par.vals = parameters_xgboost
)

rpart_Learner = makeLearner(
  "classif.rpart",
```

```
  predict.type = ptype,
  par.vals = parameters_rpart
)

binomial_Learner = makeLearner(
  "classif.binomial",
  predict.type = ptype
)

glmnet_Learner = makeLearner(
  "classif.cvglmnet",
  predict.type = ptype,
  par.vals = parameters_glmnet
)

h2ogbm_Learner = makeLearner(
  "classif.h2o.gbm",
  predict.type = "prob"
)

h2oNN_Learner = makeLearner(
  "classif.h2o.deeplearning",
  predict.type = ptype,
  par.vals = parameters_NN
)

## lijst van de learners
learners = list(
  rpart_Learner,
  RF_Learner,
  binomial_Learner,
  glmnet_Learner,
  h2ogbm_Learner,
  h2oNN_Learner
)
```

Als je categorische variabelen in je voorspel model wilt gebruiken eist het mlr package dat ze `factor` zijn. En in het geval van een classificatie probleem moet de target variabele ook een factor variabele zijn.

```
ABT = titanic_train
ABT$Target = as.factor(ifelse(ABT$Survived < 1, "N", "Y"))
ABT = ABT %>% mutate_if(is.character, as.factor)
```

3.2 Imputeren van missende waarden

Als er missende waarden zijn kan je mlr deze laten imputeren door een bepaalde waarde.

```
impObject = impute(
  ABT,
  classes = list(
    integer = imputeMean(),
    numeric = imputeMean(),
    factor = imputeMode()
    ),
```

```
  dummy.classes = "integer"
)

ABT = impObject$data
```

3.3 Het aanmaken van een task

Maak nu een 'task' aan waarin je de data, de inputs en de target specificeert. Een classificatie taak voor categorische target of een regressie task voor een numerieke target.

Wat wil je modelleren: Kans op level "Y", dan dien je positive = "Y" op te geven. Bij een binair target met Y en N levels wordt namelijk standaard "N"gebruik (alfabetisch)

```
classify.task = makeClassifTask(id = "Titanic", data = ABT, target = "Target", positive = "Y")

## Overzicht van de taak en kolom informatie

print(classify.task)
getTaskDescription(classify.task)
summarizeColumns(classify.task)
```

3.4 Variablen hard uitsluiten

Soms zijn er variabelen die je niet wilt meenemen in je model. Deze kan je hard uitsluiten.

```
vars.to.drop = c("Name", "Survived", "Ticket")

classify.task = dropFeatures(classify.task, vars.to.drop )

## Weghalen van (bijna) constante variabelen

## Je kan ook 'bijna' constante variabelen weghalen: perc iets hoger zetten
classify.task = removeConstantFeatures(classify.task, perc = 0.01)
classify.task
```

Zeldzame levels van factors samenvoegen. Het is gebruikelijk om zeldzame levels te verwijderen of te mergen

```
classify.task = mergeSmallFactorLevels (classify.task, min.perc = 0.02,  new.level = ".merged")
summarizeColumns(classify.task)
```

Welke features hebben een effect op de target? Je kan de predictive power meten per input variable. Univariate, dus per feature kan je met mlr de relatie met de target berekenen.

Onderliggend heb je Rweka en Rjava dingen nodig daar kan je op linux wat issues mee krijgen. Dingen de je kan doen:

- run in een shell: `sudo R CMD javareconf` en doe
- sudo rstudio-server stop
- export LD_LIBRARY_PATH=/usr/lib/jvm/jre/lib/amd64:/usr/lib/jvm/jre/lib/amd64/default
- sudo rstudio-server start

Zie ook stack overflow

```
## Feature selection
fv = generateFilterValuesData(classify.task,  method = c("information.gain", "chi.squared"))
```

```
## display en plot importance

importance = fv$data %>% arrange(desc(information.gain))
head(importance, n = n.importance)
plotFilterValues(fv, n.show = 2*n.importance)
```

laat nog eens variabelen weg die helemaal niks doen.

```
vars.to.drop = c("PassengerId", "Parch", "SibSp")
classify.task = dropFeatures(classify.task, vars.to.drop )
```

3.5 Sample schema

Met mlr kan je data splitsen, niet alleen in train / test maar ook cross validation. Dit heet een sample schema.

```
SampleStrageyHO = makeResampleDesc("Holdout", split=0.75)
SampleStrageyCV = makeResampleDesc("CV", iters = N_CV_iter)
```

3.6 uitvoeren machine learning becnhamrk

Nu heb je de diverse stappen gespecificeerd en kan je een benchmark uitvoeren voor de verschillende learners,

```
br1 = mlr::benchmark(learners, classify.task, SampleStrageyHO, measures = list(mlr::mmce, mlr::auc, mlr
```

3.7 Vergelijking machine learning modellen

Na het trainen van de modellen met mlr heb je een zogenaamde benchmark object, die kan je printen en plotten om wat meer info te krijgen.

```
data.frame(br1) %>% arrange(desc(auc))
plotBMRSummary(br1, measure = mlr::auc)
```

3.7.1 ROC curves

In het benchmark object zit eigenlijk nog veel meer data. Met onderstaande code pluk je alle stukjes data per model uit om deze vervolgens in een ROC grafiek te zetten.

```
NModels = length(br1$results$Titanic)
for(i in 1:NModels)
{
  tmp2  = br1$results$Titanic[[i]]$pred$data
  rocResultTEST = roc(truth  ~ prob.Y, data = tmp2 )
  if(i==1)
  {
    plot(rocResultTEST, col=i)
  }else{
    plot(rocResultTEST, col=i, add=TRUE)
  }
}
```

```
legend( 0.6,0.6, names(br1$results$Titanic), col=1:NModels,lwd=2)
title("Titanic model")
```

3.7.2 model gebruiken om te scoren

Als je een benchmark hebt gedaan heb je al de getrainde modellen in het benchmark object zitten. Die kan je al gebruiken om een data set te scoren. je dient dit model er wel 'eerst uit te halen'.

