Mathematical Foundations of ML - Linear Correlation

We have data $x_i = \left(x_{i,1},\ldots,x_{i,n}\right), y_i$. We want to understand the correlation between two features $1 \le k \ne j \le n$, or a feature k and the response variable.

Feature mean

$$\mu_k := \frac{1}{\#I}\sum_{i\in I} x_{i,k} \ .$$

Feature STD

$$\sigma_k := \sqrt{\frac{1}{\#I}\sum_{i\in I}\left(x_{i,k}-\mu_k\right)^2} \ .$$

We then compute ***correlation coefficient***

$$-1 \le \frac{\frac{1}{\#I}\sum_{i\in I}\left(x_{i,k}-\mu_k\right)\left(x_{i,j}-\mu_j\right)}{\sigma_k\sigma_j} \le 1$$

Remarks

- If $x_{i,k} = x_{i,j}$, for each i, we obviously get perfect correlation of 1.
- Application – preprocessing step of pruning out highly correlated features. This potentially will also improve the tree-based feature importance algorithms.
- Application – preprocessing step of pruning out features with low correlation to outcome.
- Application – simplest form of feature importance algorithm. Sort the features based on absolute value of correlation with response variable.