Performing Feature Selection

Xavier Morera

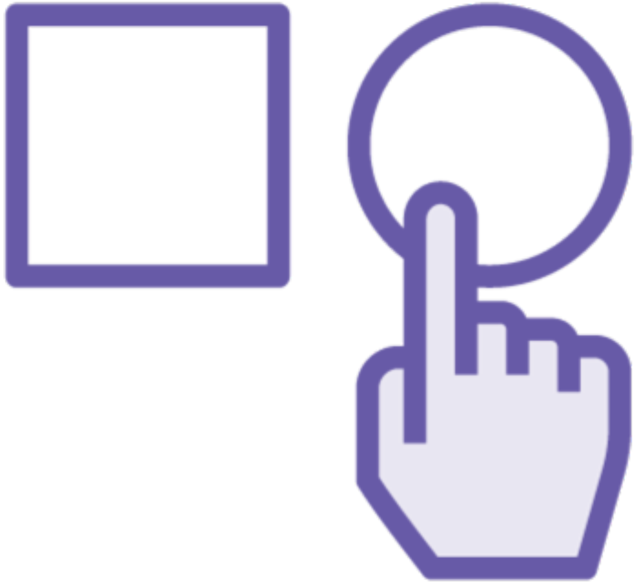
HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera   www.xaviermorera.com

Performing Feature Selection

Performing Feature Selection

Feature Selection



Correlation does not imply causation

Covariance

Choosing the right features

- Better predictions

Understanding Feature Selection

Apply statistical tests to inputs given a specified output

Determine columns that are more predictive of the output

"Not all features are created equal."

Remember this…

Understanding Feature Selection

Selection

Ranking

Azure ML Studio Feature Ranking Algorithms

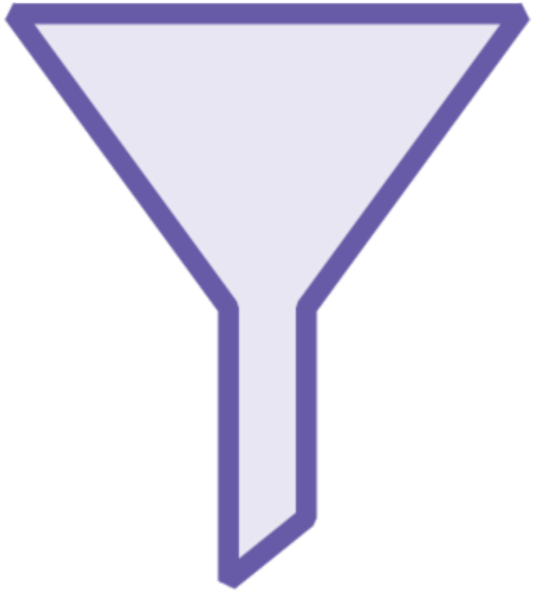
Filter Based Feature Selection

Fisher Linear Discriminant Analysis

Permutation Feature Importance

Compute Linear Correlation

Filter Based Feature Selection



Module used to filter out irrelevant features

- Redundant

Calculate a score

- Remove columns

Different algorithms available

Feature Scoring Methods

Pearson Correlation

Mutual Information

Kendall Correlation

Spearman Correlation

Chi Squared

Fisher Score

Count Based

Demo

Filter Based Feature Selection

Fisher Linear Discriminant Analysis

Identify linear combination

- Feature variables
- Best group data
- Into separate classes

Dimensionality reduction

Demo

Fisher Linear Discriminant Analysis

Permutation Feature Importance

Compute

- Permutation feature importance scores
- Feature variables
- Given a trained model
- And test dataset

