machine learn end to end helper

sian davies | March 8, 2021

The Machine Learning Process

From "Introduction to Microsoft Azure" by David Chappell

Data Preprocessing in Python Machine Learning

01 Rescaling Data
03 Standardizing Data
05 Normalizing Data
07 Binarizing Data
02 Mean Removal
04 One Hot Encoding
06 Label Encoding

Start

get your tools into your notebook eg pandas matplotlib seaborn, numpy

get data into python as data frame
read_csv or sqlalchemy
+ connect string
+MySQL query

EXPLORE
using pd/numpy commands and EDA techniques (shape, info, dtypes, describe, head/tail, iloc sampling, plot)

TIP : annotate your steps to make it easier for yourself to come back and amend the processes later

iterate

CLEAN
rename, drop dupes, drop rows/columns, deal with NULL/NAN, select columns change dtypes check/set your index

TIP: save your workbook in stages using checkpoints or save as... then you can always re start the kernel and run part of your workflow

Sense check -frame the question
-whats the unknown variable?
-what model are you going to use?
-whats the success criteria?
-where is the data?
(optional) -review data in SQL /excel /other and brainstorm at this stage

Sense check -
do you have the data you need to support the task? are some of the columns too full of unknowns to be useful? do you have too many columns to work with? look at the definitions - do you understand what each column contains? do you have 2 or more similar columns? are any columns the incorrect data type? do you have to merge data sets?

Sense check - which columns to use for this attempt?
considering the task, the data set and the type of model you are going to use, what would be a sensible starting list of features to be in your X defintion (could start small... and then add more in later iteration)

correlation matrix heatmap, use VIF, plot scatter, chi-square, regression, R matrix

check for multicollinearity

Yes

high correlation between variables?

No

split categorical and numerical fields

Sense check - which variable in your data is the dependent variable, ie y, which you are going to predict? how will you treat that in these stages?

categories

numbers

Category
feature engineering / wrangling
create buckets
assign nulls to 'unknown' bucket or drop use labels, dummies or hot encoder

numerical
feature engineering / wrangling
replace nulls with mean or interpolate remove outliers using IQR x1.5, rescale as needed

Sense check - many ML algorithms require normally distributed data, some can only work with numerical data. All algorithms have assumptions hint: check histograms

yes

do features need normalisation? eg BoxCox

Yes

attempt normaliser

No

how does the data look? more pre processing needed?

No

return to any earlier stage

Yes

Sense check
-finding the sweet spot in % train/test which allows the model to learn, but leaves enough to test it afterwards - you should have split out X and y independent and dependant variables already

train test split

do i need to iterate n times? should i review my feature selection? Can I change anything to get a more accurate model?

evaluate the model, confusion matrix, AUC, variance - R^2, P value, MSE, MAE, bias, variance, overfit

apply model (eg regression /+ KNN) => candidate model

No

Could i enhance my model with KNN?

Yes