Estimation, Detection and Learning II

Evaluation and Selection of Models

Catarina Oliveira

DEPARTAMENTO CIÊNCIA E TECNOLOGIA

UNIVERSIDADE PORTUCALENSE

IMP.GE.190.0

CONTENT

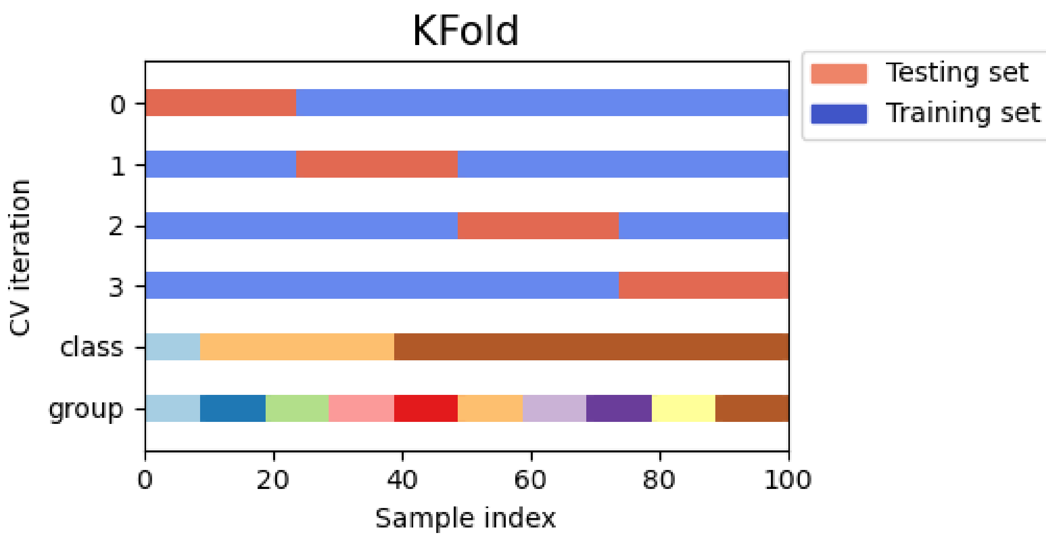
1. cross validation
2. *bootstrap*
3. ROC curves
4. feature selection
5. Regularization

cross validation

cross validation

K Fold Crossvalidation _

Divide the data into *k folds* . At each iteration, *k-1 folds* are used for training and 1 *fold* for testing



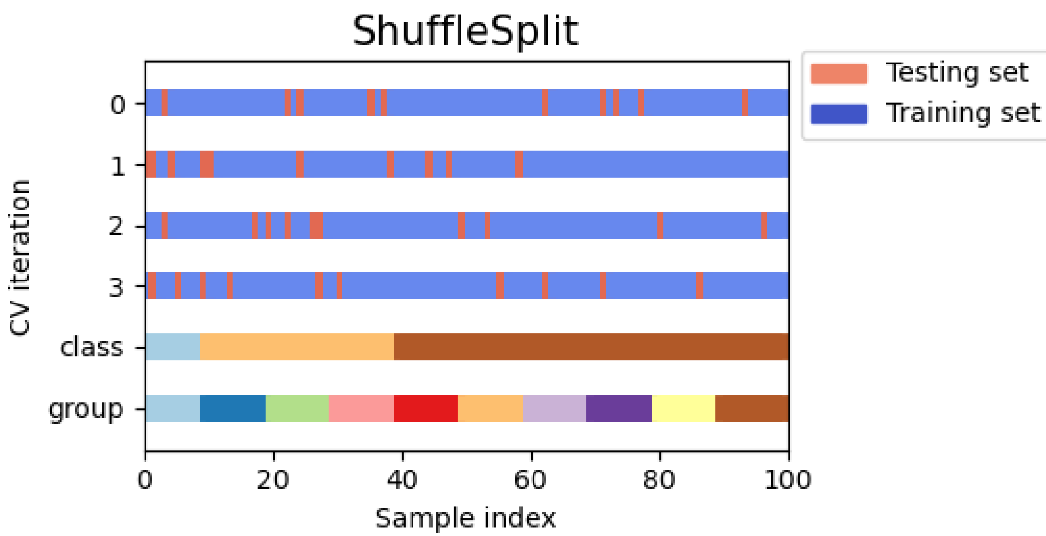
Special cases, considering a *dataset* with *n* instances

leave One Out **(LOO)** : use *n-1* instances as training and 1 instance as testing

Leave P Out **(LPO)** ***:*** use *np* instances as training and *p* instances as testing

