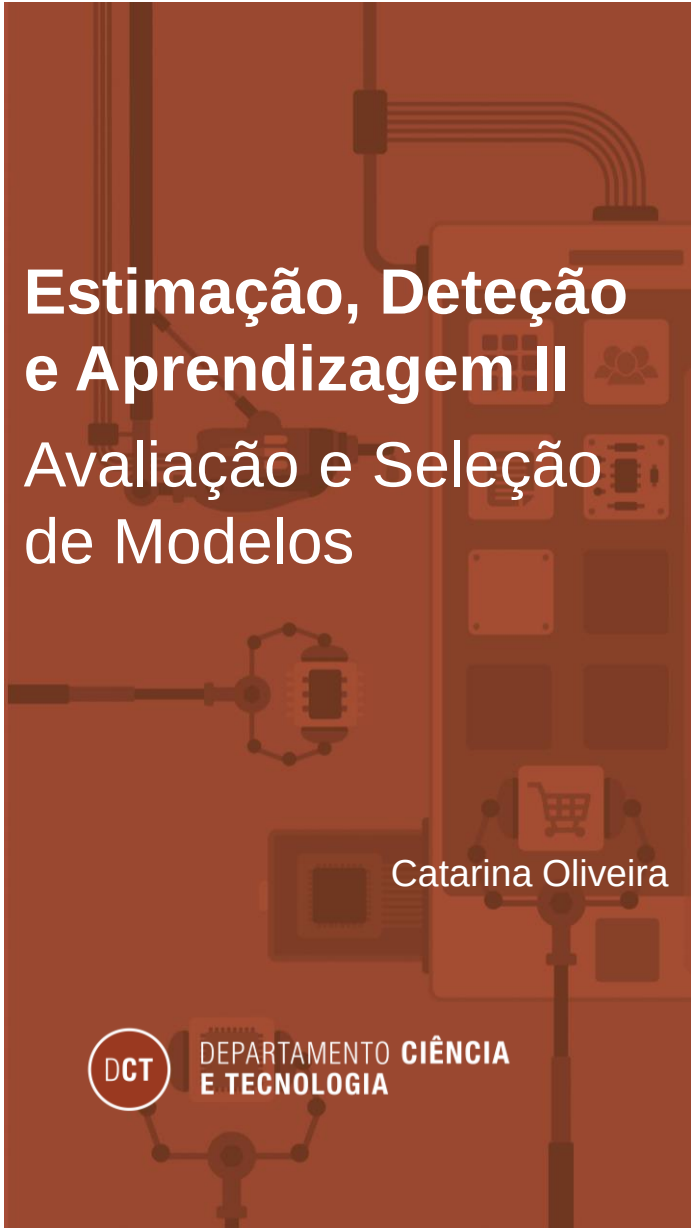


UPT UNIVERSIDADE PORTUCALENSE

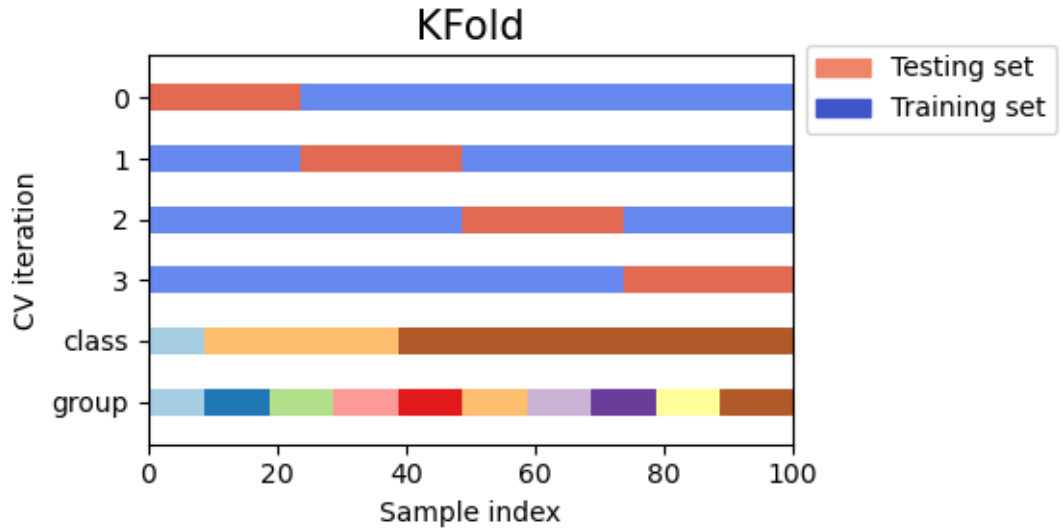
CONTEÚDO

1. Validação cruzada
2. *Bootstrap*
3. Curvas ROC
4. Seleção de características
5. Regularização

Validação cruzada
Cross validation

K Fold Cross-validation

Divide-se os dados em *k folds*. A cada iteração, usa-se *k-1 folds* para treino e 1 *fold* para teste



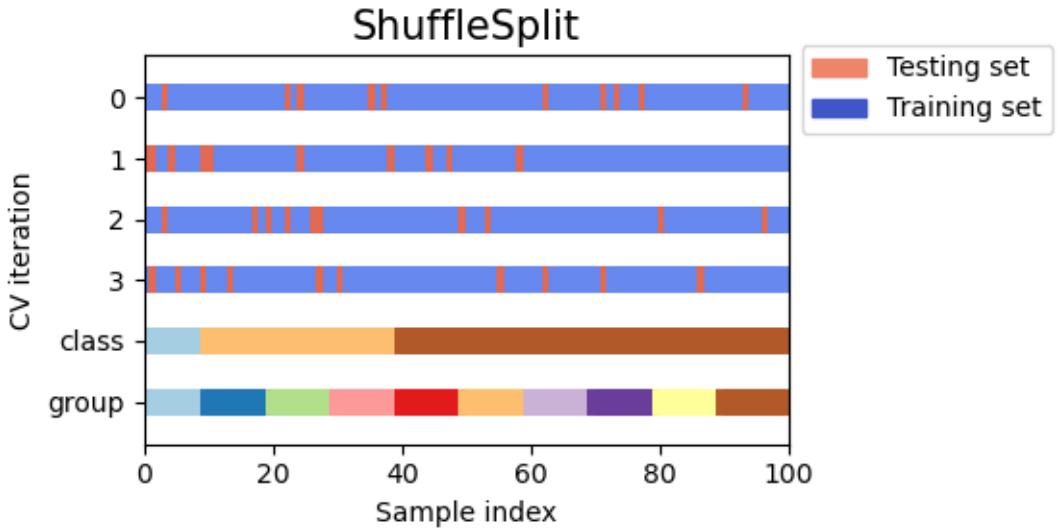
Casos especiais, considerando um *dataset* com *n* instâncias