Returns an ordered list

- Feature variables with scores

Demo

Permutation Feature Importance

Linear Correlation

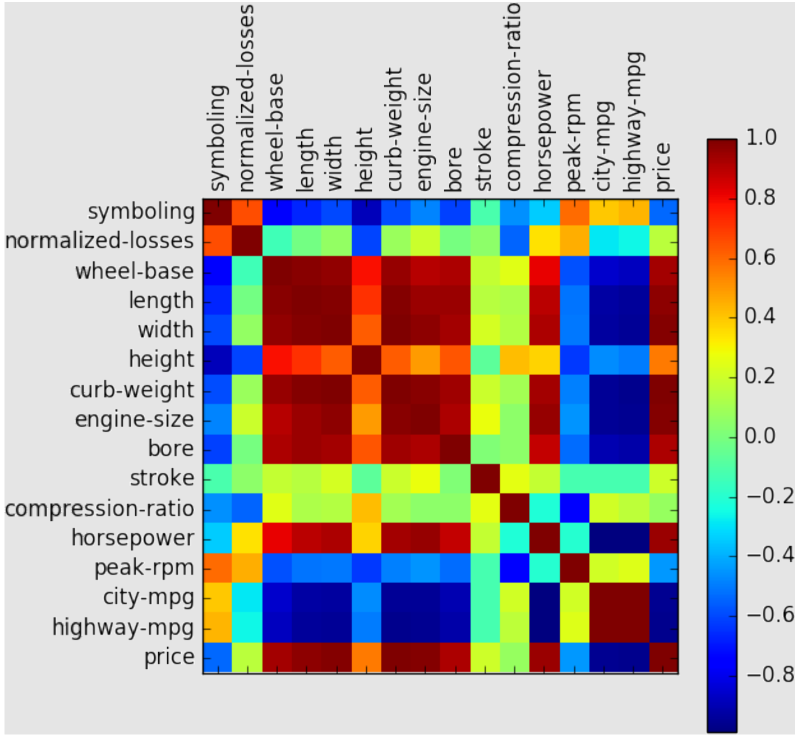
Pearson correlation coefficients

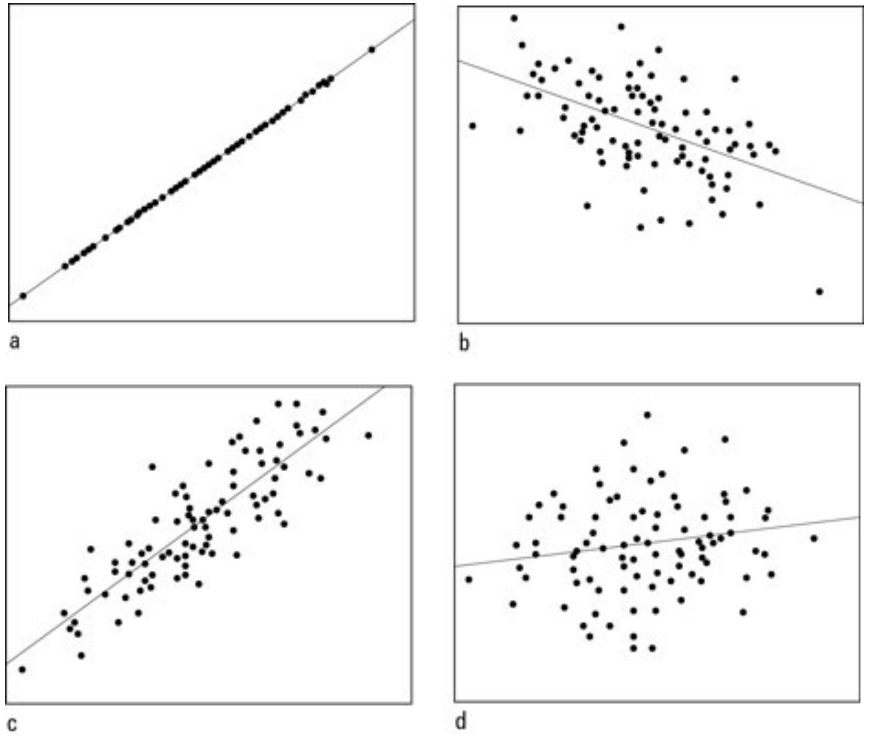
- Linear correlation
- Each pair of values in a dataset

Infer the strength of the relationship

- Two variables
- Positive or negative correlation

Linear Correlation





Demo

Compute Linear Correlation

Takeaway

Feature selection

- Select features
- More relevant for our scenario

Applying statistical inputs

- Given the specified ouput
- Determine those columns
- More predictice of the ouput

Takeaway

Not all features are created equal

Filter based feature selection

Fisher Linear Discriminant Analysis

Permutation Feature Importance

Linear Correlation