Linear Regression cheatsheet

intuition and maths

linear regression model is based on the idea that the relationship between one response variable Y and a set of explanatory variables X can be expressed in a linear form as the following:

$$y_i = \beta_0 + \beta_1 * x_i$$

having the beta coefficients measuring the level of changed on y determined from a change of x

kind of data

as X:

- categorical variables
- continuous variables

as Y:

- continuous variables
- categorical variables (but may generate out of domain estimates) estimates

how to fit in R

lm(y ~ .) fits y against all variables
stepAIC() fits linear regression with stepwise procedure
pcr() fits principal component regression

assumptions

1.absence of multicollinearity among explanatory variables
2.absence of autocorrelation in model residuals
3.absence of correlation between residual variance and fitted values

how to test them in R

1.vif(): <=10 passed, >10 ko
2.DurbinWatsonTest(lm_object): 0 positive correlation (ko), 2 absence of positive correlation (passed) 4 negative correlation (ko)
3.ncvTest(): p-value <0.05 ok, p-value >= 0.05 ko

Logistic Regression cheatsheet

intuition and maths

logistic regression is a model developed to describe a process influenced from one or more explanatory variables and having a possible outcome enclosed within an upper and a lower bound. It is based on the following formula:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

in case of a binary response variable is possible to interpret its prediction as the probability of one of the two outcomes showing up

kind of data

as X:

- categorical variables
- continuous variables

as Y:

- boolean/binary variables
- continuos variables included with and upper and a lower bound (opportunely rescaled on the 0-1 range)

how to fit in R

g*lm(y ~ ., family = 'binomial')* fits y against all variables

assumptions

1.absence of autocorrelation in model residuals
2.absence of multicollinearity between variables
3.linear relationship between explanatory variables and log odds

how to test them in R

1.DurbinWatsonTest(glm_object): 0 positive correlation (ko), 2 absence of positive correlation (passed) 4 negative correlation (ko) - *applicable to time series data only*
2.vif(): <=10 passed, >10 ko
3.fitting alternative model showing cubic and quadratic explanatory variables: if this shows being statistically significant the assumption is broken

Support Vector Machine cheatsheet

intuition and maths

SVM is classificarion algorithm based on the concept of hyperplane. This hyperplane, which in 2D population can be considered as common plane, is employed to divide into two groups the population of response variable so to minimise the number of observations grouped with the wrong group.

Beside a linear version of linear hydroplane it is also possible to define non linear and even radial hyperplanes, which usually shows better level of performance.

kind of data

as X:

- categorical variables
- continuous variables

as Y:

- boolean/binary variables

how to fit in R

svm*(y ~ ., data= data.frame, kernel = 'linear')* fits y against all variables. alternative kernels:

- radial
- polynomial
- sigmoid

assumptions

1.IID : indipendent and identically distributed variables

how to test them in R

1.indipendente should be verified looking to the nature of variables, verifying that the different outputs are not influenced one from each other. A classical example is dice rolling, when the records represented from the sequence of rolling are independent
2.identically distribution should be checked looking at frequency distribution: is it coherent with the expected one? is there any structural change that could interfere with this distribution?

Random Forest cheatsheet

intuition and maths

Random Forest is a classification algorithm based on the combination of multiple decision trees. Each decision tree is grown iteratively splitting the popupation into clusters based on the value of explanatory variable, looking for rules able to let you correctly assign each record to the right category.

Random forest is then obtained assigning to each record the category assigned from the majority of decision trees.

kind of data

as X:

- categorical variables
- continuous variables

as Y:

- categorical data (also available regression variant for continuous data)

how to fit in R

randomForest*(y ~ ., data= data.frame, ntree = n)* grows *n* trees to predict the category of y starting from values of all the explanatory variables

assumptions

1.IID : indipendent and identically distributed variables

how to test them in R

1.indipendente should be verified looking to the nature of variables, verifying that the different outputs are not influenced one from each other. A classical example is dice rolling, when the records represented from the sequence of rolling are independent
2.identically distribution should be checked looking at frequency distribution: is it coherent with the expected one? is there any structural change that could interfere with this distribution?