2. Datasets, visualization, and metrics

All the Machine Learning algorithms need data. The more the data, the better the learning. The data is required for the learning process of algorithms. A dataset represents all the information that can be processed. For example:

- One thousand images of faces, each one labeled with the person's name.
- Database of the sales department.
- A survey data.

Traditionally, the information is preprocessed to have a table with values, where the columns and rows represent the features and samples, respectively. A **feature**, also called a variable, describes a characteristic of the object. A **sample** is a concrete object. Some classic nomenclature is:

$N =$ Number of features
$M =$ Number of samples
$x_i =$ Feature i values (for all the samples)
$x^{(i)} =$ Sample i values (for all the features)
$y^{(i)} =$ Output or label of sample i
$x_j^{(i)} =$ Sample i, feature j

Example 1: House sales

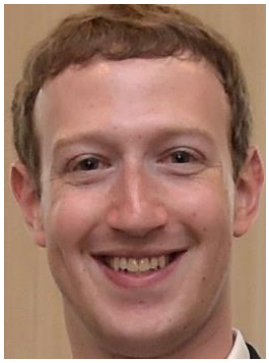
	x_1	x_2	x_3	x_4	y
	Size (feet2)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
$x^{(1)}$	2104	5	1	45	460
$x^{(2)}$	1416	3	2	40	232
$x^{(3)}$	1534	3	2	30	315
$x^{(4)}$	852	2	1	36	178
$x^{(5)}$	1035	3	2	20	250

$N = 4$, M = 5

Example 2: Face recognition


Mark Zuckenberg


Elon Musk


Mark Zuckenberg


Elon Musk

The images need to be preprocessed:

	x_1	x_2	x_3	...	y
	Eyes distance	Mouth length	Mouth nose distance	…	Person name
$x^{(1)}$	-	-	-		Mark Zuchkenberg
$x^{(2)}$	-	-	-		Elon Musk
$x^{(3)}$	-	-	-		Mark Zuchkenberg
$x^{(4)}$	-	-	-		Elon Musk

2.1 Features types

First classification

- **Numeric**: They can be integers or continuos. For example, age, price, temperature.
- **Ordinal**: The values of these features are not numbers, but their values have a specific order. For example: How do you feel today? Very unhappy, unhappy, neutral, happy, or very happy. Or scholarly: elementary school, high school, college, master, doctorate.
- **Categorical or nominal**: The values of these features represent categories. They cannot be ordered. For example, gender, hair color, birthplace.

Second classification

- **Continuous**: A feature is considered continuous when it has unlimited values. Here we can include the numeric variables.
- **Discrete**: A feature is considered discrete when it has a limited number of values. Regularly, the number of values is small. Here we can include ordinal and categorical variables.

2.2 Numeric features

Some metrics that can be helpful to understand numeric features are:

- Minimum
- Maximum
- Range: Maximum−minimum
- Average or mean: $\bar{x} = \frac{\sum_i x_i}{n}$
- Median: If we order the samples, the median is the value that lies just in the middle.
- Variance: population $\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$ sample $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$
- Standard deviation: population $\sigma = \sqrt{\sigma^2}$ sample $s = \sqrt{s^2}$. The standard deviation can be seen as the average of all the variations between the data and the average.
- Quartiles: It divides the number of data points into four parts. The data must be ordered from smallest to largest. The value of 25% of the data is below to first quartile (Q1). The second quartile (Q2) is the median of a data set; thus, 50% of the data lies below this point. The value of 75% of the data is below to first quartile (Q3).
- Interquartile range: It is useful to find outliers. $IQR = Q_3 - Q_1$. An outlier is a sample whose value x is

$$x < Q_1 - 1.5*IQR \text{ or } x > Q_3 + 1.5*IQR.$$

For measuring the dependency between two numeric features, we can use covariance and correlation.

COVARIANCE

It is a numeric value that measures the linear dependency between two numeric features.

Population $\sigma_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$ Sample $\sigma_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Interpretation (see image):

- A positive value indicates an increase in one variable results in an increase in the other variable.
- A negative value indicates an increase in one variable results in the opposite change in the other variable.
- A value close to zero indicates that there is no linear dependency. However, there could be a nonlinear dependency.

The covariance problem is that it does not have a limited range. It is difficult to know if a covariance value of 343 represents a strong or a weak dependency.

CORRELATION

Pearson's correlation coefficient, commonly called the correlation coefficient, is a numeric value that measures the linear relationship between two numeric features. The difference with the covariance is that it has a limited range between -1 and 1.

Population $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\text{Cov(x,y)}}{\text{StdDev(x) StdDev(y)}}$ Sample $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\text{Cov(x,y)}}{\text{StdDev(x) StdDev(y)}}$

Interpretation (see image):

- The sign represents the same as covariance.
- It is a number between -1 and 1
  - The closer to -1 or 1, the more the linear dependency.
  - Values close to 0 indicate that there is no linear dependency.

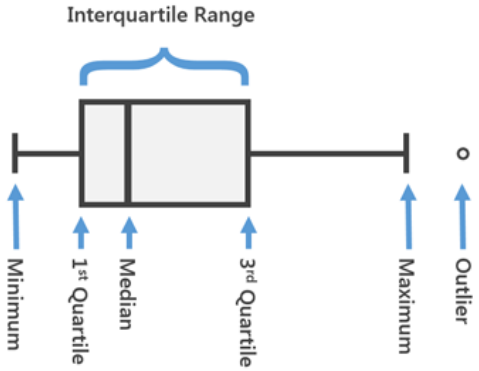
From my personal perspective, it is much better to use correlation than covariance to measure the linear dependency between numerical features.

