Introduction to Causal Inference

By Joaquín Mateos Barroso

*How to stop thinking about the **what**, and start thinking about the **why**.*

Subject: Introducción al Aprendizaje Automático

Professor: Sebastián Ventura Soto

Abstract

Data Science and Machine Learning are becoming a fundamental part of the human evolution. Any person you may find in the street has heard about "Artificial Intelligence", and it is unthinkable for a enterprise with a good size to do not organize and analyze its data. From economics to medicine, machine learning algorithms are showing that they are able to help in the daily life of billions of people.

Though, most of these tools, and specially the most modern ones (CNNs, LLMs, AGIs, ...) [1] [2], lack of an important feature. They are able to obtain, very efficiently, **what** they are "asked" to do, but they don't usually explain clearly **why** they are doing the elections.
If a doctor wants an algorithm that predicts whether a patient needs certain medication or not, he doesn't need just the output "yes" or "no". He needs a justification that explains the reason **why** the algorithm made certain prediction.

Humans are causal learners. We see events triggered by another ones, extract a rule of the kind "If A happens, then B is probably going to happen too", and extrapolate from there. It is much easier to explain to a doctor that his patient needs certain medicament "because he has these symptoms" than "because my neural network has learned these numbers as weights". Mathematics is a clear example of how far a human can get with logical causalities.

Causal inference is able to get to that point. Its tools are capable of not simply predicting what is going to happen, but of extracting implications and using them. They are able to see an event, understand it, and predict **what** may happen next, having a deep understanding of **why** it will happen.

CONTENTS

1. Brief introduction to probability theory

Probability theory is an area of mathematics with great historical relevance [3]. It is the common framework where many fields that try to model relationships between non-deterministic events in the real world work; among them, Statistical Inference and Causal Inference. However, modern science and engineering degrees tend to leave aside issues of definition and properties of random variables and other theoretical aspects to study statistical methods.

Therefore, in order to understand the starting point (statistical inference) and what are the advantages of approaching problems from a causal point of view, it is interesting to briefly review this important theory.

1.1. What is a probability?

The definition of probability is one of those fascinating questions that are wrapped around mathematics, and it is definitely a non trivial matter. Many books have wonderful chapters about this definition [4] [5] [6], and in this section there have been left some small pills on this subject.

Probability is a question about **uncertainty**. When we say that 'a coin has a probability of 50% of landing on its face', we do not mean that the future state of the coin is completely impossible to predict. A good physical modelization of the problem would able to predict exactly how the coin will fall and stay in the end.

Though, an affirmation about probability is an affirmation about **uncertainty**. It states that, according to the information we have, if we threw the coin, with unknown conditions, 'an infinite amount of times', the coin is expected to land on its face half of the times. The family of Central Limit Theorems [5] is able to make more robust this affirmation, but for the moment, we will focus on the problem of assigning certain number, between 0 and 1, to specific events, that are contained in a sample space.

Definition 1.1. *Given a measurable space* $(\Omega, \mathcal{F})$ *[7], a* ***probability*** *(or probability measure) is an application*

$$P : \mathcal{F} \to \mathbb{R}$$

such that

1. $P(A) \geq 0, \quad \forall A \in \mathcal{F}$
2. *For all numerable collection of events,* $\{A_n\} \subseteq \mathcal{F}$*, if* $A_i \cap A_j = \emptyset \;\; \forall i \neq j$*, then*

$$P(\cup_i A_i) = \sum_i P(A_i)$$

3. $P(\Omega)$ *= 1*

The only part of the definition that might not be immediate to assume is the second one, but it is easy to check that, if we have a set of events (e.g., the set of dices falling on faces $i \in \{i_1, ..., i_m\}$ in a dice of n faces) that are disjoint (the dice cannot fall at the same time in the face 2 and 3), then the probability of the union of these events (probability of obtaining any of the faces in the set) is the sum of the probabilities of each individual event $\big(P(\text{Face}_{i_1} \cup \cdots \cup \text{Face}_{i_m}) = \sum_{j=1}^{m} P(\text{Face}_{i_j})\big)$.

There are many very interesting properties of probability that can be checked in [4] [5] [6], but, in order to be concise, we will focus on those that are relevant for the task at hand.

1.2. Bayes Theorem

Bayes theorem is a classical predicate that relates conditional probabilities with more simple events. The motivation for the inclusion of it in this study is double:

- Bayes Theorem is one of the first attempts to study causality directly through probability.
- Naive Bayes classifier has been studied in the subject, and it is an interesting model, fundamentally based on Bayes Theorem.

