Train In Data

Select features based on p-values

Statistical tests in sklearn

- Information Gain
- Chi-square
- Anova

Statistical tests in sklearn

- Information Gain
- Chi-square
- Anova

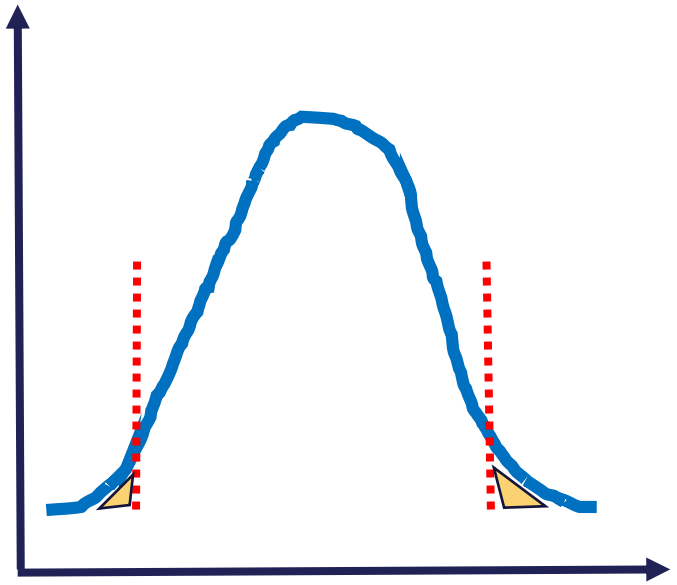
- `SelectKBest` removes all but the k highest scoring features
- `SelectPercentile` removes all but a user-specified highest scoring percentage of features

Statistical tests in sklearn

- SelectFPR
- SelectFDR
- SelectFwe

Select features based on the p-values returned by the tests (Anova, chi-square)

What is a p-value?



- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.
- A p-value of 0.05 or lower is generally considered statistically significant.

SelectFDR

```
def _get_support_mask(self):
    check_is_fitted(self)

    return self.pvalues_ < self.alpha
```

Select features which p-value is smaller than a certain threshold, typically 0.05.

SelectFwe

```
def _get_support_mask(self):
    check_is_fitted(self)

    return self.pvalues_ < self.alpha / len(self.pvalues_)
```

To minimize the false positive rate we divide alpha by the number of features.

Equivalent of the Bonferroni correction.

SelectFDR

- p-value <= alpha
- alpha changes feature per feature based on the Benjamini-Hochberg correction
- Alpha = rank / number of features * alpha
- Rank is given by the p-value

GenericUnivariateSelect

sklearn.feature_selection.GenericUnivariateSelect

class sklearn.feature_selection.**GenericUnivariateSelect**(*score_func=<function f_classif>*, ***, *mode='percentile'*, *param=1e-05*) [source]

Univariate feature selector with configurable strategy.

Read more in the User Guide.

Parameters:

score_func : *callable, default=f_classif*
Function taking two arrays X and y, and returning a pair of arrays (scores, pvalues). For modes 'percentile' or 'kbest' it can return a single array scores.

mode : *{'percentile', 'k_best', 'fpr', 'fdr', 'fwe'}, default='percentile'*
Feature selection mode.

param : *float or int depending on the feature selection mode, default=1e-5*
Parameter of the corresponding mode.

Train In Data

THANK YOU

www.trainindata.com