MA8701 Advanced methods in statistical inference and learning

L20 with Kjersti Aas

Mette Langaas IMF/NTNU

20 March, 2023

Contents

Part 4: Explainable AI

Outline

In L18 we motivated XAI, and then looked at

- Global explanation methods
  - Model specific methods
  - Model agnostic methods (PDP plots, ICE plots, ALE plots)

In L19, we covered LIME, of the local model agnostic methods.

- Local explanation methods
  - Method specific
  - Model agnostic (LIME, Shapley values, Counterfactual explanations)

In L19 we covered counterfactuals

Reading list

- Molnar (2023): Chapters 3, 6, 8 (not 8,3, 8.4,8.6,8.7, 9 (not 9.4, 9.6.3) from https://christophm.github.io/interpretable-ml-book/
- The three slidesets from Kjersti (on Blackboard)

Shapley values slide set

Bike data example

```
# download manually
#"https://github.com/christophM/interpretable-ml-book/blob/master/data/bike.Rdata"
load("bike.Rdata")
colnames(bike)
```

```
##  [1] "season"          "yr"              "mnth"            "holiday"
##  [5] "weekday"         "workingday"      "weathersit"      "temp"
##  [9] "hum"             "windspeed"       "cnt"             "days_since_2011"
```

```
n=dim(bike)[1]
bikeTrain=bike[1:600,]
bikeTest<-bike[601:n,]
```

Shapley regression with realtive weights

Show that relative weight give the same answer as the LMG-method.

```
rwa(bikeTrain[,-6],"cnt",c("temp","hum","windspeed","days_since_2011"))$result$Rescaled.RelWeight
```

```
## [1] 46.480548  3.520444  3.630216 46.368792
```

```
100*calc.relimp(cnt~.,data=bikeTrain[,8:12],type ="lmg", rela = TRUE )$lmg
```

```
##            temp             hum       windspeed days_since_2011
##       46.512680        3.578442        3.470233       46.438645
```