***Leave One Out* (LOO)**: usa-se *n-1* instâncias como treino e 1 instância como teste

Leave P Out* (LPO)*: usa-se *n-p* instâncias como treino e *p* instâncias como teste

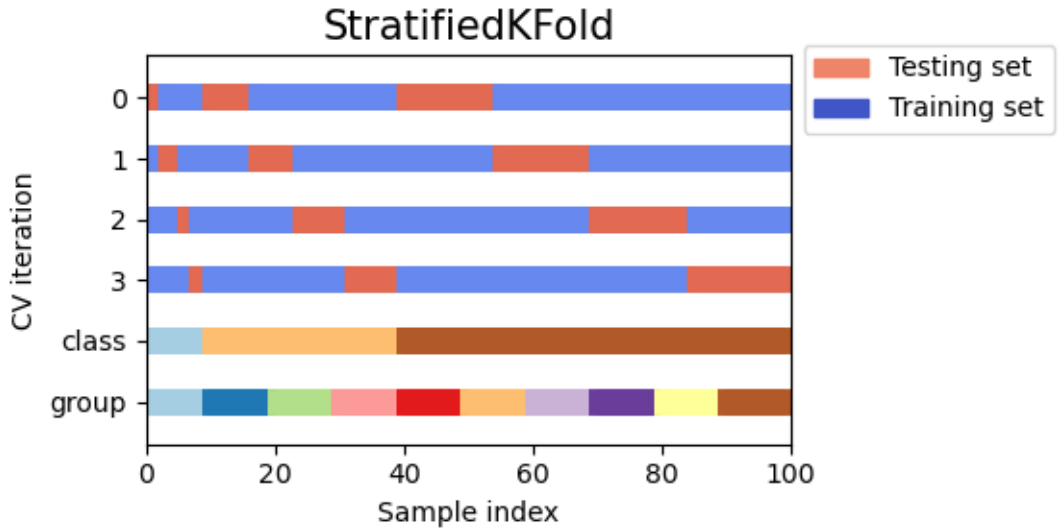
Shuffle & Split

As instâncias são “baralhadas” antes de se criar os conjuntos de treino e de teste



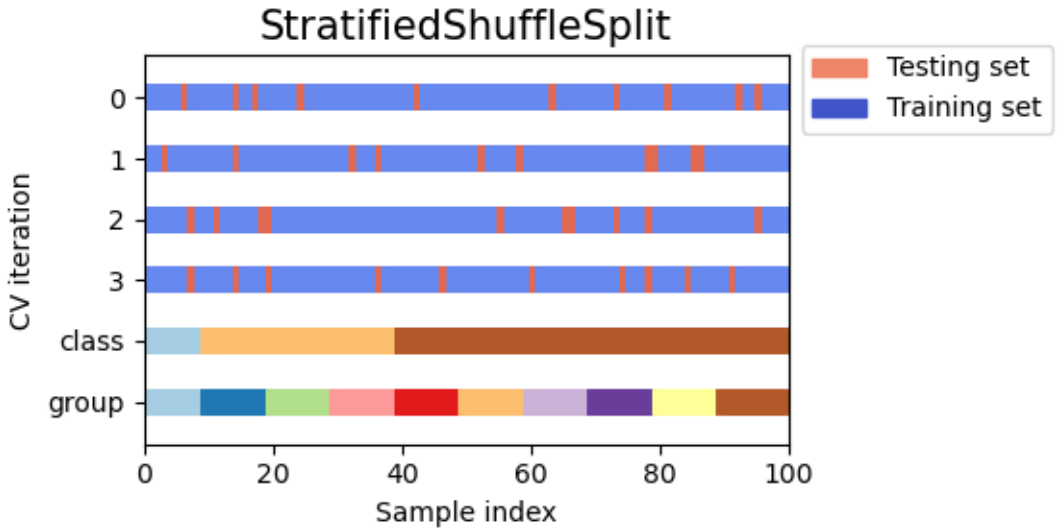
Stratified k-fold

Nos conjuntos de treino e de teste, tenta manter-se a distribuição das classes do *dataset* completo



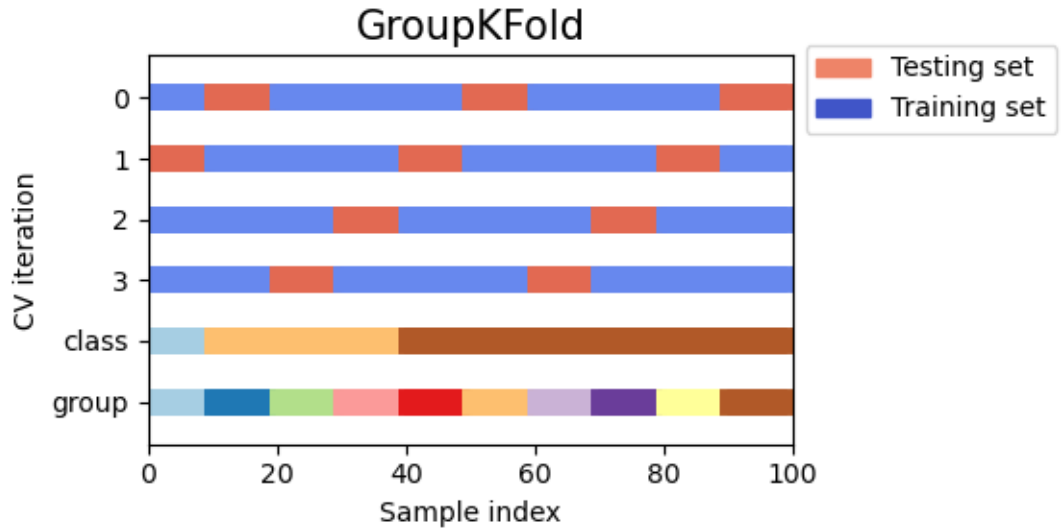
Stratified Shuffle Split

Nos conjuntos de treino e de teste, tenta manter-se a distribuição das classes do *dataset* completo e, para além disso, as instâncias são “baralhadas” antes de se criar os conjuntos de treino e de teste



Group k-fold

Quando várias instâncias podem ser agrupadas, permite assegurar que um determinado grupo não aparece ao mesmo tempo nos conjuntos de treino e de teste.



