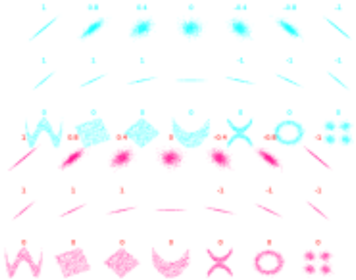
Introduction to Machine Learning

Feature Selection: Filter Methods

Feature Selection: Filter Methods



Learning goals

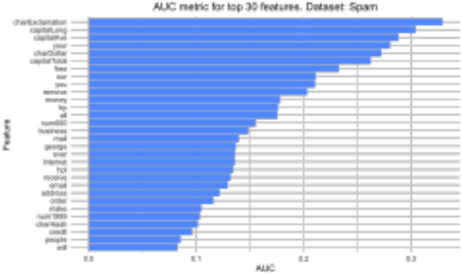
- Understand how filter methods work and how to apply them for feature selection.
- Know filter methods based on correlation, test statistics, and mutual information.

Learning goals

- Understand how filter methods work and how to apply them for feature selection.
- Know filter methods based on correlation, test statistics, and mutual information.

INTRODUCTION

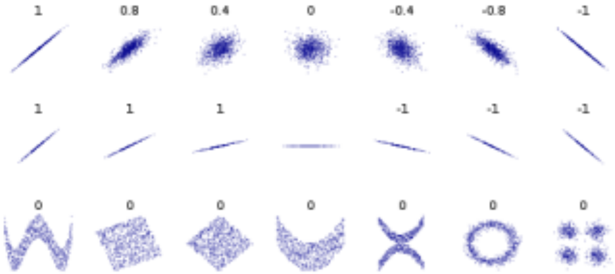
- **Filter methods** construct a measure that quantifies the dependency between features and the target variable.
- They yield a numerical score for each feature x_j, according to which we rank the features.
- They are model-agnostic and can be applied generically.



Exemplary filter score ranking for Spam data

PEARSON & SPEARMAN CORRELATION / 2

Only **linear** dependency structure, non-linear (non-monotonic) aspects are not captured:



Comparison of Pearson correlation for different dependency structures.

To assess strength of non-linear/non-monotonic dependencies, generalizations such as **distance correlation** can be used.

MUTUAL INFORMATION (MI)

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right]$$

- Each feature x_j is rated according to $I(x_j; y)$; this is sometimes called information gain (IG).
- MI measures the amount of "dependence" between RV by looking how different their joint dist. is from strict independence $p(X)p(Y)$.
- MI is zero iff $X \perp\!\!\!\perp Y$. On the other hand, if X is a deterministic function of Y or vice versa, MI becomes maximal.
- Unlike correlation, MI is defined for both numeric and categorical variables and provides a more general measure of dependence.
- To estimate MI: for discrete features, use observed frequencies; for continuous features, binning, kernel density estimation is used.