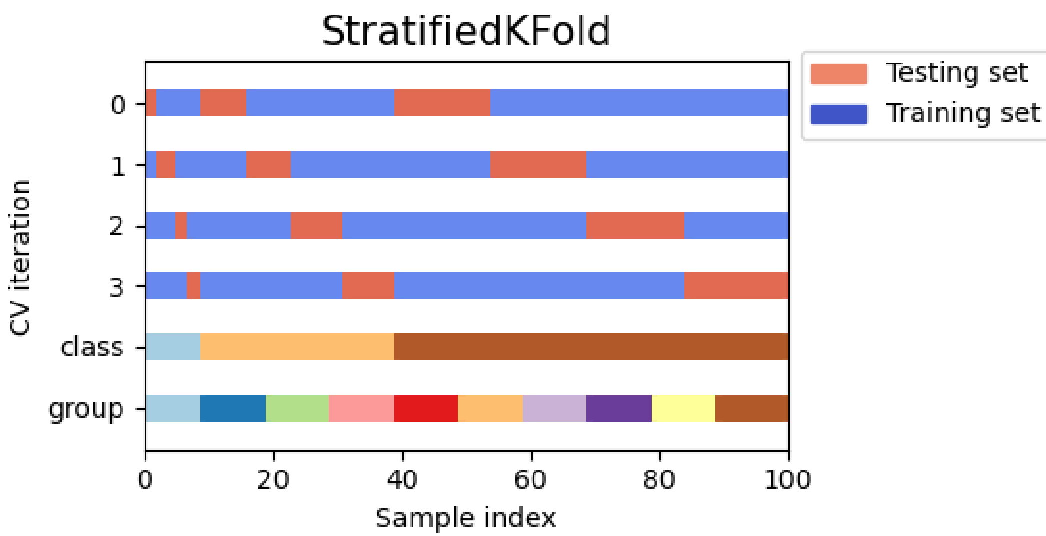
shuffle & split

The instances are “shuffled” before creating the training and test sets.



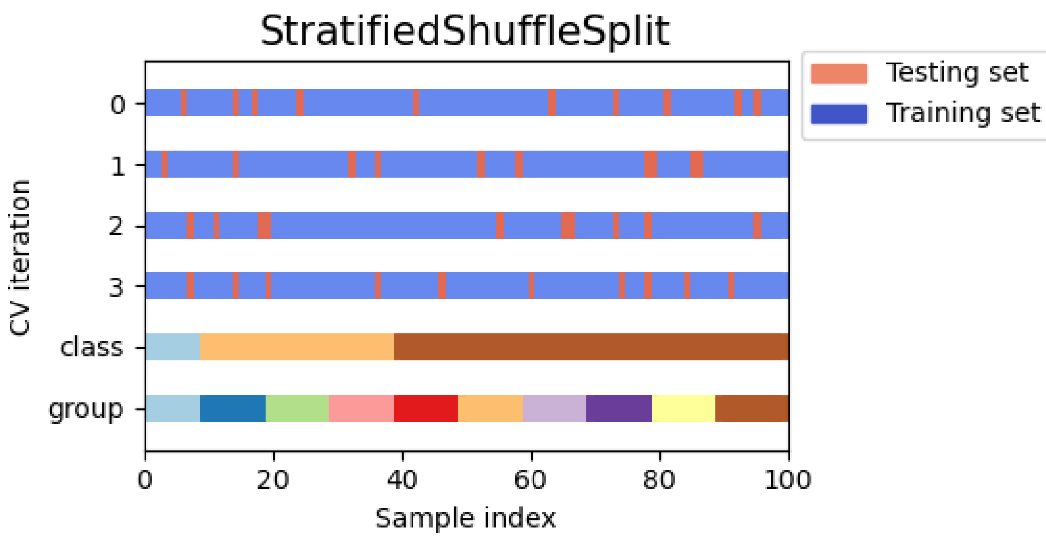
Stratified k- fold

In the training and test sets, try to keep the distribution of the classes of the complete *dataset* .



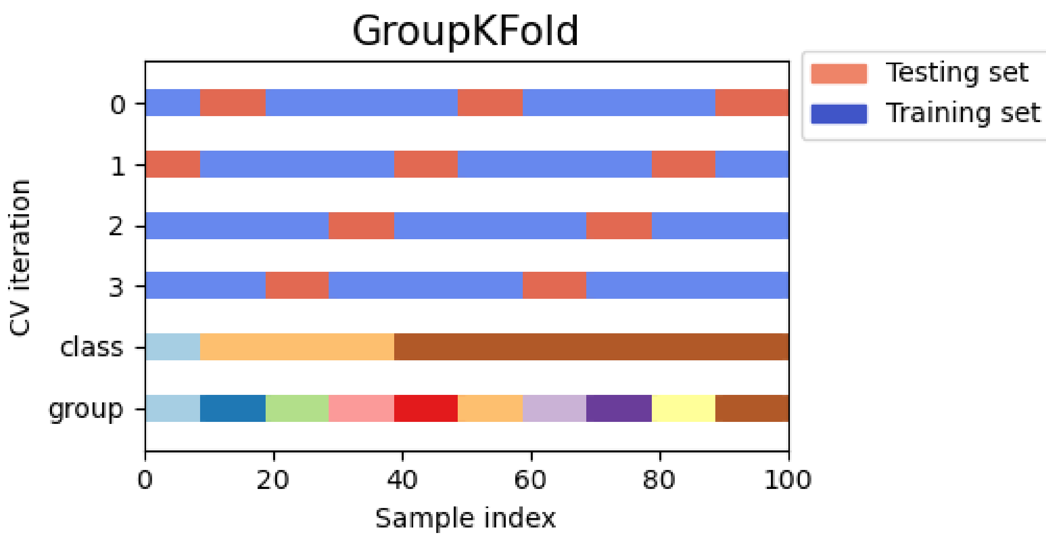
Stratified shuffle split

dataset is maintained and, in addition, the instances are “shuffled” before creating the training and test sets.



