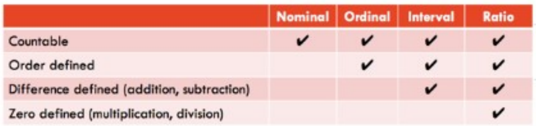
Machine Learning Data Processing

2021年1月25日 18:30

PS: Pre-Learning
先导课

(一). Data Type.

- Nominal: < mutal exclusive >.
- Ordinal: < for one catelogue. but in the order matters >
- Interval: < the difference, step between the 2 value >.
- Ratio: < has all the propertires of an interval value >.



	Nominal	Ordinal	Interval	Ratio
Countable	✓	✓	✓	✓
Order defined		✓	✓	✓
Difference defined (addition, subtraction)			✓	✓
Zero defined (multiplication, division)				✓

(二). Feature Module.

1). Feature Cleaning.

- Missing value: Method < Way > ①: Completion Algorithms.
  ②: Omit < ignore > Elements.

- Special value: Like INF. NA. INAN. Need to be cleaned.

- Outliers: Over the Limitted range. Should be detected but not necessary.

- Obvious inconsistencies: Some value can't be admitted like a man can't pregenant.

2) Feature Imputation: Like Hot - Deck、Cold Deck. Some Libraries.

I think that it may be some instrument of Data Feature.

3) Feature Selection:

① Correlation: Features should be uncorrelated < [illegible] > $\mathrm{corr}(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$

② Dimensionality Reduction: Reduce the Dimension. ↓ ND → < N-i > D.

i. PCA < Principal Component Analysis >.

ii. SVD < Singular Value Decomposition >.

③ Importance: Select the Features by these methods:

i. Filter Methods.

ii. Wrapper Methods.

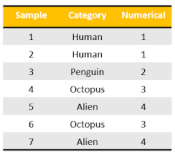
iii. Embedded. Methods.

4). Feature Encoding: All features must be numeric. Encoding help it.

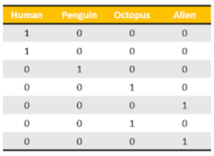
Male: 0 ; Female: 1

i. Label Encoding. ii. One Hot Encoding

i. Label Encoding. ii. One Hot Encoding



Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4



Human	Penguin	Octopus	Alien
1	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	1	0
0	0	0	1

5) Feature Normalization or Scaling :

Since the range of raw data is widely. Need to normal the data which makes the marchine learning can work properly.

$a[i] = \dfrac{a[i]}{\max(a)}$, which make the $a[i] \in [0,1]$. if it is positive.

i. Rescaling. → the simplest is the range in $[0,1]$ or $[-1,1]$.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

ii. Standardization → Make the values of each features to have zero-mean and unit variance. $x' = \frac{x - \bar{x}}{\sigma}$

iii. Scaling to unit length :→ has length one. $x' = \frac{x}{||x||}$

(三) DataSet Construction

- Training Dataset : A set of examples used for Learning.
- Test Dataset : A fully-trained classifer.
- Validation Dataset : A set of examples used to tune the param.
- Cross Validation :