FILTER METHODS

CORRELATION

FILTER METHODS - CORRELATION

- Correlation is a measure of the linear relationship of 2 or more variables
- Through correlation, we can predict one variable from the other
    - Good variables are highly correlated with the target
- Correlated predictor variables provide redundant information
    - Variables should be correlated with the target but uncorrelated among themselves

CORRELATION FEATURE SELECTION

The central hypothesis is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other

M. Hall 1999, Correlation-based Feature Selection for Machine Learning, PhD Thesis

CORRELATION AND MACHINE LEARNING

- Correlated features do not necessarily affect model accuracy per se.
- High dimensionality does
- If 2 features are highly correlated, the second one will add little information over the previous one: removing it helps reduce dimension
- Correlation affects model interpretability: linear models
- Different classifiers show different sensitivity to correlation

FILTER METHODS - CORRELATION

Pearson's correlation coefficient:

$$\frac{Sum\big(\ (X1 - X1mean) \times (X2 - X2mean) \times (Xn - Xn\_mean)\big)}{VarX1\ \times VarX2 \times\ VarXn}$$

Pearson's coefficient values vary between -1 and 1:

1 is highly correlated: the more of variable x1, the more of x2

-1 is highly anti-correlated: the more of variable x1, the less of x2