FILTER METHODS

STATISTICAL AND RANKING PROCEDURES

STATISTICS AND RANKING METHODS

Statistical – ranking methods

- Information Gain
- Fisher Score
- Univariate tests
- Univariate roc-auc / rmse

STATISTICS AND RANKING METHODS

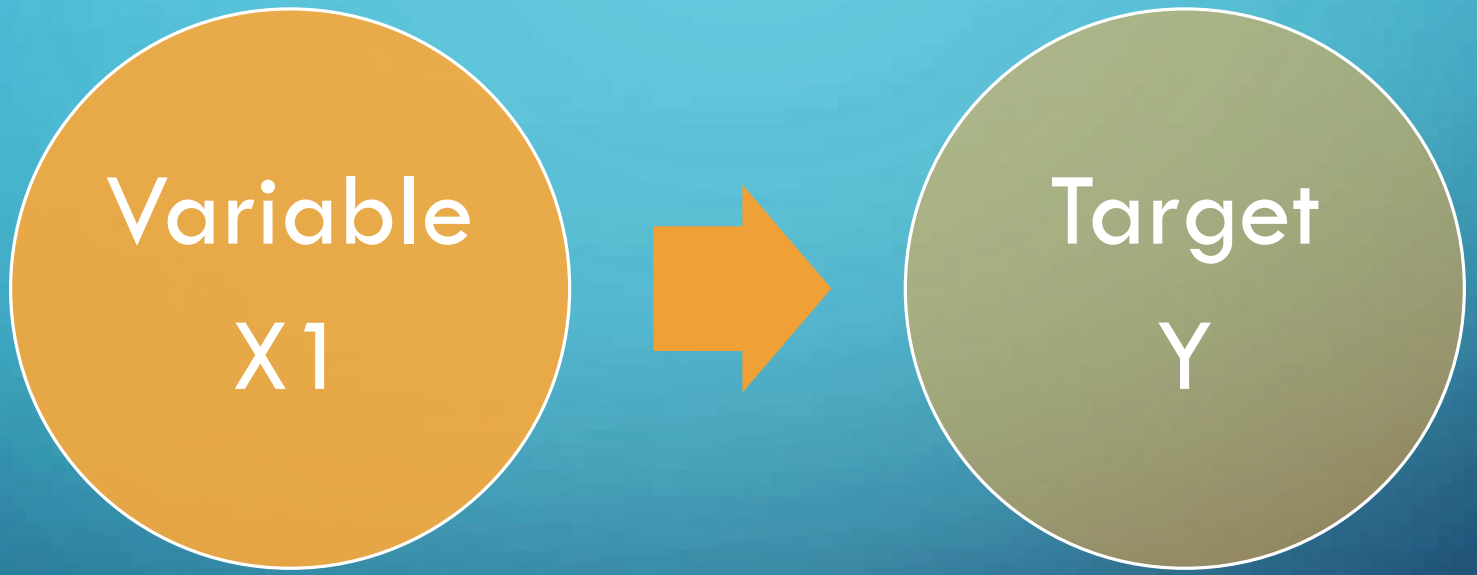
Two steps:

1. Rank features based on certain criteria / metric
2. Select features with highest rankings

Pros and Cons

- Fast
- Does not contemplate feature redundancy

STATISTICS AND RANKING METHODS



Evaluate if the variable is important to discriminate the target

MUTUAL INFORMATION

- Measures the mutual dependence of 2 variables
- Determines how similar the joint distribution $p(X,Y)$ is to the products of individual distributions $p(X)p(Y)$
- If X and Y are independent, their MI is zero
- If X is deterministic of Y, the MI is the uncertainty in X.

MUTUAL INFORMATION

$$\sum_{i,y} P(xi, yj) \times \log \frac{P(xi, yj)}{P(xi)P(yj)}$$

FISHER SCORE

- Measures the dependence of 2 variables
- Suited for categorical variables.
- Target should be binary
- Variable values should be non-negative, and typically Boolean, frequencies, or counts.
- It compares observed distribution of class among the different labels against the expected one, would there be no labels

FISHER SCORE

	Male	Female	Total Row
Survived = 1	2	9	11
Survived = 0	10	3	13
Total column	12	12	24

	Male	Female	Total Row
Survived = 1	0.17	0.75	0.46
Survived = 0	0.38	0.25	0.54
Total column	1	1	1

FISHER SCORE

Fisher Score [10]: Features with high quality should assign similar values to instances in the same class and different values to instances from different classes. With this intuition, the score for the i-th feature S_i will be calculated by Fisher Score as,

$$S_i = \frac{\sum_{k=1}^{K} n_j(\mu_{ij} - \mu_i)^2}{\sum_{k=1}^{K} n_j \rho_{ij}^2}, \tag{0.2}$$

where μ_{ij} and ρ_{ij} are the mean and the variance of the i-th feature in the j-th class respectively, n_j is the number of instances in the j-th class, and μ_i is the mean of the i-th feature.

UNIVARIATE TESTS

- Measures the dependence of 2 variables ➔ ANOVA
- Suited for continuous variables
- Requires a binary target
  - Sklearn extends the test to continuous targets with a correlation trick
- Assumes linear relationship between variable and target
- Assumes variables are normally distributed
- Sensitive to the sample size

UNIVARIATE ROC-AUC / RMSE

- Measures the dependence of 2 variables ➔ using machine learning
- Suited for all types of variables
- Makes no assumption on the distribution of the variables

UNIVARIATE ROC-AUC / RMSE

- Builds decision tree using a single variable and the target
- Ranks the features according to the model roc-auc or rmse
- Selects the features with the highest machine learning metrics

roc-auc = 0.5 means random