Before facing this theorem, it is important to understand what is a conditional probability (very different from the interventions we will see in the context of causal inference):

Definition 1.2. *Given 2 events in a probability space* $(\Omega, \mathcal{F}, P)$*,* $A, B \in \mathcal{F}$*, the probability of* A *conditioned by* B *is*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

Knowing this basic definition, we can get to the desired theorem:

Theorem 1.1 (Bayes' Theorem)**.** *Given an event* A *in a probability space, and a numerable collection of disjoint by pairs events* $\{B_i\} \subseteq \mathcal{F}$*, with* $P(B_i) > 0$*, and* $\cap_i \{B_i\} = \Omega$*, then*

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)} = \frac{P(B_j)P(A|B_j)}{\sum_i P(B_i)P(A|B_i)}$$

Proof:

The first equality is trivial, checking that $P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(B_j)P(A|B_j)}{P(A)}$.
For the second one, the only step done is the application of the total probability formula:

$$P(A) = P(A \cap \Omega) = P\big(A \cap (\cup_i \{B_i\})\big) = P\big(\cup_i (A \cap B_i)\big)$$

So, applying the second property of the probability definition:

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(B_i)P(A|B_i) \quad \blacksquare$$

In simple terms, this theorem states that, if we know the probabilities that the occurrence of certain events $\{B_i\}$ 'implies' on a event A, then we can obtain the probability that the event A 'implies' on each of the events B_i.

When trying to extract conclusions about similar, but different questions, such as 'What would be the probability of obtaining certain medical result, B_j, if on a patient over which certain medicament A was applied, we had applied the medicament A'?' (**counterfactual**; [8]), we might think that obtaining the probability $P(B_j|A')$ is a good idea. Though, as Judea Pearl explains magnificently in his book [9], this is not the best idea.

When calculating $P(B_j|A')$, we are considering many situations that have not necessarily happened. When using, for example, the Bayes' Theorem to obtain it, we are using all the probabilities $P(A'|B_i)$, which are obtained in cases in which a doctor decided that the best medication for certain patient was A'.

This affirmation is much clearer with an example: If B_1 is the event of of the patient's survival, and B_2 is its complementary ($B_2 = B_1^c$), then, if we have applied certain simple medicament, A, to a patient that had a slight cough, and he has survived, we might think about calculating the probability of survival when applying another, also simple, medicament, A', that is only applied on patients that have a terminal illness.
We would obtain that $P(B_1|A') = 0$, and conclude that the patient, who arrived to the hospital with a slight cough, is going to die with a probability of 100% if we give to him a simple medicament.

This case seems ridiculous, but when using inferencist and machine learning algorithms, we are constantly making similar assumptions, and these examples were one of the main motivations for the growth in recent years of **Causal Inference**.

1.3. Random Variables

The last theme to treat before entering in Causal Inference is, probably, one of the biggest tools that made statistics and data science what it is nowadays: Random Variables.
In simple terms, they are the tool through which we are able to study the behavior, in probabilistic terms, of real (or any subset of $\mathbb{R}^k$) encoding problems, and combinations of various of these problems.

Definition 1.3 (Random Variable). *A 1-dimensional,* ***random variable*** *in a probability space $(\Omega, \mathcal{F}, P)$ is any function*

$$X : \Omega \to \mathbb{R}$$

that is measurable [1]*.*

Every absolutely continuous random variable is associated with a ***probability***

[1] I.e., such that $\vec{X}^{-1}(B) \in \mathcal{F}$, for all $B \in \mathbb{B}$ (Borel σ-algebra [7]).

***distribution**:*

$$P_X(B) := P\{w \in \Omega | X(w) \in B\}$$

*and with the corresponding **distribution function**:*

$$F_X(x) := P\{w \in \Omega | X(w) < x\}$$

and its associated density/probability function $f_X : \Omega \to \mathbb{R}$.

If $\Omega = \mathbb{R}$, $\mathcal{F} = \mathbb{B}$ and the variable is absolutely continuous[6], then we have:

$$F_X(x) = \int_{-\infty}^{x} f(t)dt$$

This definition is usually joined to the classical notation

$$\{w \in \Omega | X(w) \in B\} := \{X \in B\}$$

which allows obtaining probabilities in a more understandable way.

With this notation, given a random variable X with a distribution $\mathcal{N}(\mu, \sigma^2)$ $\big($notated $X \sim \mathcal{N}(\mu, \sigma^2)\big)$, the probability of X being lower than certain value $x \in \mathbb{R}$ is:

$$F_X(x) = P\{X < x\} = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

It is easy to proof that

$$P(\Omega) = F_X(\infty) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

using the multivariable, integration variable change [10], by applying the Jacobian of $(x, y) = (r\cos\theta, r\sin\theta)$ (details are left as an exercise for the reader).