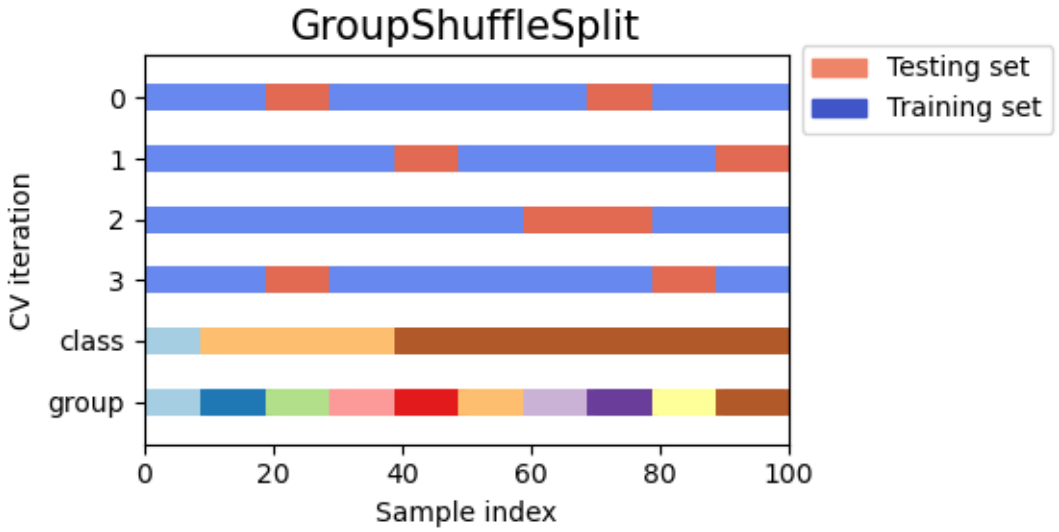
Casos especiais, considerando um *dataset* com *g* grupos de instâncias

Leave One Group Out: usa-se *g-1* grupos como treino e 1 grupo como teste

Leave P Groups Out: usa-se *g-p* grupos como treino e *g* grupos como teste

Group Shuffle Split

Combinação de *Shuffle & Split* e de *Leave P Groups Out*: gera uma sequência de partições aleatórias em que um subconjunto de grupos são reservados para teste (*held out*) em cada *split*



Bootstrap

Bootstrap

Técnica de reamostragem usada para estimar estatísticas numa população por amostragem de um conjunto de dados com reposição.

Pode ser usado para estimar estatísticas de resumo (ex: média ou desvio padrão).

Usado em *machine learning* para estimar a performance preditiva dos modelos sobre dados não incluídos no conjunto de treino.

A performance estimada pode ser apresentada com intervalos de confiança (não disponível com outros métodos, como validação cruzada).

Bootstrap: construção de uma amostra

1. Escolher o tamanho da amostra.
2. Enquanto o tamanho da amostra for menor do que o tamanho escolhido
    1. Selecionar aleatoriamente uma observação do conjunto de dados
    2. Adicione à amostra

Boostrap: estimação de estatísticas

1. Escolher o número de amostras
2. Escolher o tamanho da amostra
3. Para cada amostra
    1. Criar a amostra (método anterior)
    2. Calcule as estatísticas da amostra
4. Calcular a média das estatísticas calculadas das amostras

Bootstrap: estimar a performance preditiva de modelos de *machine learning*

Treina-se o modelo na amostra

Avalia-se a performance (testa-se) nas instâncias não incluídas na amostra (*out-of-bag sample* - OOB)

1. Escolher o número de amostras
2. Escolher o tamanho da amostra
3. Para cada amostra
    1. Criar a amostra (método anterior)
    2. Treinar o modelo na amostra
    3. Calcular a performance do modelo na amostra OOB
4. Calcular a média das performances obtidas com as amostras

Bootstrap: parâmetros

Tamanho da amostra

É comum usar amostras do mesmo tamanho do *dataset*. Algumas instâncias vão aparecer mais do que uma vez na amostra, enquanto que outras não vão aparecer.

Em *datasets* muito grandes, podemos usar amostras mais pequenas (ex: 50% ou 80%)

Nº de repetições

Deve ser grande o suficiente para assegurar significância estatística

Mínimo: 20 ou 30 repetições (valores mais pequenos aumentam a variância)

Idealmente, dependendo dos recursos existentes, deverão ser centenas ou milhares

Curvas ROC

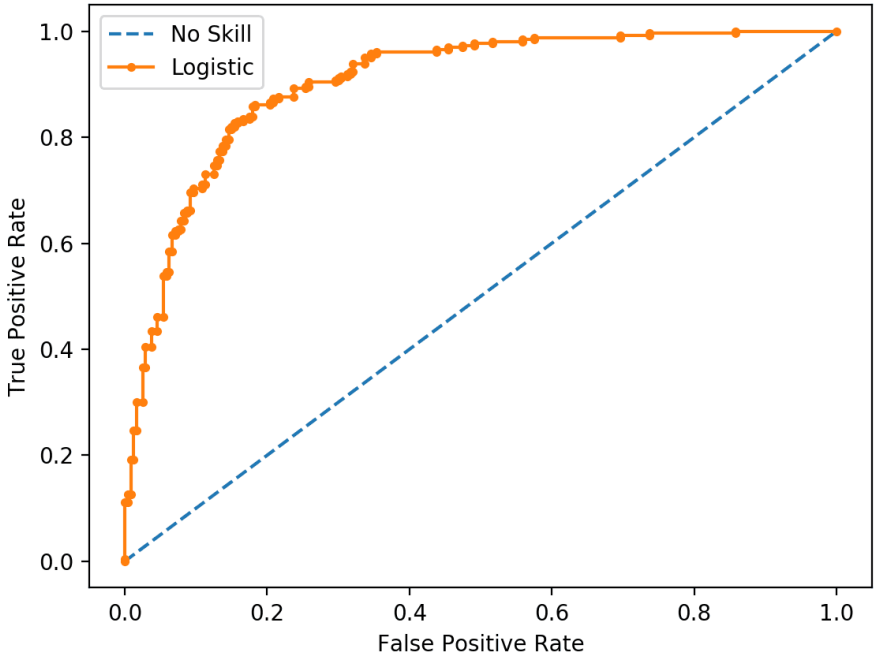
Curva ROC

ROC: *receiver operating characteristic*

Gráfico que ilustra a capacidade preditiva de um classificador binário com a variação do *discrimination threshold* (valor a partir do qual consideramos que a previsão é positiva)

