Aprendizado de Máquinas
Pré-Processamento

Douglas Rodrigues

Universidade Federal Fluminense

Padronização dos dados

- Ideia: evitar que o algoritmo fique enviesado para as variáveis com maior ordem de grandeza.
- Para isso, efetuamos uma simples transformação para obter variáveis com média 0 e desvio padrão 1. Ou seja, para cada observação X_i da variável X, realizamos a transformação

$$X_i^{(padr)} = \frac{X_i - \bar{X}}{sd(X)}$$

Padronização dos dados

```
> library(caret)
> library(kernlab)
> data(spam)

#Criar amostras treino/teste
> set.seed(100)
> inTrain <- createDataPartition(y=spam$type,p=0.75,list=F)
> training <- spam[inTrain,]
> testing <- spam[-inTrain,]
```

Padronização dos dados

- Podemos criar alterações nas variáveis de estudo, para tornar a analise mais eficiente. Vamos transforma-las em variáveis de média 0 e desvio padrão 1.

```
> padr <- preProcess(training, method=c("center","scale"))
```

Aplicamos o pré-processamento nas amostras TREINO e TESTE.

```
> training_pdr <- predict(padr,training)
> testing_pdr <- predict(padr,testing)
```

Padronização dos dados com `train()`

- Utilizando função `train()`
- Sem padronização:

```
> set.seed(100)
> ctrl<-trainControl(method="repeatedcv", number=10, rep=3)
> modelFit<-train(type∼ ., data=spam,method="knn"',
                  trControl=ctrl)
```

- Com padronização:

```
> set.seed(100)
> ctrl<-trainControl(method="repeatedcv", number=10, rep=3)
> modelFit2<-train(type∼ . , data=spam,method="knn",
                  trControl=ctrl, preProcess=c("center", "scale"))
```

Exercício: compare o desempenho dos classificadores com e sem a padronização dos dados.

Normalização dos dados - Box-Cox

- O método de Box-Cox é o método mais simples e o mais eficiente computacionalmente.
- A transformação de Box-Cox só pode ser utilizada com dados positivos.
- Essa transformação é dada pela seguinte forma, onde o parâmetro λ é estimado utilizando o método de máxima verossimilhança.:

$$X_i^{(box)}(\lambda) = \frac{X_i^{\lambda} - 1}{\lambda} \text{ , se } \lambda \neq 0.$$

Normalização dos dados - Box-Cox

- O método de Box-Cox é o método mais simples e o mais eficiente computacionalmente.
- A transformação de Box-Cox só pode ser utilizada com dados positivos.
- Essa transformação é dada pela seguinte forma, onde o parâmetro λ é estimado utilizando o método de máxima verossimilhança.:

$$X_i^{(box)}(\lambda) = \frac{X_i^{\lambda} - 1}{\lambda} \text{ , se } \lambda \neq 0.$$

- Para aplicar a normalização com o `preProcess()`, basta executar o comando

```
> norm_box <- preProcess(training, method = "BoxCox")
> training_box <- predict(norm_box,training)
```

Normalização dos dados - Yeo-Johnson

- A transformação de Yeo-Johnson é semelhante à transformação de Box-Cox, porém ela aceita preditores com dados nulos e/ou dados negativos.
- Também podemos aplicá-la aos dados através da função `preProcess()`.

```
> norm_YJ <- preProcess(training, method = "YeoJohnson")
> training_YJ <- predict(norm_YJ,training)
```

Normalização dos dados - Transformação Exponencial de Manly

- O método exponencial de Manly também consiste em estimar um λ tal que as variáveis transformadas se aproximem de uma distribuição normal.
- Essa transformação é dada pela seguinte forma:

$$X_i^{(exM)}(\lambda) = \frac{e^{X\lambda} - 1}{\lambda} \text{ , se } \lambda \neq 0$$

- Também podemos aplicá-la aos dados através da função `preProcess()`.

```
> norm_exM <- preProcess(training, method = "expoTrans")
> training_exM <- predict(norm_exM,training)
```