This means that, using the properties in Definition 1.1 and 1.3, the normal distribution defines a correct random variable over a sample space Ω.[2]

[2] This sample space cannot be any set. In particular, it must contain, over a surjective application, to $\mathbb{R}$.

1.4. Likelihood Ratio

Lastly, the Likelihood Ratio[3] is a very powerful tool, that in our case, will needed to construct the conditional independence test between random variables, in section 3.2. This important tool is very intuitively interpretable; we are going to contrast how 'litle' is the 'probability of the null hypotheses', compared to the 'probability of the whole space'. The main theorem about Likelihood Ratios is[4]:

Theorem 1.2. *Let X be a random variable over a sample space Ω, with density $f_\theta(x)$, whose parameter θ is in a space $\Theta \subseteq \mathbb{R}^k$, over which we want to contrast the null hypothesis $H_0 : \theta \in \Theta_0 = \{\theta \in \Theta | \theta_i = g_i(\omega_1, ..., \omega_q), with (\omega_1, ..., \omega_q) \in \Omega\}$, being Θ_0 an arbitrary subset of Θ. If we denote as likelihood ratio of a simple random sample, $(X_1, ..., X_n)$,*

$$\Lambda := \frac{\max_{\theta \in \Theta_0} f_\theta(x_1, ..., x_n)}{\max_{\theta \in \Theta} f_\theta(x_1, ..., x_n)}$$

Then, if the actual parameter is $\theta_0 \in \Theta_0$, we have the following distribution limit:

$$-2 \log \Lambda(X_1, ..., X_n) \xrightarrow{d_{\theta_0}} \chi^2_{k-q}$$

This means that if we are able to obtain the Maximum Likelihood estimator[5] of θ, then we will be able to find the statistical distribution that a function of our sample space, supposing $\theta \in \Theta_0$, follows. This, as will be seen in 3.2, is easily usable for the construction of statistical tests[6].

[3] In Spain, this ratio is named after 'Razón de Verosimilitud', and the author personally thinks this is a more appropriate term.

[4] An elegant proof of this theorem can be consulted at [11], page 418.

[5] The estimator of θ, T, that is defined by the rule $f_T(x_1, ..., x_n) = \max_{\theta \in \Theta} f_\theta(X_1, ..., X_n)$ is, by itself, an important measure of our space, and is named after 'Maximum Likelihood estimator' [12].

[6] In particular, this kind of tests are usually known as 'Pearson's χ^2 tests'.

2. Motivation and basic definitions

2.1. Motivation of Causal Inference

Classical Statistical Inference [12] studies a probable theoretical distribution, using as basis a sample of this distribution, estimating parameters and testing hypotheses. Its techniques do not necessarily rely on causal relations; the covariance between random variables, for example, is a symmetrical statistic.

On the other hand, **Causal inference** aims to understand and answer questions like ’Does treatment A cause outcome B?’ (causal discovery), or ’What would have happened to the variable A if B had happened?’ (counterfactuals).

It is important to note that these questions are significantly different from traditional questions such as ’Is B statistically dependent (non-independent) of A?’ (what would just indicate that, in presence of certain value of A, B has a different distribution; $f_{A,B} \neq f_A \cdot f_B \Leftrightarrow f_B \neq f_{B|A}$) or ’What is the best estimator of A knowing that B has happened?’ (what is obtained with Bayes estimators via a conditioned distribution [12], $\pi(a|B = b) = \frac{\pi(a) f_a(b)}{\int_{W_A} f_t(b)\pi(t)dt}$, and inside the integral considers different events that are more probable due to the fact $B = b$).

These tools are very useful for prediction, and modern machine learning uses them constantly, but causal inference allows extracting causal implications that might be useful for both a better comprehension of the data and relationships between variables, and obtaining, in some cases, even better estimators than the ones obtained via traditional methods, such as Bayes or minimal risk.