Group k- fold

When several instances can be grouped, it ensures that a certain group does not appear in the training and test sets at the same time.



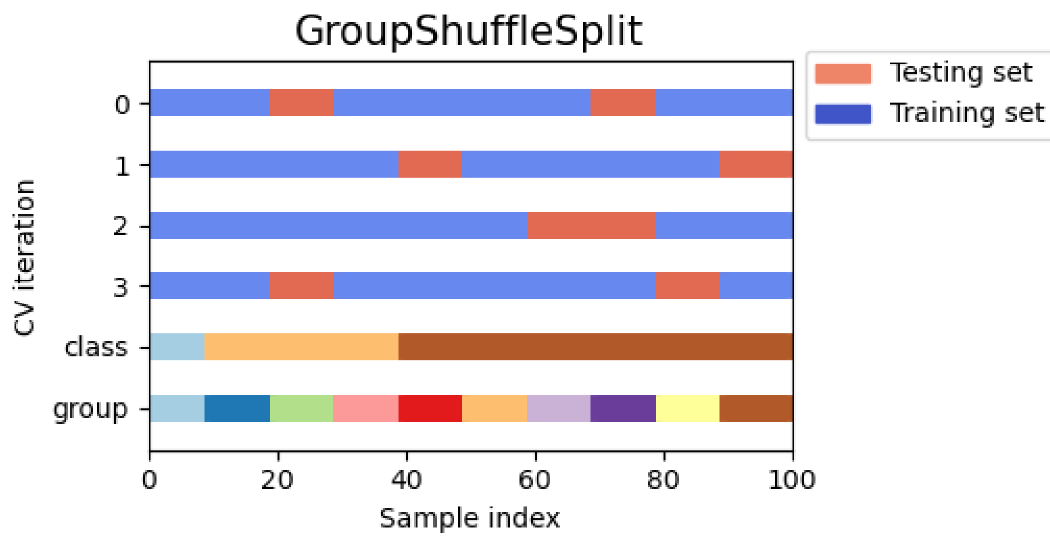
Special cases, considering a *dataset* with *g* groups of instances

leave One Group Out : use *g-1* groups as training and 1 group as test

Leave P Groups Out: use *gp* groups as training and *g* groups as testing

group shuffle split

Combination of *Shuffle & Split* and *Leave P Groups Out* : generates a sequence of random partitions in which a subset of groups are reserved for testing (*held out*) in each *split*



bootstrap

bootstrap

Resampling technique used to estimate statistics in a population by sampling a dataset with replacement.

Can be used to estimate summary statistics (eg mean or standard deviation).

used in *machine learning* to estimate the predictive performance of the models on data not included in the training set.

Estimated performance can be presented with confidence intervals (not available with other methods such as cross-validation).

Bootstrap : building a sample

1. Choose sample size.
2. As long as the sample size is smaller than the chosen size
    1. Randomly select an observation from the dataset
    2. add to sample

Boostrap : stat estimation

1. Choose the number of samples
2. Choose sample size
3. for each sample
    1. Create the sample (previous method)
    2. Calculate sample statistics
4. Calculate the average of the statistics calculated from the samples

Bootstrap : estimating the predictive performance of *machine models learning*

The model is trained on the sample

Performance is evaluated (tested) in instances not included in the sample (*out- of - bag sample* - OOB)

1. Choose the number of samples
2. Choose sample size
3. for each sample
    1. Create the sample (previous method)
    2. Train the model on the sample
    3. Calculate the performance of the model in the OOB sample
4. Calculate the average of the performances obtained with the samples

DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Bootstrap : parameters

Sample size

It is common to use samples of the same size as the *dataset* . Some instances will appear more than once in the sample, while others will not.

In very large *datasets* , we can use smaller samples (eg 50% or 80%)

number of repetitions

It must be large enough to ensure statistical significance.

Minimum: 20 or 30 repetitions (smaller values increase variance)