Gráfico: true positive rate

$$TPR = \frac{TP}{TP + FN}$$



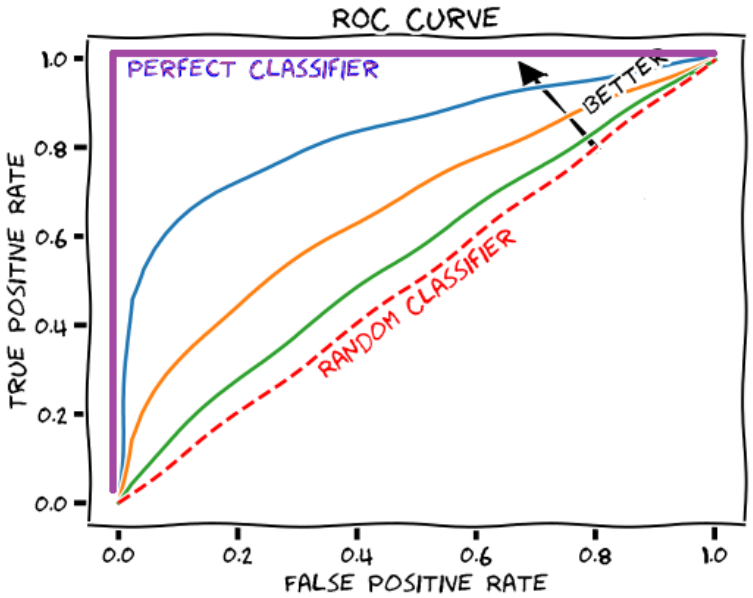
$$FPR = \frac{FP}{FP + TN}$$

AUC - ROC

AUC: *Area Under the Curve*

AUC – ROC: *Area Under the ROC Curve*

Tanto melhor quanto mais próximo de 1



Cálculo da curva ROC

#	Classe
1	P
2	P
3	P
4	N
5	N
6	P
7	P
8	P
9	N
10	N
11	N
12	N
13	P
14	N
15	P
16	N
17	N
18	P
19	P
20	N

Calcular previsões (probabilidade de ser P)

#	Classe	Prob(P)
1	P	0,9
2	P	0,51
3	P	0,34
4	N	0,33
5	N	0,36
6	P	0,54
7	P	0,55
8	P	0,4
9	N	0,39
10	N	0,7
11	N	0,35
12	N	0,37
13	P	0,8
14	N	0,505
15	P	0,6
16	N	0,1
17	N	0,53
18	P	0,38
19	P	0,3
20	N	0,52

Ordenar por probabilidade

#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Cálculo da curva ROC

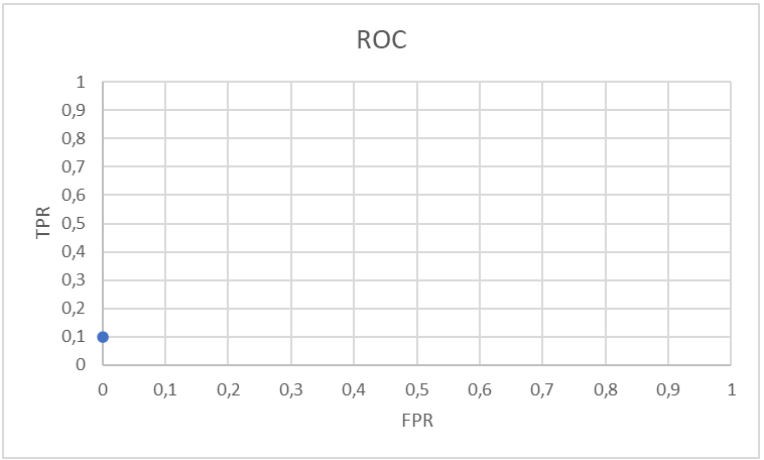
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,9

$$TPR = \frac{TP}{TP + FN} = \frac{1}{1 + 9} = 0{,}1$$

$$FPR = \frac{FP}{FP + TN} = \frac{0}{0 + 10} = 0$$

Thr	**0,9**
TP	1
TN	10
FP	0
FN	9
TPR	0,1
FPR	0



Cálculo da curva ROC

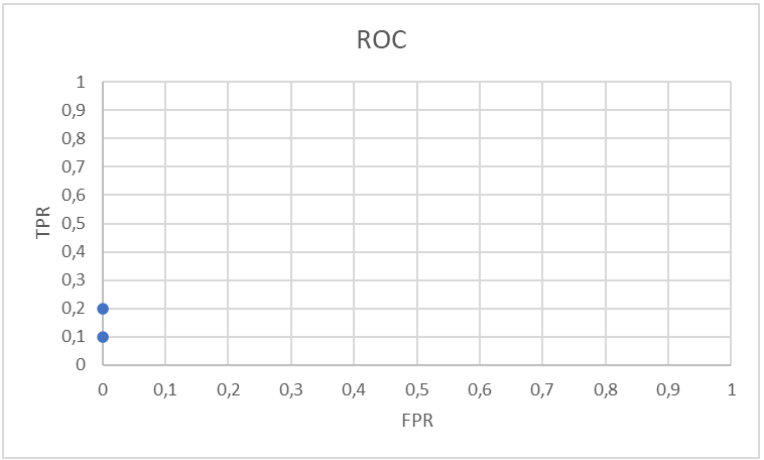
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,8