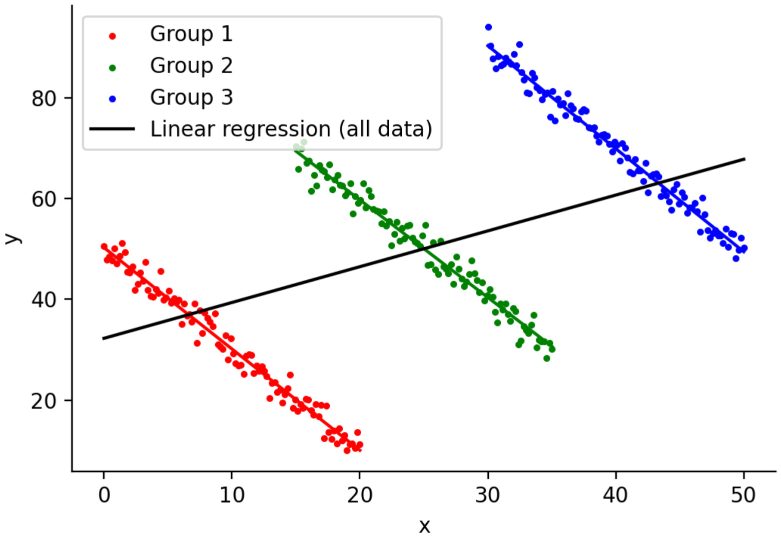

Figure 2.1: Example of Simpsons' paradox

A classical illustration of the saying 'correlation does not imply causality' is the **Simpson's paradox**, which states that, for example, we could obtain a positive Pearson correlation between smoking and getting good grades in a difficult math test. Though, this correlation might be caused by people with higher age smoking more and also getting better grades. In Figure 2.1 it can be seen a representation of this happening, where the x axis could be the quantity of cigarettes taken per month, and y axis the score, over 100 in the math test. If the groups are ordered according to their age (Group 1 has lowest age and 3 the highest), then we can see that the colored regressions make much more sense than the black line.

While statistical inference could use the black line, and it would be very useful for the task at hand, causal inference tries to find these causalities rather than correlations.

2.2. Pearson vs Spearman correlation

In this section there will be explained the differences between two classical and powerful correlation measures, or coefficients. Despite Pearson correlation is used in Causal Inference and Spearman Correlation has great implications in Statistical Inference, the first has a clear inferencist justification, and the second is constructed taking as basis a causal assumption, so it might be interesting, as first insight into the Causal Inference field, to contrast both views.

Before doing this comparison, we need a few more probability tools.

2.2.1. Expectation and Covariance

When we speak about the expectation of a Random Variable, we are speaking about what is the "outcome that will happen on average". This does not mean that it is the most probable outcome (mode); in fact, it doesn't even have to be a possible outcome.

Definition 2.1 (Expectation). *Given an absolutely continuous random variable, X, with density $f_X(x)$, its* ***Expectation*** *is*

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f_X(x) dx$$

On the other hand, it X is a discrete random variable, with distribution $p_X\{X = i\}$, its ***Expectation*** *is*

$$\mathbb{E}[X] := \sum_{x \in Image(X)} x \cdot p_X\{X = x\}$$

The ***variance*** *of any random variable, X, is the expectation of the square of the difference between X and $\mathbb{E}[X]$:*

$$VAR(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

The ***covariance*** *between 2 random variables X, Y is the expectation of the square of the difference between XY and $\mathbb{E}[XY]$:*

$$COVAR(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

This means that the covariance between X and Y is the expectation of the product of the differences of these random variables with their expectations.
The covariance definition could seem a little arbitrary if we didn't check the proof of the following

Theorem 2.1. *2 random variables X, Y, have null covariance iff,*

$$VAR(X + Y) = VAR(X) + VAR(Y)$$

Proof:

($\Rightarrow$) If we denote $\mu_X = \mathbb{E}[X], \mu_Y = \mathbb{E}[Y]$, then it is $\mathbb{E}[X+Y] = \mu_X + \mu_Y$ and we have:

$$\text{VAR}(X+Y) = \mathbb{E}[(X+Y-\mu_X-\mu_Y)^2] =$$

$$= \mathbb{E}[(X-\mu_X)^2] + \mathbb{E}[(Y-\mu_Y)^2] + 2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] =$$

$$= \mathbb{E}[(X-\mu_X)^2] + \mathbb{E}[(Y-\mu_Y)^2]$$

($\Leftarrow$) It is important to note that we have been able to perform the last equality just due to the supposition that $2\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = 2\text{COVAR}(X,Y) = 0$. ■

This means that the variance of the random variable $X + Y$[1] is explained by the addition of the variances of X and Y iff their covariance is null.
Now we are able to understand the Pearson Correlation.

2.2.2. Pearson Correlation

As seen in previous section, the covariance between random variables indicates how well the covariance of the sum of these random variables is explained by the particular variances of each random variable.