Ideally, depending on existing resources, there should be hundreds or thousands

ROC curves

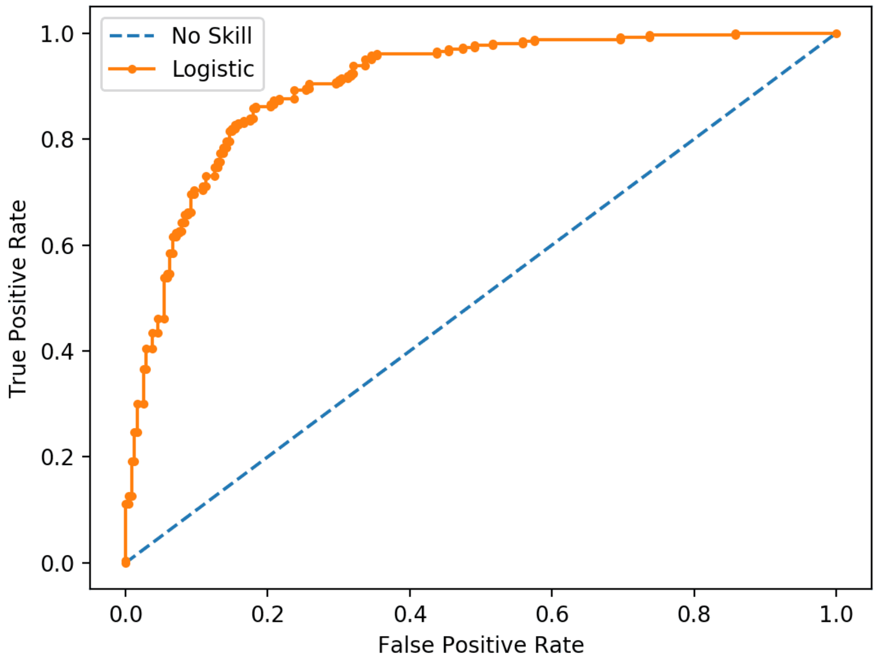
ROC curve

ROC: *receiver operating characteristic*

Graph illustrating the predictive ability of a binary classifier with variation in *discrimination threshold* (value from which we consider that the forecast is positive)

Graph : true positive rate

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



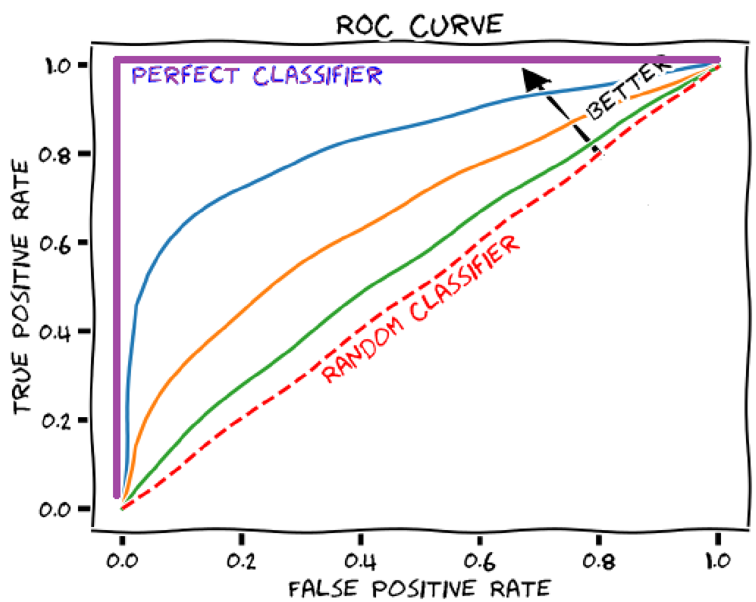
$$FPR = \frac{FP}{FP + TN}$$

AUC - ROC

AUC: *Area Under the Curve*

AUC - ROC: *Area Under the ROC Curve*

The better the closer to 1



Calculation of the ROC curve

#	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

calculate forecasts

(probability of being 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

Order by

probability

#	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

Calculation of the ROC curve

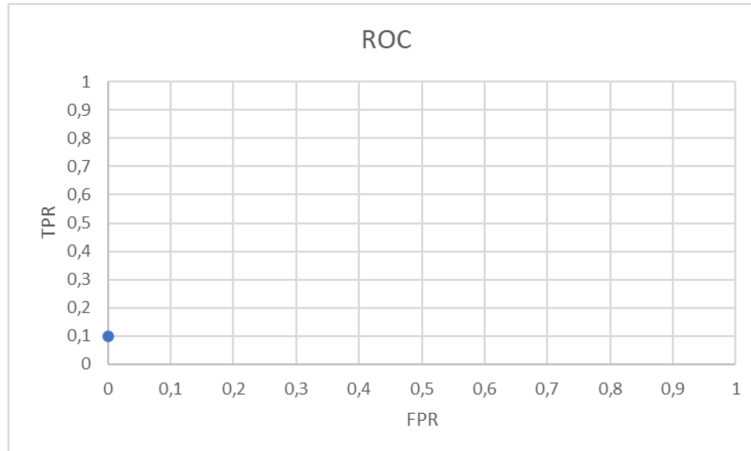
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



UPT DCT DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Calculation of the ROC curve

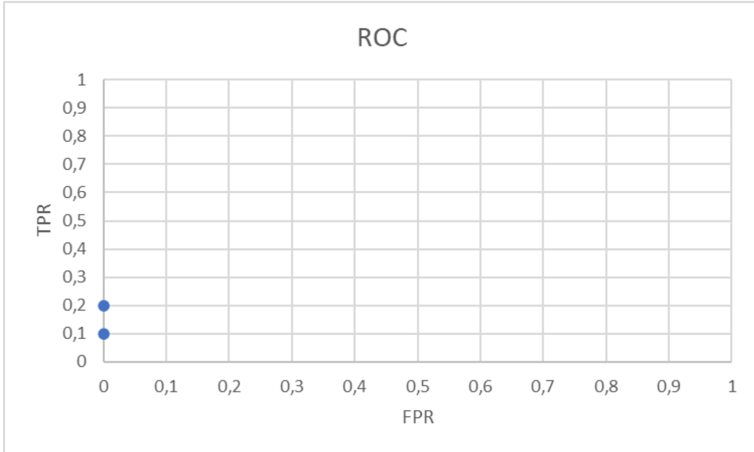
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