$$TPR = \frac{TP}{TP + FN} = \frac{2}{2 + 8} = 0{,}2$$

$$FPR = \frac{FP}{FP + TN} = \frac{0}{0 + 10} = 0$$

Thr	**0,9**	**0,8**
TP	1	2
TN	10	10
FP	0	0
FN	9	8
TPR	0,1	0,2
FPR	0	0



Cálculo da curva ROC

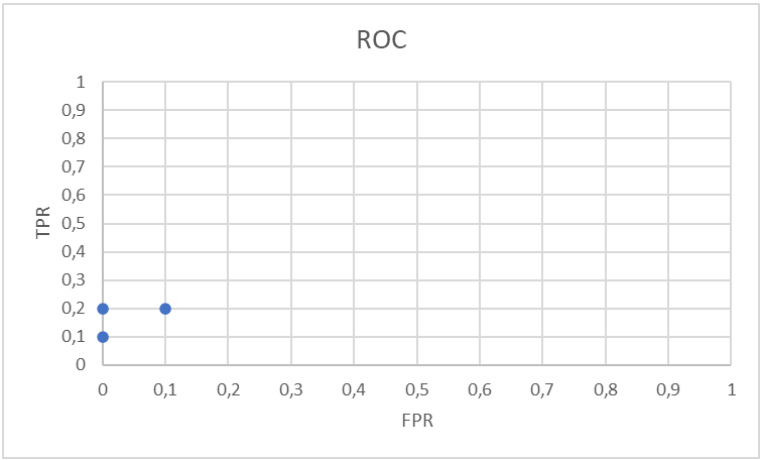
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,7

$$TPR = \frac{TP}{TP + FN} = \frac{2}{2 + 8} = 0{,}2$$

$$FPR = \frac{FP}{FP + TN} = \frac{1}{1 + 9} = 0{,}1$$

Thr	**0,9**	**0,8**	**0,7**
TP	1	2	2
TN	10	10	9
FP	0	0	1
FN	9	8	8
TPR	0,1	0,2	0,2
FPR	0	0	0,1



Cálculo da curva ROC

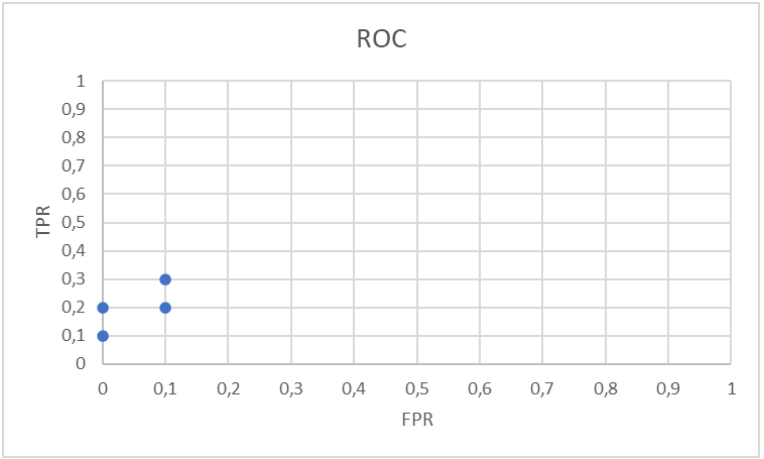
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,6

$$TPR = \frac{TP}{TP + FN} = \frac{3}{3 + 7} = 0{,}3$$

$$FPR = \frac{FP}{FP + TN} = \frac{1}{1 + 9} = 0{,}1$$

Thr	**0,9**	**0,8**	**0,7**	**0,6**
TP	1	2	2	3
TN	10	10	9	9
FP	0	0	1	1
FN	9	8	8	7
TPR	0,1	0,2	0,2	0,3
FPR	0	0	0,1	0,1



Cálculo da curva ROC

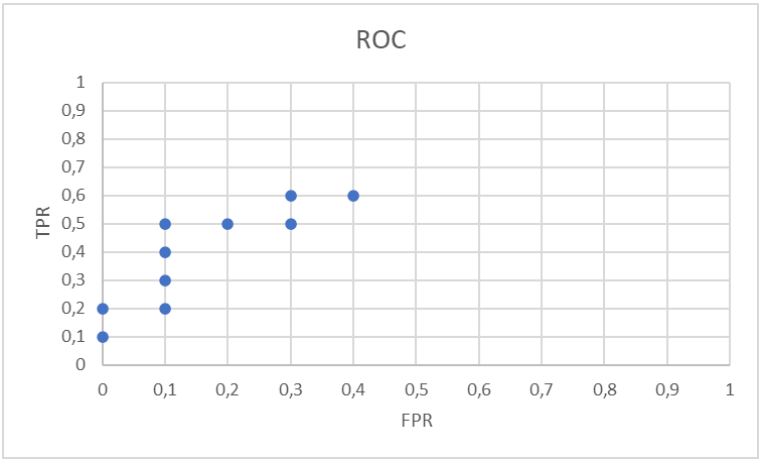
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,5

$$TPR = \frac{TP}{TP + FN} = \frac{6}{6 + 4} = 0{,}6$$

$$FPR = \frac{FP}{FP + TN} = \frac{4}{4 + 6} = 0{,}4$$

Thr	**0,9**	**0,8**	**0,7**	**0,6**	**0,5**
TP	1	2	2	3	6
TN	10	10	9	9	6
FP	0	0	1	1	4
FN	9	8	8	7	4
TPR	0,1	0,2	0,2	0,3	0,6
FPR	0	0	0,1	0,1	0,4



Cálculo da curva ROC

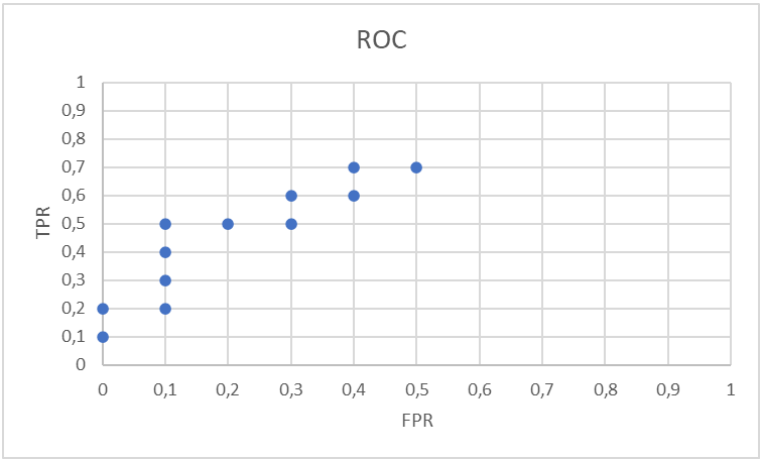
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,4

$$TPR = \frac{TP}{TP + FN} = \frac{7}{7 + 3} = 0{,}6$$

$$FPR = \frac{FP}{FP + TN} = \frac{4}{4 + 6} = 0{,}4$$

Thr	**0,9**	**0,8**	**0,7**	**0,6**	**0,5**	**0,4**
TP	1	2	2	3	6	7
TN	10	10	9	9	6	6
FP	0	0	1	1	4	4
FN	9	8	8	7	4	3
TPR	0,1	0,2	0,2	0,3	0,6	0,7
FPR	0	0	0,1	0,1	0,4	0,4