This idea was used by Karl Pearson, together with a normalization of the measure, to define a coefficient that describes how well a random variable Y can be defined by a linear function of X (and the reciprocal):

Definition 2.2. *Given 2 random variables X, Y, the **Pearson correlation coefficient** is the normalization of their covariance to the product of the standard deviation of each particular variable, i.e.,*

$$\rho_{X,Y} = CORR(X,Y) := \frac{COV(X,Y)}{std(X)std(Y)}$$

2.2.3. Spearman Correlation

Spearman correlation has a more causal motivation.
It tries to find monotonous relations between random variables by measuring if the

[1] Or, equivalently, the variance given by any random variable $aX + bY + c$, with $a, b, c \in \mathbb{R}$

increase in one of them implies the increase of the other one.

For any person who has studied basic real analysis this idea must sound familiar, since a function $f : D \subseteq \mathbb{R} \to \mathbb{R}$ is said to be monotonous when $\forall x, y \in D$ s.t. $x \leq y$ it is true that $f(x) \leq f(y)$ (or $\geq$).

Definition 2.3. *In order to compute this idea for a simple random sample*[2] $(X_1, ..., X_n)$, $(Y_1, ..., Y_n)$ *of 2 random variables* X, Y*, the values of the samplings are sorted, and their indexes in this new sorting are named the* ***ranks*** *of these random variables;* $R(X_i), R(Y_i)$*. That way, the* ***Spearman correlation coefficient*** *is the correlation between the ranks of* X *and* Y*:*

$$r_s = \rho_{R(X),R(Y)} = \frac{COV(R(X), R(Y))}{std(R(X)std(R(Y)))}$$

This means that the Spearman coefficient measures how well can we sort Y if we know how X is sorted. In other words, it measures whether a increase on X *implies* a increase on Y.

2.2.4. Practical comparison

These correlations are very easy to calculate using a high level programming language, such as python.

The Pearson correlation code would be:

```python
def pearsonCorrelation(x, y):
    n = len(x)
    mean_x = sum(x) / n
    mean_y = sum(y) / n
    std_x = (sum([(i - mean_x)**2 for i in x]) / n)**0.5
    std_y = (sum([(i - mean_y)**2 for i in y]) / n)**0.5
    return sum([(x - mean_x) * (y - mean_y) for x, y in zip(x,y)]) /
                    (n * std_x * std_y)
```

And, having this, the one for the Spearman correlation would be even simpler:

[2] In order to simplify the problem, along this section we will be considering just simple random samplings, despite most of the shown results extrapolate to more general samplings. The whole definition of Spearman correlation has been simplified so that, despite being losing certain formality, it is easier to get to the point.
The properties of these samplings can be checked at [12].

```
def spearmanCorrelation(x, y):
    R_x = [sorted(x).index(i) for i in x]
    R_y = [sorted(y).index(i) for i in y]
    return pearsonCorrelation(R_x, R_y)
```

Now it would be correct to perform certain comparison between them to see what kind of relation is each one of them able to find.

The first case tried is the data obtained from a simple linear relation with some noise. Particularly, the data shown in Figure 2.2, extracted from the file *Linear relation.py*. The obtained results are:

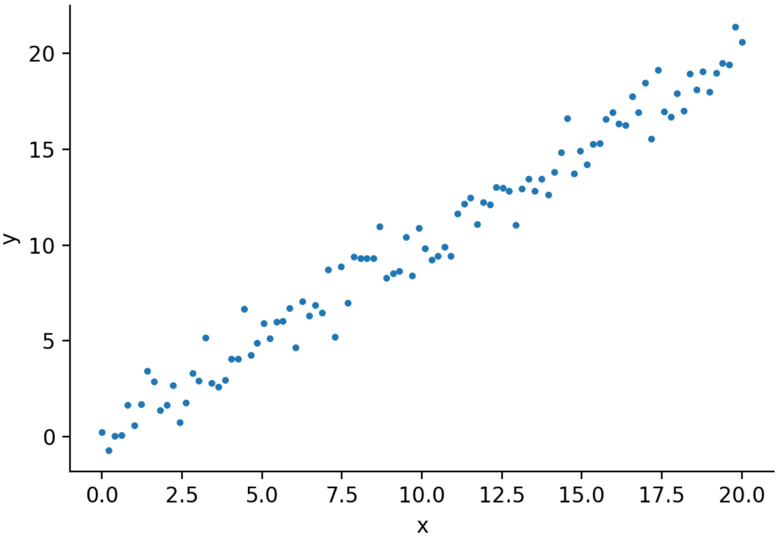

Figure 2.2: Basic linear relation with some noise between 2 variables.

```
Pearson correlation -> 0.9865806673673898
Spearman correlation -> 0.9872427242724272
```

Without showing significant differences, these results simply are a practical demonstration of the already known capacity of both Pearson and Spearman coefficients to capt linear relations. The first one is able because it is its main focus, and the latter because the linear relations are a subset of the monotonous ones.