DEPARTAMENTO **CIÊNCIA E TECNOLOGIA**

Calculation of the ROC curve

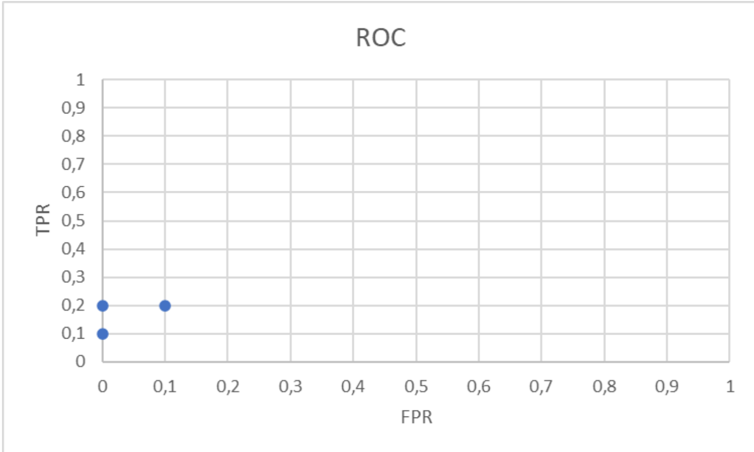
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Calculation of the ROC curve

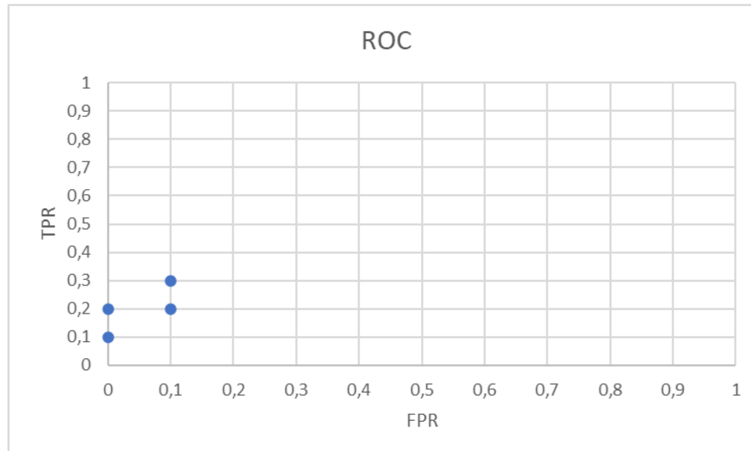
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Calculation of the ROC curve

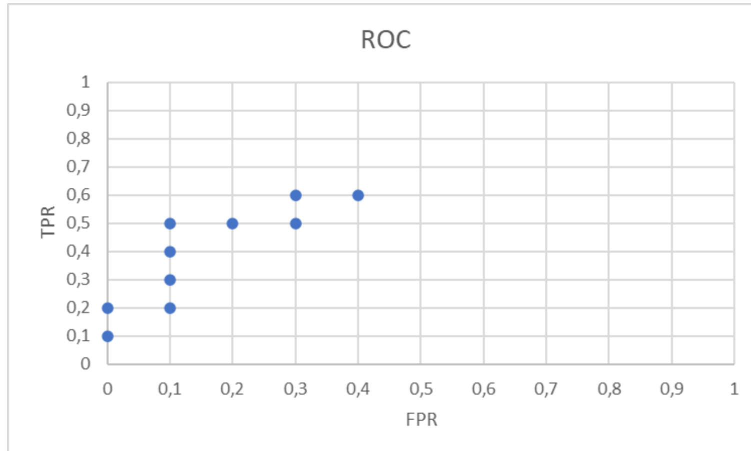
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



DEPARTAMENTO CIÊNCIA E TECNOLOGIA

Calculation of the ROC curve

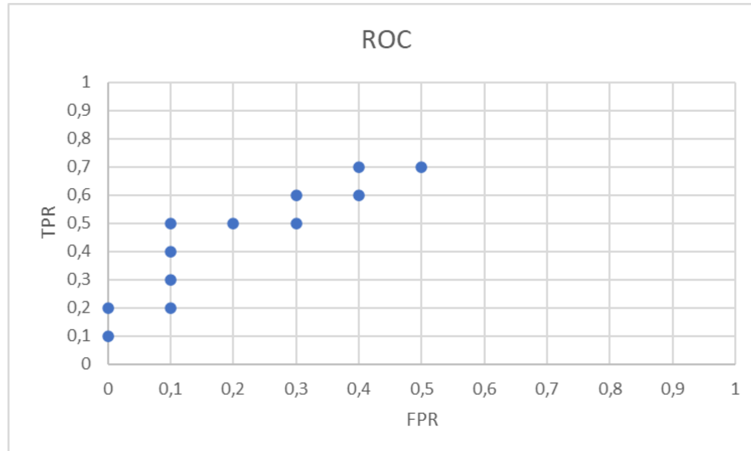
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



