Aprendizado de Máquinas
Pré-processamento

Douglas Rodrigues

Universidade Federal Fluminense

Dependência Linear

- Ideia: remover variáveis que são combinações lineares de outras variáveis.

A	B	C	D	E	F
1	1	0	1	0	0
1	1	0	0	1	0
1	1	0	0	0	1
1	0	1	1	0	0
1	0	1	0	1	0
1	0	1	0	0	1

- Observe que

  C = A-B

  F = A-D-E

- Ou seja, as colunas C e F são combinações lineares das outras colunas, ou seja, podem ser removidas.

Dependência Linear

- Para detectar e remover as dependências lineares, utilizaremos o comando `findLinearCombos` do pacote `caret`.

```
> dl<-caret::findLinearCombos(testData2)
> testData<-testData2[ ,-dl$remove]
```

Função preProcess()

- O pré-processamento por ser realizado:
  1. Por funções próprias. Ex.: `nearZeroVar()`, `findLinearCombos()`, `findCorrelation()`, todos do pacote `caret`.
  2. Dentro do `caret::train()`.
  3. Utilizando a função `preProcess()`, também do pacote `caret`.

Função preProcess()

- Quando realizamos pré-processamento utilizando `train()`, ele utiliza o comando `preProcess()`, ou que vai facilitar bastante, pois os argumentos são os mesmos.

```
>library(caret)
> Wage<-ISLR::Wage
> set.seed(100)
# Separar em amostras treino/teste
> inTrain <- createDataPartition(y=Wage$wage,p=0.75,list=F)

> training <- Wage[inTrain,]
> testing <- Wage[-inTrain,]
```

Função preProcess()

```
#Criar pré-processamento
> tratamento<-preProcess(training, method = c("nzv","corr"),
                freqCut = 95/5, uniqueCut = 10, cutoff = .75)

#Agora precisamos APLICAR o pré-processamento na amostra treino.
> training_pp<-predict(tratamento,training)

#Aplicamos o mesmo pré-processamento na amostra teste.
> testing_pp<-predict(tratamento,testing)
```

Função `preProcess()`

- Posso criar um novo pré-processamento para as amostras TESTE?

  NÃO! Deve aplicar O MESMO pré-processamento aplicado na amostra TREINO, sem recalcular os parâmetros.

- Por que estamos aprendendo a fazer pré-processamento fora do `train()`?

  Você pode realizar treinamentos fora do `train()`. Há métodos que permitem maior edição dos hiperparâmetros quando realizados diretamente, fora do `train()`.