The second case is the one shown in Figure 2.3, created in *Non-linear relation*, which clearly is non-linear. Though, if we wanted to check whether X and Y are related, we would expect to get a very high result, because Y doesn't seem to be very difficult to predict from X. Though, the results were:

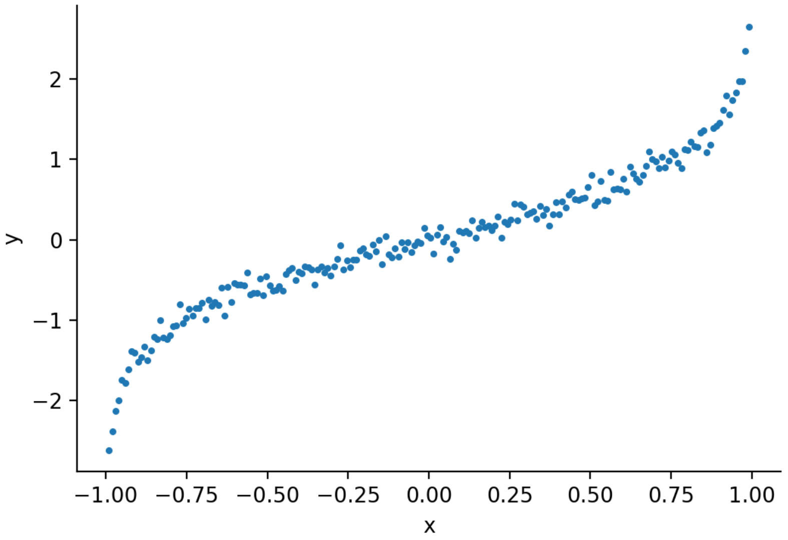

Figure 2.3: Non-linear relation between x and y with certain noise. In particular, the used function was $y = \text{arctanh}(x)$.

```
Pearson correlation -> 0.960780650717893
Spearman correlation -> 0.9921743043576091
```

2.2.5. Conclusion

Previous results show exactly what we were expecting.
Pearson correlation finds certain linearity between the variables, but the Spearman correlation, due to being able to capt any kind of monotonous relation, shows us that, effectively, the variables have a very high correlation.

This simply is an example of how, by means of causal assumptions and ideas, we can use classical tools in order to create a new concept that is able to obtain more information. In particular, it is able to obtain information of the kind *cause implies effect* (higher values in X imply higher values in Y); $C \rightarrow E$, using the notation that will be introduced in section 2.4.

2.3. Languages for causality

Nowadays, there are three different useful languages for representing causal relations, that are advantageous for different purposes [13]:

1. Graphs: Easy to visualise the causal assumptions; Difficult for statistical inference because model is nonparametric.

2. Structural equations: Bridge between graphs and counterfactuals; Easy to operationalise; Danger to be confused with regressions

3. Counterfactuals: Easy to incorporate additional assumptions; Elucidation of the meaning of statistical inference; not as convenient if system is complex.

Sometimes, some of them can be used together in order to create a more "powerful" model. We are going to study one of these cases (a model that I personally find really powerful and intuitive) in the following section.

2.4. Structural Causal Model

We will begin explaining the simplest, non-trivial, relationship between 2 variables, C and E; cause and effect[14]:

Definition 2.4 (Basic Structural Causal Model)**.** *Given two random variables, C, E, an Structural Causal Model (SCM), $\mathbb{M}$ with graph $C \to E$ consists of two assignments:*

1. $C := N_C$

2. $E := f_E(C, N_E)$

where $N_C \perp N_E$ are two independent random variables.

We say C is the **cause** and E the **effect**.
In this particular case, C is a **direct cause** of E, and $C \to E$ is a **causal graph**.
This definition is directly related, and can be better understanded, with the following one:

Definition 2.5. *An **intervention** in a SCM, $\mathbb{M}$ is the modification of one of the two definition assignments.*

*A **hard intervention** is the replacement of the second assignment so that E is independent of C. One example would be $do(E := k)$, for $k \in \mathbb{R}$, and the new C distribution (that in this case would remain unchanged) would be $P_C^{do(E:=3)}$, with density $p^{do(E:=3)}(c)$.*

*A **soft intervention** is an intervention on an assignment that keeps a function dependence on C over E, e.g., $do(E := g_E(C) + N'_E)$.*

The following example helps clarifying the difference between this model and the use of the conditional probability:

- Example: Cause-effect intervention.