```
## haal model er uit
titanmodel = br1$results$Titanic$classif.ranger$models[[1]]

## dit zijn de feauteures in het model
FT = titanmodel$features

## Maak even een score set van de ABT met alleen de features
ScoreSet = ABT[, FT]

outpredict = predict(titanmodel, newdata = ScoreSet)
outpredict
```

4 Unsupervised learning

De bovenstaande code was gericht op predictive modeling, ook wel supervised learning genoemd: met input variabelen een target variable proberen te voorspellen. In deze sectie zullen we een tweetal technieken laten zien waar geen target variabele is, ook wel unsupervised learning genoemd.

4.1 k-means Clustering

Dit is een van de bekendste clustering methode. Je dient een aantal clusters van te voren op te geven, de k, het algoritme gaat dan elke observatie aan een van de k clusters toekennen.

```
mycars = mtcars %>% select (mpg, wt)
cars.cluster = kmeans(mycars, 5)
cars.cluster

# in het ouput object zit informatie over het gefitte kmeans algoritme
mycars$cluster = cars.cluster$cluster
mycars

plot(mycars$mpg, mycars$wt, col = cars.cluster$cluster )
points(cars.cluster$centers, col = 1:5, pch = 8, cex = 2)
```

Met het `h2o` package kan je ook k-means clustering doen, dit is niet alleen sneller maar kan ook meteen factor variabelen aan, in de `kmeans` van R kan dat niet. Start indien nodig h2o.

```
library(h2o)
#h2o.init(nthreads=-1, port=54323, startH2O = FALSE)
h2o.init()
```

Breng data naar h2o, we gebruiken nu de sample data set mtcars in R maar we maken nog 1 extra factor kolom aan.

```
# am is de transimssie: 0 is automat en 1 is handgeschakeld, is eig
mycars = mtcars %>% mutate(am = as.factor(am))
cars.h2o = as.h2o(mycars)
```

Laat het algoritme zelf bepalen hoeveel clusters er in de data zijn.

```
cars_clustering = h2o.kmeans(cars.h2o,  k = 10, estimate_k = TRUE)
cars_clustering
```

na het trainen heb je een h2o cluster object met diverse informatie

```
cars_clustering@model
h2o.cluster_sizes(cars_clustering)
h2o.centers(cars_clustering)

## met h2o.predict kan je data scoren: bepalen tot welk cluster een observatie hoort en weer terug naar
cluster_membership = h2o.predict(
  cars_clustering,
  newdata = cars.h2o
  ) %>%
  as.data.frame()
```

5 Market basket analyse

Met market basket analyse (ook wel association rules mining genoemd) kan je uit "transacties van klanten" vaak voorkomende combinaties of juiste hele "sterke combinaties" van producten bepalen. Hieronder volgt een voorbeeldje op een fictief grocery data setje.

```
library(arules)
library(datasets)

## voorbeeld R object Groceries, dit is een tranaction object
data(Groceries)
summary(Groceries)
class(Groceries)

## De producten die een klant zou kunnen kopen
Groceries@itemInfo

### Klant 10 heeft gekocht:
Groceries@data[,10]
```

Met de functie `apriori` uit het arules package kan je nu vaak voorkomende combinaties van producten identificeren.

```
rules <- apriori(Groceries, parameter = list(supp = 0.001, conf = 0.8))
rules = sort(rules, decreasing = TRUE, na.last = NA, by = "lift")
inspect(head(rules, n = 15))
```

De fictieve voorbeelddata `Groeceries` in R is al in een zogenaamde transactions object, normaal gesproken zal je data niet meteen zo hebben in R. Vanuit transactionele data moet je dit converteren naar een transaction object. Dit kan je ook doen met het `arules` package.

```
## De meest simpele transactionele data set
trxDF = readRDS("data/boodschappen.RDs")
```

```
## Transormeer naar een transaction object
Groceries2 = as(
  split(
    trxDF$item,
    trxDF$id
    ),
  "transactions"
)
Groceries2

## Visuele Item informatie
itemFrequencyPlot(Groceries2, topN = 35, cex.names = 0.75)

rules2 <- apriori(Groceries2, parameter = list(supp = 0.001, conf = 0.8))
```

Nu je de regels hebt kan je filteren op regels. Welke regels bevatten bepaalde producten.

```
rules.subset2 <- subset(rules, lhs %in% c("cereals", "curd"))
rules.subset2
inspect(head(rules.subset2,n=15))
```

Of als iemand een bepaalde reeks transacties heeft welke regels horen daar bij en welk product kan je dan aanraden.

```
PersoonA = data.frame(
  id = rep(1,3),
  item2 = c("butter","curd","domestic eggs")
)

trxs_trans = as(
  split(
    PersoonA$item2,
    PersoonA$id
    ),
  "transactions"
)
inspect(trxs_trans)

rulesMatch <- is.subset(rules@lhs,trxs_trans)

## er zijn meerdere regels, je zou degene met de hoogste lift kunnen kiezen
inspect(rules[rulesMatch[,1]])
inspect(rules[rulesMatch[,1]]@rhs)
```

Een ander manier om regels weer te geven is in een network graph, de verzameling regels vormen in feite een netwerk. A –> B, B –> C, D –> B bijvoorbeeld.

```
library(arulesViz)
plot(head(sort(rules2, by = "lift"), n=50), method = "graph", control=list(cex=.8))
```