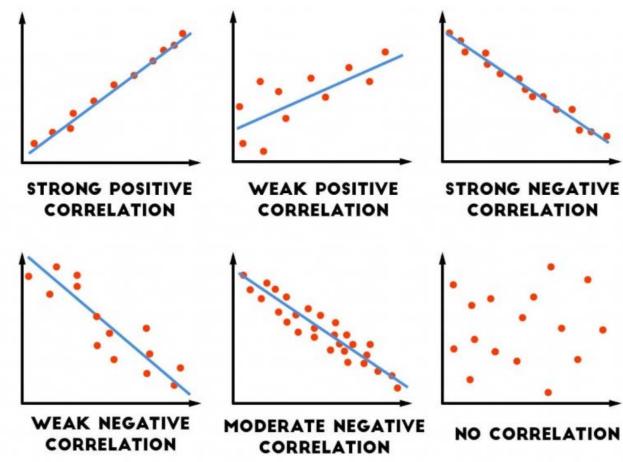
#### MACHINE LEARNING WITH PYTHON

# FEATURE SELECTION

Themistoklis Diamantopoulos

#### Correlation

Similar information



Source: http://danshiebler.com/2017-06-25-metrics/

#### **Correlation Statistics**

- Variables X, Y
- Variance  $var(X) = E \left[ (X E[X])^2 \right]$
- Covariance

$$cov(X,Y) = E[(X-E[X])(Y-E[Y])]$$

Correlation

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}}$$

Values from [-1, +1]+: Positive correlation-: Negative correlation0: No correlation

### **Mutual Information**

- Variables X, Y
- Measures the dependence between random variables

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left( \frac{p(x,y)}{p(x) \, p(y)} \right)$$

- Higher dependence → Higher mutual information
- If the variables are independent then mutual information is 0

## Chi-Squared Statistic

$$\chi^2 = \sum_{i=1}^n \frac{\left(O_i - E_i\right)^2}{E_i}$$

O<sub>i</sub>: actual count

*E*<sub>i</sub>: expected count

	Play Chess	Don't Play Chess	Sum
Like Science	250	200	450
Fiction	90	360	
Don't Like	50	1000	1050
Science Fiction	210	840	
Sum	300	1200	1500

Example:

$$\chi^{2} = \frac{\left(250 - 90\right)^{2}}{90} + \frac{\left(50 - 210\right)^{2}}{210} + \frac{\left(200 - 360\right)^{2}}{360} + \frac{\left(1000 - 840\right)^{2}}{840} = 507.93$$

Source: https://chrisalbon.com/machine\_learning/feature\_selection/chi-squared\_for\_feature\_selection/http://www.learn4master.com/machine-learning/chi-square-test-for-feature-selectionhttps://www.slideshare.net/GajanandSharma1/data-preprocessing-46881206