  Given the SCM $\mathbb{M} \equiv \{C := N_C, E := 4 \cdot C + N_E\}$, being $N_C, N_E \sim \mathcal{N}(0, 1)$, we have:

  $$P_E^{\mathbb{M}} = 4N_C + N_E = \mathcal{N}(0, 5)$$

  $$P_E^{do(C:=2)} = 4 \cdot 2 + N_E = \mathcal{N}(8, 1) = P_{E|C=2}$$

  $$P_C^{do(E:=4)} = N_C = P_C^{\mathbb{M}} \neq P_{C|E=4}^{\mathbb{M}}$$

  Interventions on C change the distribution of E, but interventions on E do not have effect on C, despite C and E may be dependent, what implies that $P_C^{\mathbb{C}} \neq P_{C|E=2}^{\mathbb{C}}$.

  This assymetry can also be formulated from the independence of C and E when intervening with $do(E := N'_E)$, but remaining dependent when intervening with $do(C := N'_C)$

Despite this causal model is very illustrative, and simple to understand, there is a more general definition that allows constructing more powerful models[8]:

Definition 2.6 (Structural Causal Model)**.** *An Structural Causal Model, that might be associated with any kind of DAG (Directed Acyclic Graph) between variables, is defined by three numerable sets:*

- *Exogenous variables, U, that are obtained from external methods.*
- *Endogenous variables, V, that are obtained from a combination of variables from both U and V.*

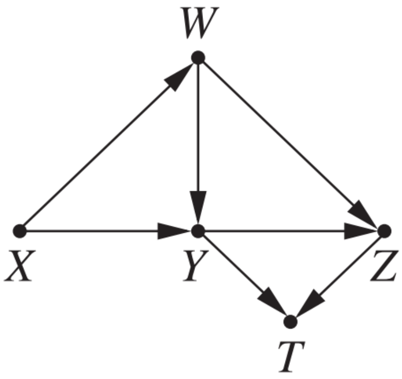

Figure 2.4: SCM basic example

- *Causal relations, **F**, that is a set of* $\#V$ *functions,* $f_i : U \times V \backslash \{E_i\} \to V$*, one for each variable,* $E_i \in V$*.*

That way, each $f_i \in \mathbf{F}$ would be a generalization of the second equation in Definition 1.1, E_i the effect, and $U \times V \backslash \{E_i\}$ the cause. Though, now a variable $C \in U \cup V \backslash \{E_i\}$ is considered to be a cause of E_i iff there does not exist a function, f_i', s.t. $f_i(C, X_1, X_2, ...) = f_i'(X_1, X_2 ...)$.

These models are usually represented via a DAG, where there are edges that part from causes and end in effects. E.g., in Figure 2.4, an SCM is represented, and we can see that $U = \{X\}$, because it is the only variable without cause, $V = \{Y, W, Z, T\}$, and the causes of Z would be W, Y.

2.5. Common cause principle

One of the main intuitions that make causal inference an important field of study is [15]:

Definition 2.7 (Richenbach's common cause principle)**.** *A probability* P *is said to accomplish the Richenbach's common cause principle if for any two random variables* X*,* Y *that are statistically dependent* $(X \not\perp_P Y)$ *there exists a random variable* Z *that "causally influences both" and such that* $(X \perp_P Y | Z)$*.*

From now and on we will suppose that all probabilities accomplish Richenbach's common cause principle.

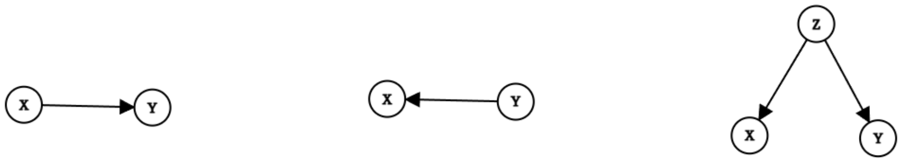

Figure 2.5: 3 possible DAG in the Common Cause Principle hypotheses.