VISUALIZATION

The graphs that we can use to visualize numeric features are:

- **Histogram** (1 feature). It is similar to a bar plot, but the difference is that a bar plot is used for categorical features and histograms for numeric features. It means the bars' ranges represent values, they are ordered, and we can change the width or number of bars. It is very useful for recognizing the distribution of the variables or if it has one or two modes. See examples.

- **Boxplot** (1 feature). It represents the distribution of a numerical feature based on quartiles. It is very useful to recognize outliers. The box width represents how sparse the samples are. See examples.



- **Scatter plot** (2 features). It is beneficial for understanding the relationship between two numerical features. See examples.

- **Bubble plot** (3 features). It is similar to a scatter plot, but using colors and sizes. It can represent more than two features. See examples.

2.3 Categorical features

Some metrics that can be helpful to understand categorical features are:

- Mode. It is the value that appears most often.

ENTROPY

The entropy, also called the entropy of the information or Shannon's entropy, measures the **surprise** or the **chaos** in the values of a variable. Formally, it measures the uniformity of the variable. It can be calculated as follows:

$$H(X) = -\sum_{x} P(x) \log_2 P(x)$$

Entropy: eyes color

Case 1: Brown: 1/3 Green: 1/3 Blue: 1/3

$$H(X) = -\left[\frac{1}{3}\log\left(\frac{1}{3}\right) + \frac{1}{3}\log\left(\frac{1}{3}\right) + \frac{1}{3}\log\left(\frac{1}{3}\right)\right] = 1.0986$$

Case 2: Brown: 90% Green: 3% Blue: 7%

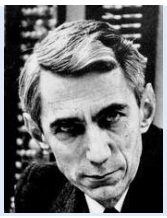
$$H(X) = -[.9\log(.9) + .03\log(.03) + .07\log(.07)] = 0.3863$$

Case 3: Brown: 100% Green: 0% Blue: 0%

$$H(X) = -[1\log(1) + 0\log(0) + 0\log(0)] = 0$$

The problem is that there is no defined a maximum value for the entropy. For each variable, the maximum value is:

$$H_{unif}(X) = -\log_2\left(\frac{1}{\#\ categories}\right)$$



Claude Shannon (1916-2001)

He was an American mathematician, electrical engineer, and cryptographer. He graduated from the University of Michigan and MIT. He is known as the father of information theory.

MUTUAL INFORMATION

Mutual Information measures how much one random variable tells us about another.

$$IM(X,Y) = \sum_{x}\sum_{y} P(x,y)\log_2\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

If X, Y are independent between them $\rightarrow \forall_{x,y} P(x,y) = P(x)P(y)$

Example:

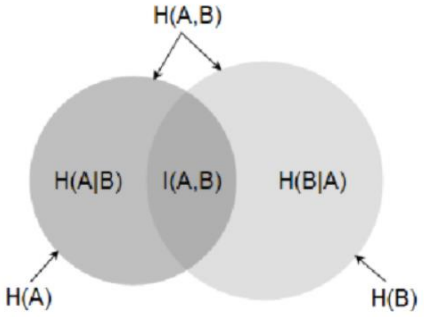
	1	2	3	4	5	6	7	8	9	10
Eyes color	Blue	Brown	Brown	Blue	Brown	Blue	Brown	Brown	Brown	Blue
Skin color	White	Brown	Brown	Brown	White	White	Brown	Brown	Brown	White
Career	IIA	IIA	IIA	IIA	IIA	IM	IM	IM	IM	IM

Entropy

Eyes color	frecuency	prob	log(prob)	p*log
blue	4	0.4	-0.529	-0.212
brown	6	0.6	-0.442	-0.265
				0.477
Skin Color	frecuency	prob	log(prob)	p*log
White	4	0.4	-0.529	-0.212
brown	6	0.6	-0.442	-0.265
				0.477
Career	frecuency	prob	log(prob)	p*log
IA	5	0.5	-0.5	-0.25
IM	5	0.5	-0.5	-0.25
				0.5

Mutual Information

MI(eyes,skin)	frecuency	prob XY	probX	probY	pXY/(pXpY)
blue,white	3	0.3	0.4	0.4	0.272
blue,brown	1	0.1	0.4	0.6	-0.126
brown,brown	5	0.5	0.6	0.6	0.237
brown,white	1	0.1	0.6	0.4	-0.126
					0.256
MI(skin,career)	frecuency	prob XY	probX	probY	pXY/(pXpY)
white,ia	2	0.2	0.4	0.5	0
white,im	2	0.2	0.4	0.5	0
brown,ia	3	0.3	0.6	0.5	0
brown,im	3	0.3	0.6	0.5	0
					0



$$IM(A,B) = H(A) - H(A|B)$$

$$IM(A,B) = H(\mathrm{B}) - H(A|B)$$

$$IM(A,B) = H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{A},\mathrm{B})$$

$$IM(A,B) = H(A,B) - H(A|B) - H(\mathrm{B}|\mathrm{A})$$

VISUALIZATION

The graphs that we can use to visualize categorical features are:

- **Bar plot**. Each bar length represents the frequency of each category. See examples.
- **Pie plot**. Each section represents the proportion of elements in each category. See examples. Be careful because if pie plots are represented with different effects, they can cause misunderstanding. See example.
- **Rectangular pie chart**. It is like a pie plot, but we use rectangles instead of a circle. See examples.