Cálculo da curva ROC

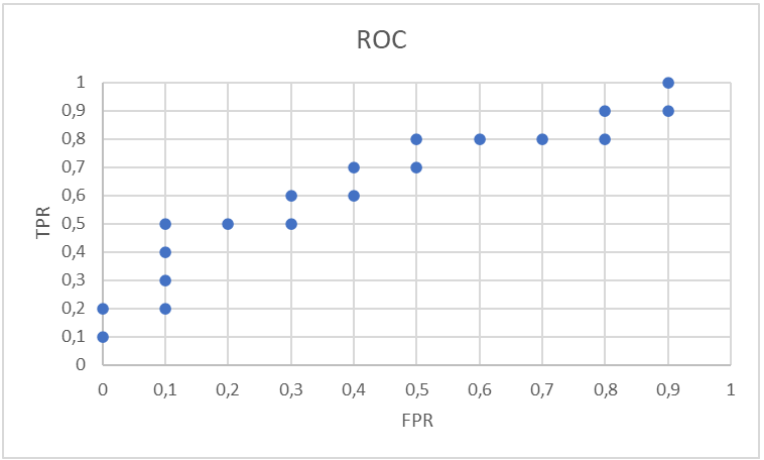
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,3

$$TPR = \frac{TP}{TP + FN} = \frac{10}{10 + 0} = 1$$

$$FPR = \frac{FP}{FP + TN} = \frac{9}{9 + 1} = 0{,}9$$

Thr	**0,9**	**0,8**	**0,7**	**0,6**	**0,5**	**0,4**	**0,3**
TP	1	2	2	3	6	7	10
TN	10	10	9	9	6	6	1
FP	0	0	1	1	4	4	9
FN	9	8	8	7	4	3	0
TPR	0,1	0,2	0,2	0,3	0,6	0,7	1
FPR	0	0	0,1	0,1	0,4	0,4	0,9



Cálculo da curva ROC

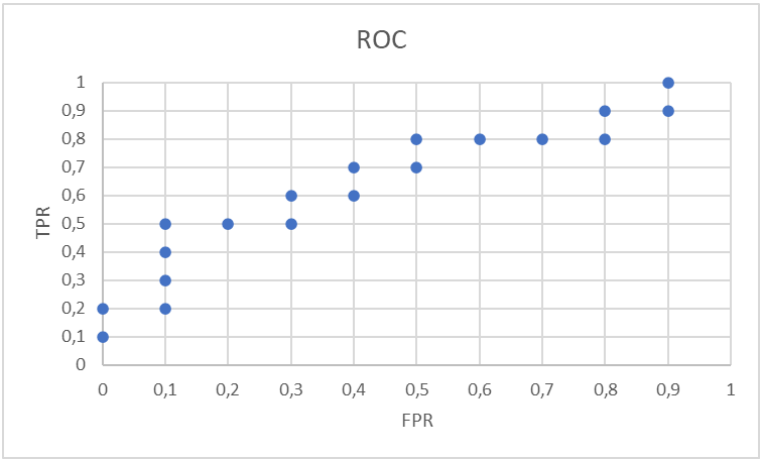
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,2

$$TPR = \frac{TP}{TP + FN} = \frac{10}{10 + 0} = 1$$

$$FPR = \frac{FP}{FP + TN} = \frac{9}{9 + 1} = 0{,}9$$

Thr	**0,9**	**0,8**	**0,7**	**0,6**	**0,5**	**0,4**	**0,3**	**0,2**
TP	1	2	2	3	6	7	10	10
TN	10	10	9	9	6	6	1	1
FP	0	0	1	1	4	4	9	9
FN	9	8	8	7	4	3	0	0
TPR	0,1	0,2	0,2	0,3	0,6	0,7	1	1
FPR	0	0	0,1	0,1	0,4	0,4	0,9	0,9



Cálculo da curva ROC

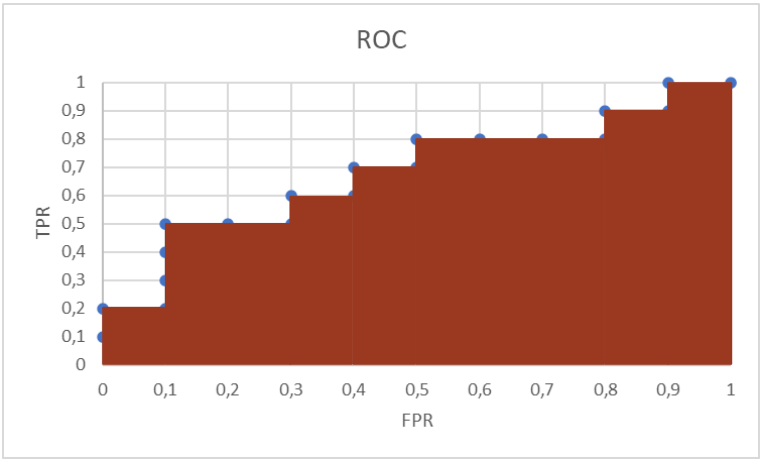
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