Kernel SHAP to calculate Shapley values

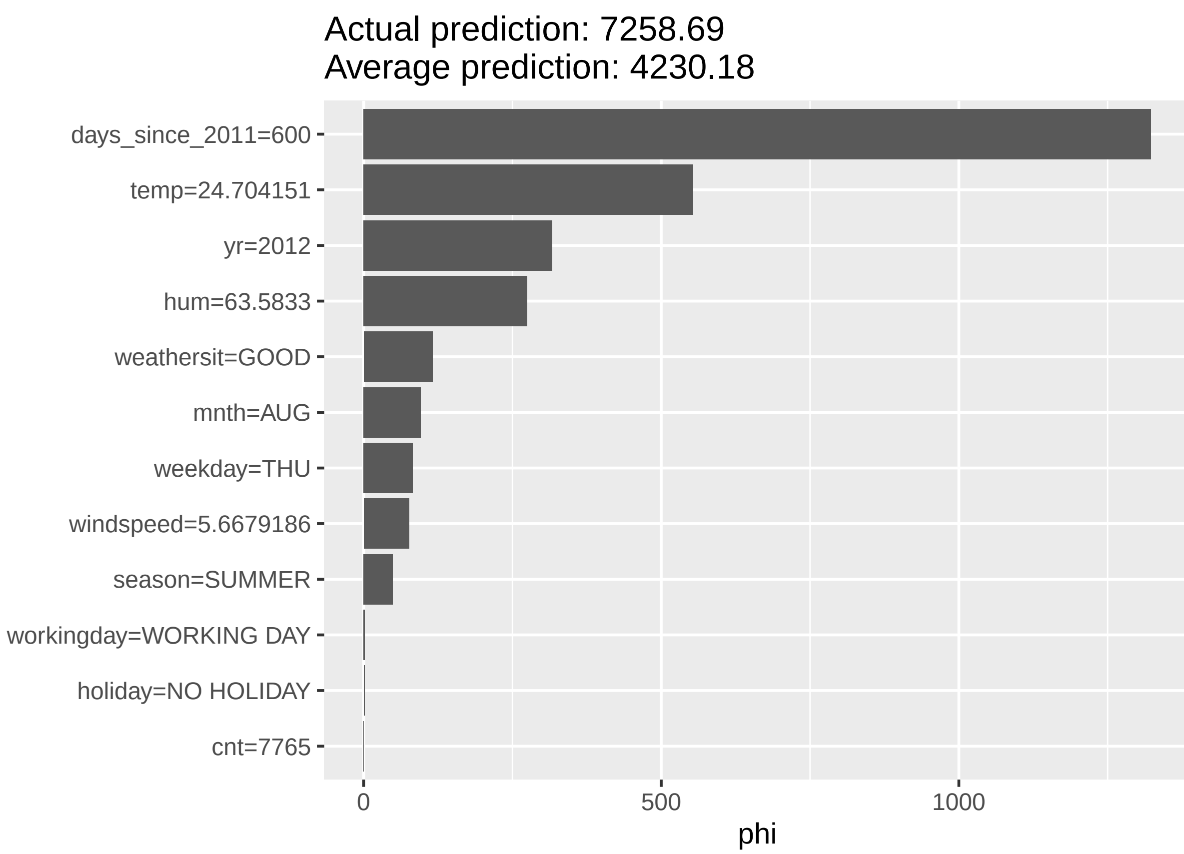
with taxi example

```
y=c(0,6,12,42,12,42,42,42)
m <- 3
xMat <- NULL
for (i in 1:m)
{ # compute all possible combinations of i features
coalitions <- combn(m, i)
tmpMat <- matrix(0,ncol=m,nrow=ncol(coalitions))
for(j in 1:ncol(coalitions)) tmpMat[j,coalitions[,j]] <- 1
xMat <- rbind(xMat,tmpMat)
}
#Add row for intercept
xMat <- rbind(rep(0,m),xMat)
d <- dim(xMat)[1]
w <- array(0,d)
for(i in 1:d)
{
  s <- length(which(xMat[i,]==1))
 # w[i] = (m-1)/(mCooseS(m,s)*s*(m-s))
  w[i] = (m-1)/(choose(m,s)*s*(m-s))

}
w[1] <- 10^6
```

```r
w[d] <- 10^6
lm(y ~ ., data=as.data.frame(cbind(y,xMat)),weights=w)
```

```
##
## Call:
## lm(formula = y ~ ., data = as.data.frame(cbind(y, xMat)), weights = w)
##
## Coefficients:
## (Intercept)           V2           V3           V4
##   5.333e-06    2.000e+00    5.000e+00    3.500e+01
```

```r
# probably other model used in slide set
model<- ranger(cnt ~ ., data = bikeTrain,
num.trees = 50, num.threads = 6,
verbose = TRUE,
probability = FALSE,
importance = "impurity",
mtry = sqrt(27))

pfun <- function(object, newdata)
predict(object, data = newdata)$predictions
mod <- Predictor$new(model = model, data = bikeTrain, predict.fun = pfun)
x.interest <- bikeTest[1, ]
shapley <- Shapley$new(mod, x.interest = x.interest)
plot(shapley)
```



ctree approach in shapr

```
explainer <- shapr(bikeTrain[,-11], model)
p <- mean(bikeTrain[,11])
explain <- shapr::explain(bikeTest[1,],
explainer,
approach = "ctree",
prediction_zero = p,
mincriterion = 0.95,
minsplit = 20,
minbucket = 7,
sample = TRUE)

print(explain$dt)

if (requireNamespace("ggplot2", quietly = TRUE))
{plot(explain)}
```

```
#Independence
pfun <- function(object, newdata)
predict(object, data = newdata)$predictions
mod <- Predictor$new(model = model, data = bikeTrain, predict.fun = pfun)
x.interest <- bikeTest[1, ]
shapley <- Shapley$new(mod, x.interest = x.interest)
plot(shapley)
```

References for further reading Shapley

- Relative weights: Johnson (2000)
- Shapley values with dependent features: Aas, Jullum, and Løland (2021)
- Kernel SHAP: Lundberg and Lee (2017)

Software

- R shapr: https://cran.r-project.org/web/packages/shapr/shapr.pdf
- Python kernel SHAP: https://github.com/slundberg/shap

References

Aas, Kjersti, Martin Jullum, and Anders Løland. 2021. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence*. https://doi.org/10.1016/j.artint.2021.103502.

Johnson, J W. 2000. "Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression." *Multivariate Behavioral Research*. doi:10.1207/S15327906MBR3501_1.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Molnar, Christoph. 2023. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.*