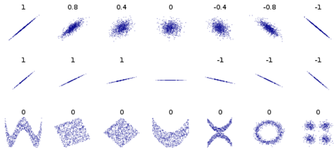
Introduction to Machine Learning

Feature Selection
Feature Selection: Filter Methods

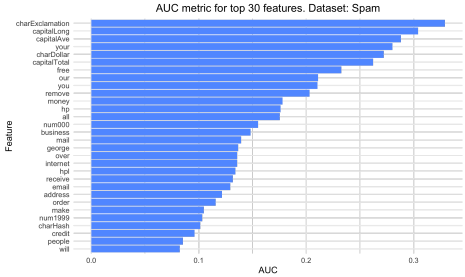


Learning goals

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

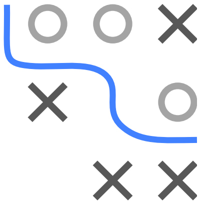
INTRODUCTION

- **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

χ^2-STATISTIC

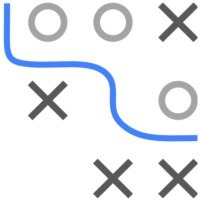


- Test for independence between categorical x_j and cat. target y. Numeric features or targets can be discretized.
- Hypotheses:
  $H_0 : p(x_j = m, y = k) = p(x_j = m)\, p(y = k)\, \forall m, k$
  $H_1 : \exists\ m, k : p(x_j = m, y = k) \neq p(x_j = m)\, p(y = k)$
- Calculate χ^2-statistic for each feature-target combination:

$$\chi_j^2 = \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\frac{e_{mk} - \tilde{e}_{mk}}{\tilde{e}_{mk}}\right) \quad \underset{approx.}{\overset{H_0}{\sim}} \quad \chi^2((M-1)(K-1)),$$

  where e_{mk} is observed relative frequency of pair (m, k), $\tilde{e}_{mk} = \frac{e_{m\cdot} e_{\cdot k}}{n}$ is expected relative frequency, and M, K are number of values x_j and y can take
- The larger χ_j^2, the more dependent is the feature-target combination $\rightarrow$ higher relevance

PEARSON & SPEARMAN CORRELATION



Pearson correlation $r(x_j, y)$:

- For numeric features and targets only
- Measures linear dependency
- $r(x_j, y) = \frac{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)}\sqrt{(\sum_{i=1}^{n} y^{(i)} - \bar{y})}}, \qquad -1 \leq r \leq 1$

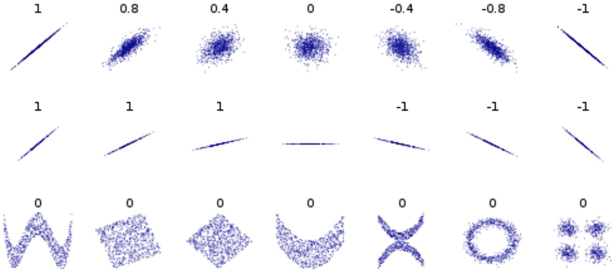
Spearman correlation $r_{SP}(x_j, y)$:

- For features and targets at least on ordinal scale
- Equivalent to Pearson correlation computed on ranks
- Assesses monotonicity of relationship

Use absolute values $|r(x_j, y)|$ for feature ranking:
higher score indicates a higher relevance

PEARSON & SPEARMAN CORRELATION

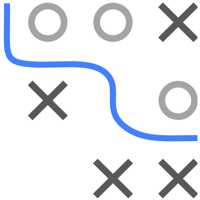
Only **linear** dependency structure, non-linear (non-monotonic) aspects are not captured:



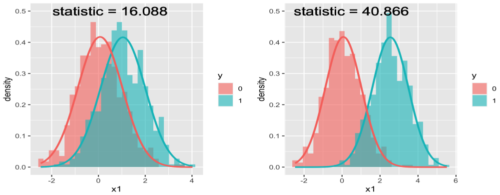
Comparison of Pearson correlation for different dependency structures.

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

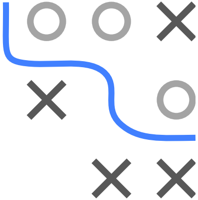
WELCH'S t-TEST



- For binary classification with $\mathcal{Y} = \{0, 1\}$ and numeric features
- Two-sample t-test for samples with unequal variances
- Hypotheses: $H_0: \quad \mu_{j_0} = \mu_{j_1} \qquad$ vs. $\qquad H_1: \quad \mu_{j_0} \neq \mu_{j_1}$
- Calculate Welch's t-statistic for every feature x_j
  $t_j = (\bar{x}_{j_0} - \bar{x}_{j_1}) / \sqrt{(S^2_{x_{j_0}}/n_0 + S^2_{x_{j_1}}/n_1)}$
  ($\bar{x}_{j_y}$, $S^2_{x_{j_y}}$ and n_y are the sample mean, variance and sample size)
- Higher t-score indicates higher relevance

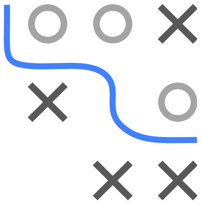


AUC/ROC



- For binary classification with $\mathcal{Y} = \{0, 1\}$ and numeric features
- Classify samples using single feature (with thresholds), compute AUC per feature as proxy for its ability to separate classes
- Features are then ranked; higher AUC scores $\rightarrow$ higher relevance.

F-TEST



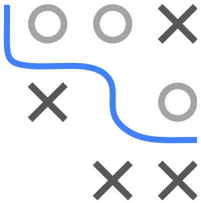
- For multiclass classification ($g \geq 2$) and numeric features
- Assesses whether the expected values of a feature x_j within the classes of the target differ from each other
- Hypotheses:
  $H_0 : \mu_{j_0} = \mu_{j_1} = \cdots = \mu_{j_g}$ vs. $H_1 : \exists\, k, l : \mu_{j_k} \neq \mu_{j_l}$
- Calculate the F-statistic for each feature-target combination:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

$$F = \frac{\sum_{k=1}^{g} n_k (\bar{x}_{j_k} - \bar{x}_j)^2 / (g-1)}{\sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{j_k}^{(i)} - \bar{x}_{j_k})^2 / (n-g)}$$

  where $\bar{x}_{j_k}$ is the sample mean of feature x_j where $y = k$ and $\bar{x}_j$ is the overall sample mean of feature x_j
- A higher F-score indicates higher relevance of the feature

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
- 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