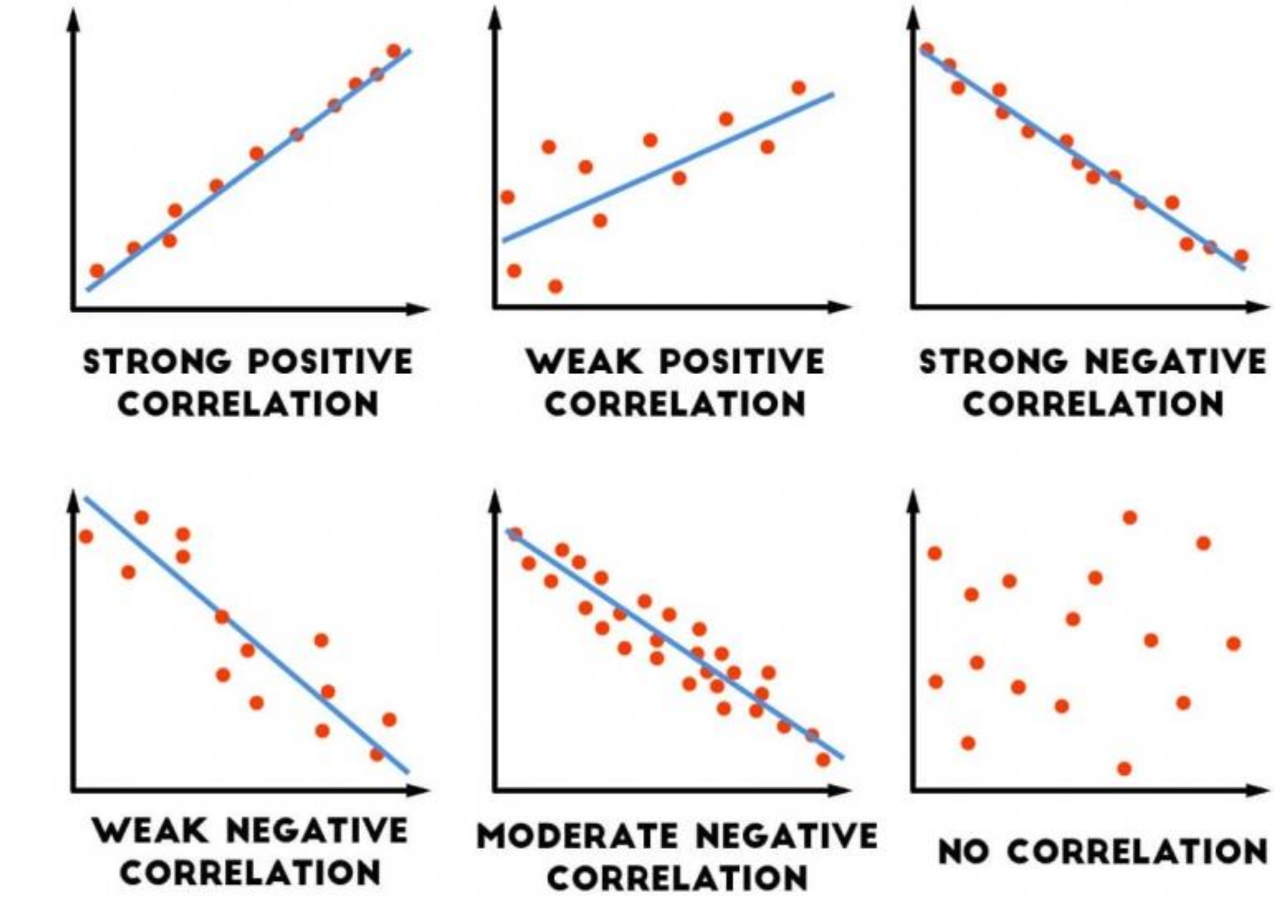
MACHINE LEARNING WITH PYTHON

FEATURE SELECTION

Themistoklis Diamantopoulos

Correlation

- Similar information



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

Correlation Statistics

- Variables X, Y
- Variance

$$\mathrm{var}(X) = E\left[\left(X - E[X]\right)^2\right]$$

- Covariance

$$\mathrm{cov}(X,Y) = E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right]$$

- Correlation

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

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

Source: https://en.wikipedia.org/wiki/Covariance

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

Source: https://www.slideshare.net/guo_dong/feature-selection
https://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html

Chi-Squared Statistic

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

O_i: actual count

E_i: expected count

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

- Example:

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

- Higher dependence → Higher chi-squared

Source: https://chrisalbon.com/machine_learning/feature_selection/chi-squared_for_feature_selection/
http://www.learn4master.com/machine-learning/chi-square-test-for-feature-selection
https://www.slideshare.net/GajanandSharma1/data-preprocessing-46881206