Calculation of the ROC curve

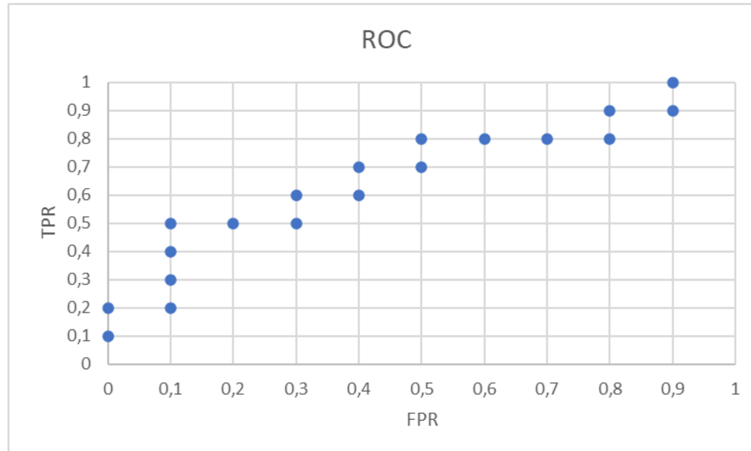
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



Calculation of the ROC curve

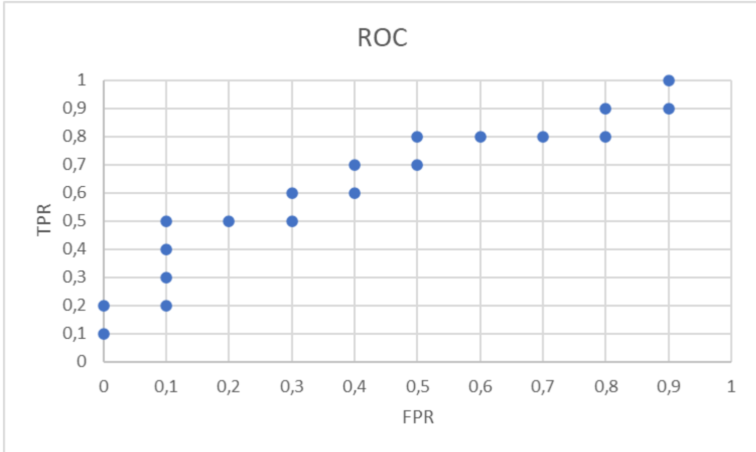
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



DEPARTAMENTO **CIÊNCIA E TECNOLOGIA**

Calculation of the ROC curve

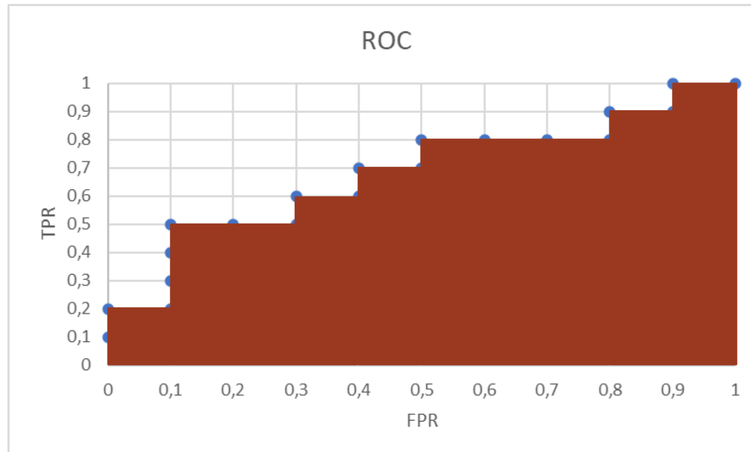
#	Classe	Prob(P)
1	P	0,9
13	P	0,8
10	N	0,7
15	P	0,6
7	P	0,55
6	P	0,54
17	N	0,53
20	N	0,52
2	P	0,51
14	N	0,505
8	P	0,4
9	N	0,39
18	P	0,38
12	N	0,37
5	N	0,36
11	N	0,35
3	P	0,34
4	N	0,33
19	P	0,3
16	N	0,1

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



$$\begin{aligned} AUC = & (0{,}1-0)\times(0{,}2-0) \\ & +(0{,}3-0{,}1)\times(0{,}5-0) \\ & +(0{,}4-0{,}3)\times(0{,}6-0) \\ & +(0{,}5-0{,}4)\times(0{,}7-0) \\ & +(0{,}8-0{,}5)\times(0{,}8-0) \\ & +(0{,}9-0{,}8)\times(0{,}9-0) \\ & +(1-0{,}9)\times(1-0) = 0.68 \end{aligned}$$

feature selection

feature selection

feature selection

Objective: reduction of the number of *features* to be considered by the model

- Reduce computational cost
- increase performance

Methods that evaluate a statistical relationship between the *features* and the target and choose the *features* with the strongest relationship

feature

- redundant
- not informative

feature techniques *selection*

- ***Feature selection*** **:** Select a subset of the
  - ***Unsupervised*** **:** Do not use the target variable (eg : remove redundant) → Correlation
  - ***Supervised*** **:** Use the target variable (eg remove irrelevant)
    - ***Wrapper*** **:** Look for subsets of *features* that lead to high performance → RFE [1]
    - ***Filter*** **:** Select subsets of features based on their relationship to the target
      - → Statistical methods
      - → *feature importance* [2][3]
    - ***Intrinsic*** **:** Algorithms that automatically execute *feature selection* → decision trees, Random Forests [4]

