



UNIVERSITY OF EDINBURGH  
Business School

# Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

**Dr Xuefei Lu**

The University of Edinburgh Business School

A photograph of the University of Edinburgh Business School building, a modern structure with a glass and metal facade. The building is shown from a low angle, emphasizing its height. The sky is blue with some light clouds. A large blue rectangle is overlaid on the right side of the image, containing the text 'Feature Selection'.

# Feature Selection



UNIVERSITY OF EDINBURGH  
Business School

20 Buccleuch Place



# Feature Selection Techniques

Three types of FS techniques

## 1. **Filter methods**

To find a single measure that relates each independent variable to the dependent variable, which can be used to score the importance of the independent variable, e.g. Correlation. The outcome can be used for ranking.  
computationally fast and model-free

## 2. **Wrapper methods**

to generate a vast number of models with various subsets of independent variables to check which subset give the best performance.

Advantages: testing various combinations of predictors; interaction taken into account

Drawback: Model specific and computationally expensive.

## 3. **Embedded methods**

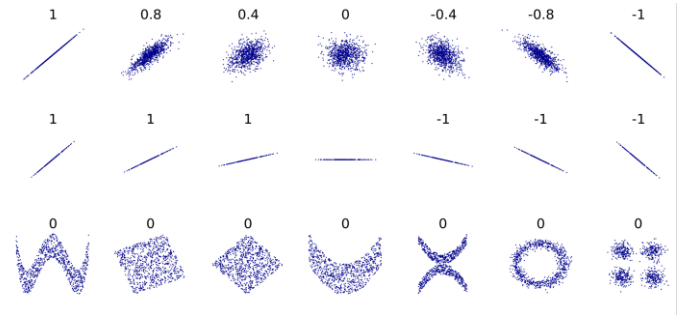
Some algorithms have built-in feature selection thanks to their way of modelling, e.g. Random Forest and Lasso regression.

# Filter Method – Correlation Coefficient

The most widely known is the correlation coefficient, more precisely, the **Pearson correlation coefficient**  $r$ :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- $r \in [-1, 1]$ : 1 (-1) for perfect positive (negative) correlation
- It captures the strength of the linear relationship between  $x$  and  $y$ .



# Filter Method – Rank correlation coefficient

The Spearman's Rank correlation coefficient  $R$ :

$$R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)},$$

where  $d_i = R(x_i) - R(y_i)$  is the difference in ranks between the rank of each observation  $R(x_i)$  and  $R(y_i)$

- Non-parametric
- Test for linear relationships
- Suitable for both ordinal and continuous variables



# Filter Method – $\chi^2$ -statistic

For discrete observations, another measure of association that is commonly used is the **Pearson's  $\chi^2$  -statistic**:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \sum_{i=1}^k \sum_{j=1}^l \frac{\left(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n}\right)^2}{\frac{n_{i+} \cdot n_{+j}}{n}},$$

with  $0 \leq \chi^2 \leq n \cdot (\min(k, l) - 1)$

$k$  and  $l$  the number of levels of the first and second variables, respectively,

where  $n_{ij}$  is the number of observations where value  $i$  of the first variable and value  $j$  of the second value are present in the dataset;

$\tilde{n}_{ij}$  denotes the expected absolute frequency when both variables are independent, with  $n_{i+}$  ( $n_{+j}$ ) the number of all observations with value  $i$  (value  $j$ ) as the first variable (second variable)

	2 <sup>nd</sup> var, value 1	2 <sup>nd</sup> var, value 2	Sum
1 <sup>st</sup> var, value 1	$n_{11}$	$n_{12}$	$n_{1+}$
1 <sup>st</sup> var value 2	$n_{21}$	$n_{22}$	$n_{2+}$
Sum	$n_{+1}$	$n_{+2}$	$n$

- $\chi^2$ -stat is symmetric, both variables can be reversed;
- Higher the value, stronger the relationship.
- The value ranges between 0 and the size of the number of observations times the minimum of the number of values either variable can take.