Threshold = 0,1

$$TPR = \frac{TP}{TP + FN} = \frac{10}{10 + 0} = 1$$

$$FPR = \frac{FP}{FP + TN} = \frac{9}{9 + 1} = 0{,}9$$

Thr	**0,9**	**0,8**	**0,7**	**0,6**	**0,5**	**0,4**	**0,3**	**0,2**	**0,1**
TP	1	2	2	3	6	7	10	10	10
TN	10	10	9	9	6	6	1	1	0
FP	0	0	1	1	4	4	9	9	10
FN	9	8	8	7	4	3	0	0	0
TPR	0,1	0,2	0,2	0,3	0,6	0,7	1	1	1
FPR	0	0	0,1	0,1	0,4	0,4	0,9	0,9	1



$$\begin{aligned} AUC = {} & (0{,}1 - 0) \times (0{,}2 - 0) \\ & + (0{,}3 - 0{,}1) \times (0{,}5 - 0) \\ & + (0{,}4 - 0{,}3) \times (0{,}6 - 0) \\ & + (0{,}5 - 0{,}4) \times (0{,}7 - 0) \\ & + (0{,}8 - 0{,}5) \times (0{,}8 - 0) \\ & + (0{,}9 - 0{,}8) \times (0{,}9 - 0) \\ & + (1 - 0{,}9) \times (1 - 0) \ = 0{,}68 \end{aligned}$$

Seleção de características
Feature selection

Feature selection

Objetivo: redução do número de *features* a serem consideradas pelo modelo

- Reduzir o custo computacional
- Aumentar a performance

Métodos que avaliam uma relação estatística entre as *features* e o target e escolhem as *features* com relação mais forte

Eliminação de *features*

- Redundantes
- Não informativas

Técnicas de *feature selection*

***Unsupervised*:**
Não utilizar a variável target
(ex: remover redundantes) → Correlação

***Feature selection*:**
Selecionar um subconjunto
das features do dataset

***Supervised*:**
Utilizar a variável target
(ex: remover irrelevantes)

***Wrapper*:**
Procurar subconjuntos de *features*
que levam a alta performance → RFE[1]

***Filter*:**
Selecionar subconjuntos de features
com base na sua relação com o target
→ Métodos estatísticos
→ *Feature importance*[2][3]

***Intrinsic*:**
Algoritmos que automaticamente
executam *feature selection*
→ Árvores de decisão,
Random Forests[4]

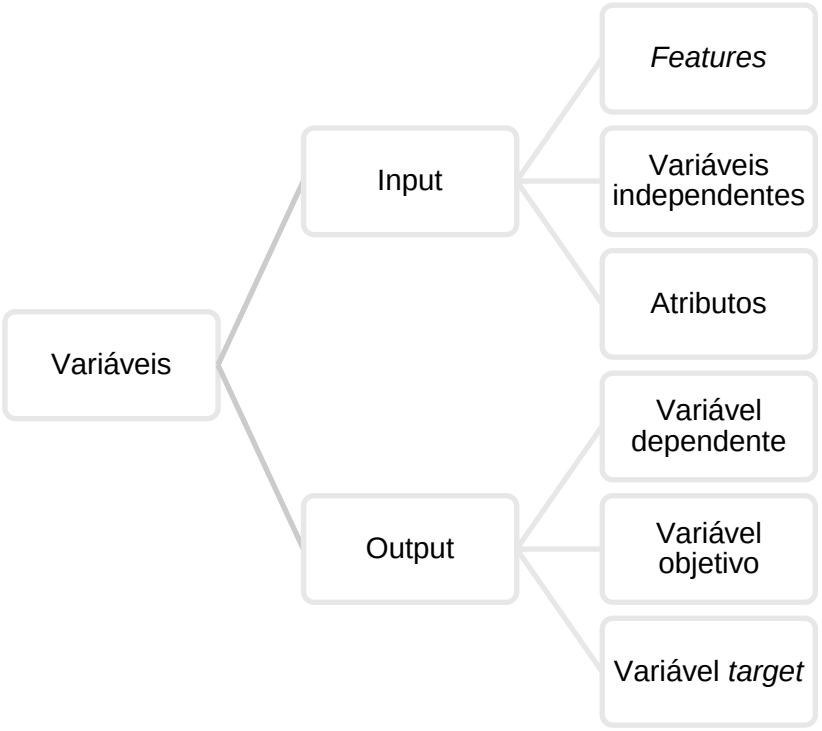
[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html#sklearn.feature_selection.RFECV

[2] Escolher as *k features* mais importantes: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

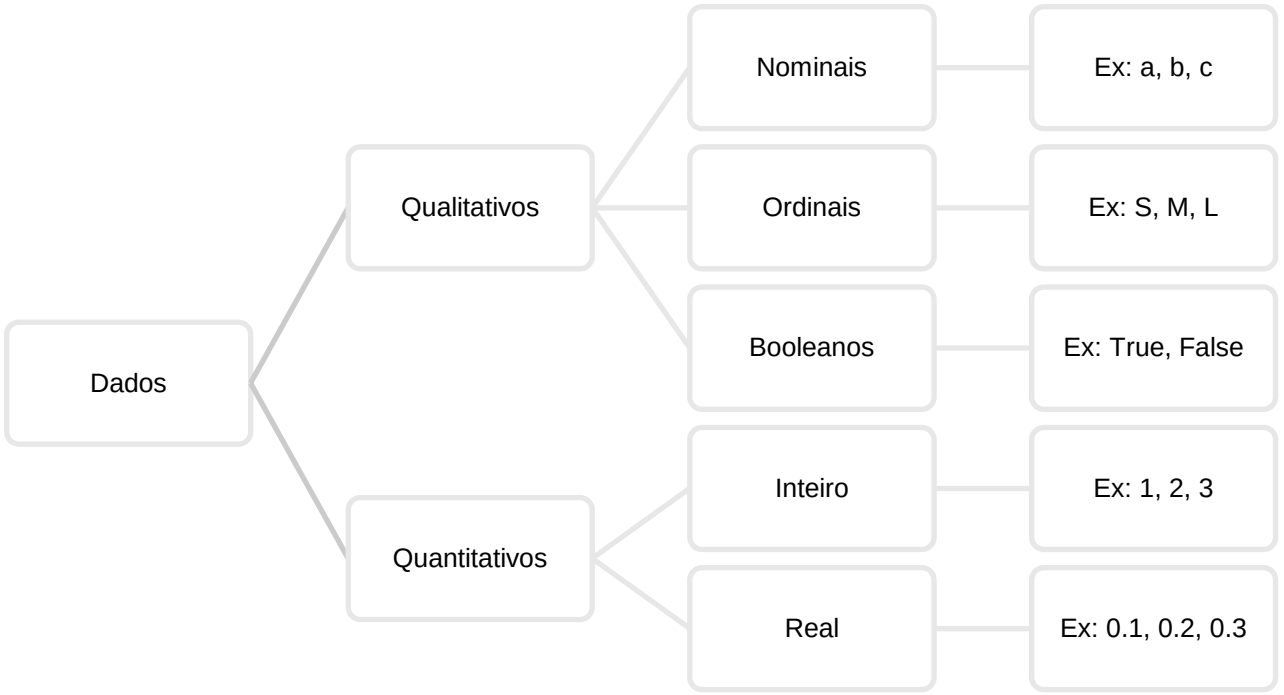
[3] Escolher as *p% features* mais importantes: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.feature_importances_