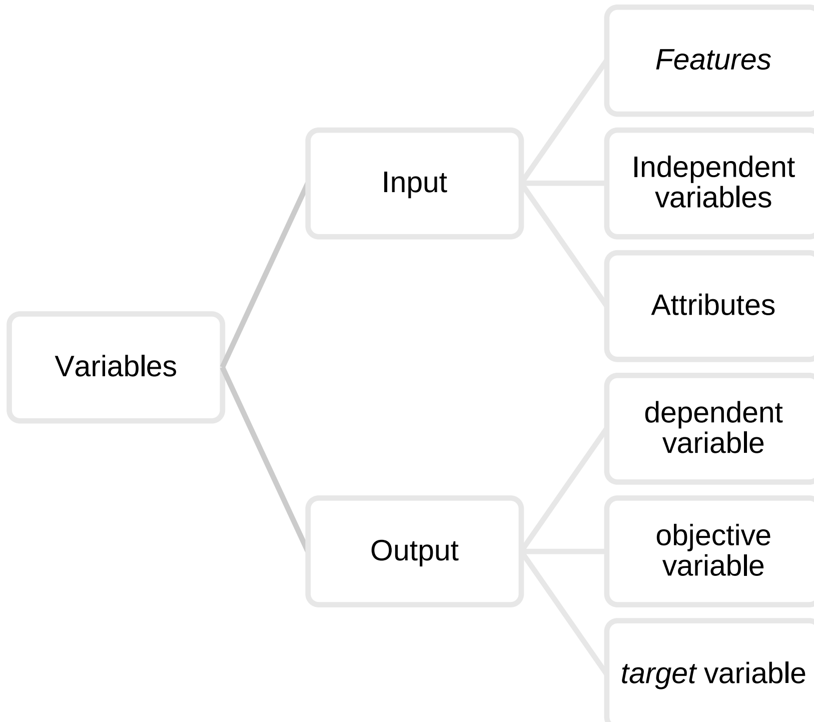
[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html#sklearn.feature_selection.RFECV

[2] Choose the *k features* most important: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

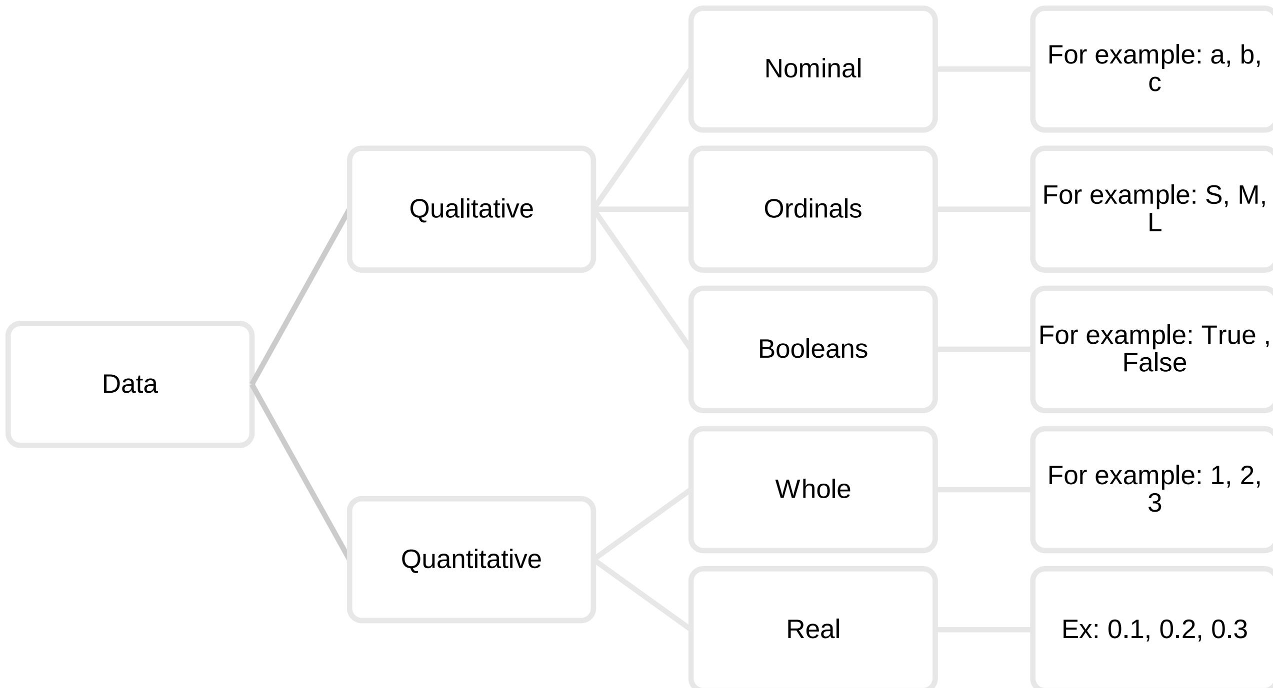
[3] Choose the *p% features* most important: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html

