Feature Engineering

Moriarty

April 17, 2020

1 Continuous Variable

In this section, We tackle the continuous variables.

1.1 Nondimensionalization

Standarlization / Normalization

1.2 Data transformation

Transform data to adapt the assumption we supposed, so as to better statistical inference. Intuitively, the transofrmation is to

1.3 Discretization of continuous variables

2 Discrete Variable

3 Categorical Variable

3.1 Ordered

Label encoding / Count encoding / Target encoding(aka Mean encoding)

Label encoding : encode categories resolving around their rank.

Count encoding : each categories is replaced with the count for samples having that category.

Target encoding : each categories is replaced with the mean target value for samples having that category.

3.2 Disorderd

Dummy variable encodeing / one-hot encoding Dummy variable encoding/one-hot encoding are exactly the same thing;the former term comes from statistics and the latter from computer science(borrowed from electronics). The dummy variable trap: In linear regression, the normal equation describes thita = (XTX)(t(X)Y).Noticed that t(X)X

may not be invertible when the column vectors of X are correlative.For a instance, if we want to encode gender(aka "sex") to 0 and 1, we might get two columns like

Name	sex_0	sex_1
Alice	0	1
Bob	1	0
Cindy	0	1
Donald	1	0

Obviously the two columns, precisely the two column vectors, are correlative, incurring the result that the dataset matrix is non-invertible.(But it tends to be less of a concern in ML because multicollinearty is only a big issue in linear regression, and can be minimized with regularization.) Tikhonov regularization(aka. ridge regression) is a practical regularization to solve the problem above.As a biased estimation method of linear regression specialized in collinearity data analysis, the essence of ridge regression is a promoted OLS(Ordinary Least Squares), enhancing reliability of the regression coefficients by losing some information and reducing the precision.It fits in ill-conditioned data better than OLS.

4 Time/Date type Variable

5 Missing Value Processing

6 Feature Combination

More text.