

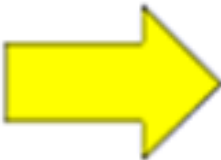


# Count or frequency encoding

# Count / frequency encoding: definition

- Categories are replaced by the count or percentage of observations that show that category in the dataset.
- Captures the representation of each label in a dataset
- Very popular encoding method in Kaggle competitions.
- Assumption: the number observations shown by each category is predictive of the target.

# Count encoding: example



Color
Red
Red
Yellow
Green
Yellow

Color
2
2
2
1
2

# Frequency encoding: example

Color	Color
Red	0.4
Red	0.4
Yellow	0.4
Green	0.20
Yellow	0.4



# Count / frequency encoding: Advantages

- Straightforward to implement
- Does not expand the feature space
- Can work well enough with tree based algorithms

# Count / frequency encoding: Limitations

- Not suitable for linear models
- Does not handle new categories in test set automatically
- If 2 different categories appear the same amount of times in the dataset, that is, they appear in the same number of observations, they will be replaced by the same number:
  - may lose valuable information.

# THANK YOU

[www.trainindata.com](http://www.trainindata.com)