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

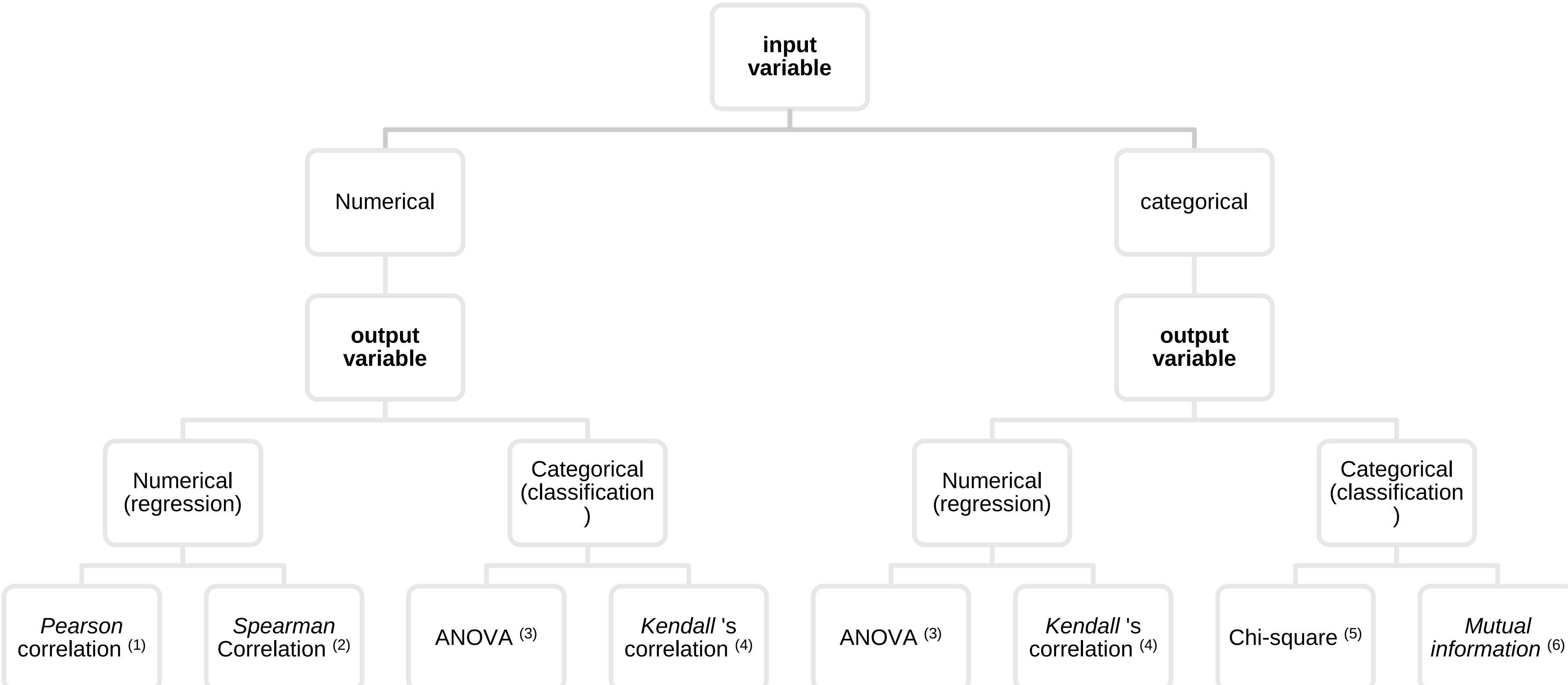
Review of variable types



Review of variable types



feature method *selection* ?



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

(two) 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

Regularization

Regularization

Regression form that regularizes (constrains, “shrinks”) the coefficients determined by the regression model

In regression, considering

- A *dataset* with independent variables X and dependent Y
- *fitting* process chooses the β coefficients in order to minimize a *loss works*
- at *loss function* is *Residual Sum of squares* (RSS):

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

The coefficients are adjusted taking into account the totality of the data

If there is noise in the data, the model is more flexible, but the estimated coefficients will not generalize well to new data.

Regularization “shrinks” these coefficients (they tend to zero)

Ridge regression

The function to be minimized becomes

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

On what λ

- It is the regularization factor that determines how much we want to penalize model flexibility.
- It is a parameter to be defined, for example, with *cross validation* .
    - $\lambda=0$: the penalty has no effect and the coefficients produced are the estimated ones
    - $\lambda\rightarrow\infty$: the impact of the regularization factor increases and the coefficients β will approach zero

The coefficients produced by this method are known as *L2 norm*

The coefficients no longer obey the scale of the *features* , so it is necessary to standardize them , using:

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

LASSO regression

The function to be minimized becomes

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

Similar to *Ridge regression*

As it uses the modulus instead of the square, it does not penalize large coefficients as much

The coefficients produced by this method are known as *L1 norm*

As it penalizes low coefficients to the same extent as high coefficients, it ends up executing *feature selection*

UNIVERSIDADE
PORTUCALENSE

Do conhecimento à prática.