Train In Data

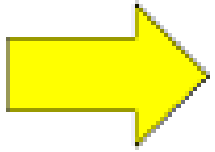
One hot encoding

One hot encoding: definition

- One hot encoding, consists in encoding each categorical variable with a set of boolean variables which take values 0 or 1, indicating if a category is present for each observation.

One hot encoding : example
k dummy variables

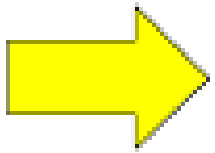
Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

One hot encoding : example k-1 dummy variables

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow
1	0
1	0
0	1
0	0

One hot encoding into k -1 variables

- More generally, a categorical variable should be encoded by creating k-1 binary variables, where k is the number of distinct categories.
- In the case of binary variables, like gender where k=2 (male / female)we need to create only 1 (k - 1 = 1) binary variable.

One hot encoding into k-1 binary variables takes into account that we can use 1 less dimension and still represent the whole information:

if the observation is 0 in all the binary variables, then it must be 1 in the final (not present) binary variable.

One hot encoding into k -1 variables

Most machine learning algorithms, consider **the entire data set** while being fit.

Therefore, encoding categorical variables into k - 1 binary variables, is better, as it avoids introducing redundant information.

One hot encoding into k variables

There are a few occasions when it is better to encode variables into k dummy variables:

- when building tree based algorithms
- when doing feature selection by recursive algorithms
- when interested in determine the importance of each single category

One hot encoding: Advantages

- Makes no assumption about the distribution or categories of the categorical variable
- Keeps all the information of the categorical variable
- Suitable for linear models

One hot encoding: Limitations

- Expands the feature space
- Does not add extra information while encoding
- Many dummy variables may be identical, introducing redundant information

Train In Data

THANK YOU

www.trainindata.com