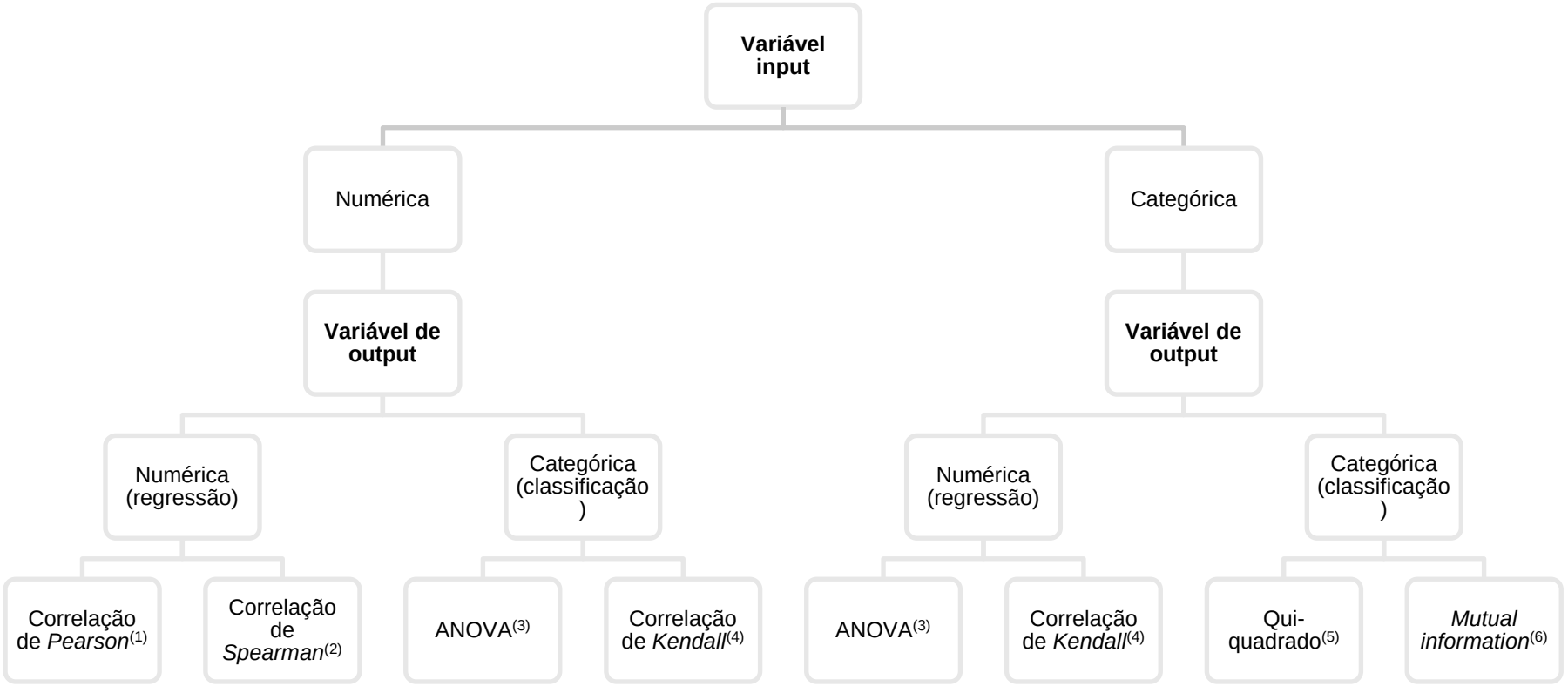
Revisão de tipos de variáveis



Revisão de tipos de variáveis



Como escolher o método de *feature selection*?



(1) https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html

(2) https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

(3) https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html

(4) https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html

(5) https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

(6) https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html, https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

Regularização

Regularização

Forma de regressão que regulariza (restringe, “encolhe”) os coeficientes determinados pelo modelo de regressão

Em regressão, considerando

- Um *dataset* com variáveis independentes X e dependente Y
- O processo de *fitting* escolhe os coeficientes β de forma a minimizar uma *loss funcion*
- A *loss function* é *Residual Sum of Squares* (RSS):

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Os coeficientes são ajustados tendo em conta a totalidade dos dados

Caso haja ruído nos dados, o modelo fica mais flexível, mas os coeficientes estimados não irão generalizar bem para dados novos

A regularização “encolhe” estes coeficientes (tendem para zero)

Ridge regression

A função a minimizar passa a ser

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}{\beta_j}^2=RSS+\lambda\sum_{j=1}^{p}{\beta_j}^2$$

Em que λ

- É o fator de regularização que determina quanto queremos penalizar a flexibilidade do modelo
- É um parâmetro a definir, por exemplo, com *cross validation*.
  - $\lambda=0$: a penalização não tem efeito e os coeficientes produzidos são os estimados
  - $\lambda\to\infty$: o impacto do fator de regularização aumenta e os coeficientes β irão aproximar-se de zero

Os coeficientes produzidos por este método são conhecidos por *L2 norm*

Os coeficientes deixam de obedecer à escala das *features*, pelo que é necessário standardizá-las, usando:

$$\tilde{x}_{ij}=\frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{ij})^2}}$$

LASSO regression

A função a minimizar passa a ser

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

Semelhante a *Ridge regression*

Como usa o módulo em vez do quadrado, não penaliza tanto coeficientes grandes

Os coeficientes produzidos por este método são conhecidos por *L1 norm*

Como penaliza coeficientes baixos na mesma medida dos coeficientes altos, acaba por executar *feature selection*

UNIVERSIDADE
PORTUCALENSE

Do conhecimento à prática.