# Filter Method – Rank correlation coefficient



The Cramer's V-statistic normalizes the  $\chi^2$ -static:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(k, l) - 1)'}}$$

The closer this metric is to one, the stronger the relationship.



# Filter Method – Mutual Information

The mutual information is related to *information gain*. It is a measure of similarity between two different variables, based on the concept of entropy, i.e., the unstructuredness of data, from the area of information retrieval. **Entropy** of a variable can be defined as

$$H(X) = -\sum p(x) \log_2(p(x))$$

with  $p(x)$  the probability distribution function of  $X$ .

Example:

If  $X$  has 10 potential outcomes (e.g., 10 types of umbrellas) and they are equally likely, we have  $H(X) = -\sum_{i=1}^{10} \frac{1}{10} \log_2\left(\frac{1}{10}\right) = 3.32$ . In information theory, this would mean we need 3.32 bits to encode the variable.

If two outcomes with probabilities  $1/10$  and  $9/10$ ,  $-\left(\frac{1}{10} \cdot \log_2\left(\frac{1}{10}\right) + \frac{9}{10} \cdot \log_2\left(\frac{9}{10}\right)\right) = -0.47$ , as there is less unstructuredness in the data.



# Filter Method – Mutual Information



The **mutual information** is the reduction in unstructuredness when one variable is used in conjunction with the other given their joint probability, and is expressed as

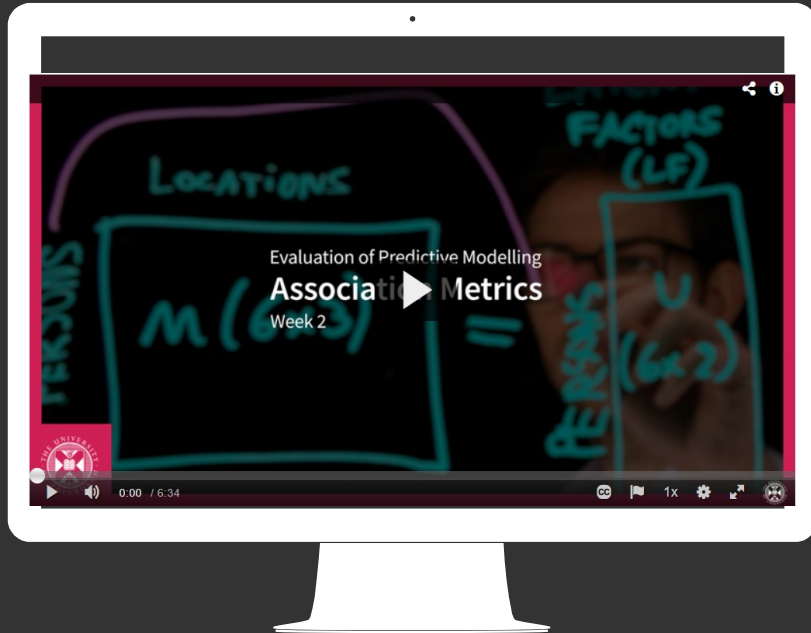
$$I(X, Y) = H(X) - H(X|Y) = \sum_{x,y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)$$


- Notice that the relation is symmetric, just like the Pearson correlation coefficient.
- $I(X, Y) \geq 0$
- $I(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent

## Association Metrics

Please watch the video via this [link](https://media.ed.ac.uk/media/Association+Metrics/1_4u5lu2dn/114521421)

[https://media.ed.ac.uk/media/Association+Metrics/1\\_4u5lu2dn/114521421](https://media.ed.ac.uk/media/Association+Metrics/1_4u5lu2dn/114521421)



- 
- Please study the following file:  
5 - code\_for\_correlation\_analysis.ipynb
  - Then try the following exercise  
6 - Activity\_5\_finding\_correlated\_variables.ipynb

In the above exercise, you will be using a new dataset containing information about different cars (cars.csv). Your goal is to find correlated variables and remove them.

## Activity: Correlation analysis & Finding correlated variables



UNIVERSITY OF EDINBURGH  
Business School