It is easy to obtain an intuitive idea of the power this idea has.
Knowing that [6] $X \perp Y \Rightarrow \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, using the counter-reciprocal, we obtain that always that 2 variables have a Pearson correlation different to 0, they are dependent; $X \not\perp Y$, so the common cause principle allows us stating that there are 3 main possibilities, shown in Figure 2.5:

1. X is the cause of Y (in case $Z = X$).

2. Y is the cause of X (in case $Z = Y$).

3. The relation between X and Y is due to a third variable.

e.g., in the Simpsons' Paradox explained previously, we would be in the third case, and the variable Z would be the age of the person taking the test.

This important result is the theoretical fundament of the PC algorithm for causal discovery that will be explained in the following chapter.

3. Application for Causal Discovery

3.1. PC Algorithm implementation

Once the basic definitions and the important needed theorem are presented, we can start with the objective of the current work: Causal Discovery.

Causal discovery is aimed at, having a dataset in which we infer there must be relationships, and making a series of assumptions, to obtain and model these relationships in a SCM, $\mathbb{M}$, either just focusing on the DAG, or on the functions $f_i \in \mathbf{F}$.

These are the necessary assumptions to perform the PC algorithm [16]:

- (Causal Markov Condition) An SCM satisfies the Causal Markov Condition iff all the variables are independent of their non-descents in the distribution conditioned by the knowledge of all its parents.

- (Faithfulness Condition) An SCM satisfies the Faithfulness Condition iff all variables are dependent unless entailed by the Causal Markov Condition.

- (Causal sufficiency) For every pair of variables which have their observed values in a given dataset, all their common causes also have observations in the dataset.

And, based on these conditions, the following algorithm, named after its authors, Peter and Clark (PC), emerges very naturally:

1. Make a node for each observed variable

2. Start with all of them being connected to each other.

3. Eliminate as many edges as possible using conditional independence tests. More specifically, remove edges $X - Y$ if X is independent of Y given a conditioning set Z. Step 3 is a repetitive procedure, starting with S as the empty set Z= and increasing its size (cardinality) by 1 for every iteration.

4. Establish (causal) directions for each remaining edge using colliders, the assumption that there are no cycles, and any other assumptions you can make use of, such as time order.

Perhaps the only non-obvious step is the third one, but it is easy to understand, at least intuitively, if we note that, due to the common cause principle, a pair of variables X and Y with non-trivial correlation, has a dependence of the style $X \to Y$ or $Y \to X$ iff there does not exist a third, different variable, Z, that influences both and $(X \perp Y|Z)$. Due to the causal sufficiency assumption, Z must be a combination of variables in the dataset (particularly, if $Z = \{\}$, then, in the first step, we will find that (X, Y) are independent, ergo there is no relation to look for. This case does also solve the problem of trivial correlations between X and Y).

Now, speaking about implementation, first and second steps are trivial, and for the application of the third one, we can obtain easily the following pseudocode, that simply iterates over all possible tuples (X, Y, Z), where X ans Y are nodes, ans Z is a set of nodes that might 'influence' (is connected to) both X and Y.

```
depth = 0
repeat
    for each ordered pair of adjacent vertices X and Y in G do
            if (|adj(X, G)\{Y}| >= depth) then
                for each subset Z in adj(X, G)\{Y } and |Z| = depth do
                    if Independent(X, Y |Z) then
                        Remove edge between X and Y
                        Save Z as the separating set of (X, Y)
                        Update G
                        break
    depth = depth + 1
until |adj(X, G)\{Y }| < depth for every pair of adjacent vertices in G
```

Where I(X, Y| Z) is an independence test for X, Y conditioned by Z. We have to decide one.

3.2. Independence test

3.2.1. χ^2 Independence test

The study of different independence test is a very extensive and formal theme, so here there will be explained a simple one; the χ^2 independence test[1].

This test requires the discretization of our sample space, which usually is $\Omega = \mathbb{R}^2$, in k sets, $\{A_1, ..., A_k\}$ for the first variable, X, and r sets, $B_1, ..., B_r$ for the second variable, Y.

Once a simple sampling is performed over our space, its information can be compiled in a *contingency table*, which stores, in the position (i, j), the number of instances from the sampling, that, being (x, y), comply with $x \in A_i$ and $y \in B_j$.

	B_1	B_2	$\cdots$	B_r	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1\cdot}$
A_2	n_{21}	n_2	$\cdots$	n_{2r}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{kr}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot r}$	n

Table 3.1: Representation of a general contingency table.

That way, if we denote $p_{ij} := P\{(x, y) \in \Omega | x \in A_i, y \in B_j\}$, the probability of obtaining the contingency table in Table 3.1 can be easily obtained via a multinomial distribution, and it is the following:

$$n! \prod_{i=1}^{k} \prod_{j=1}^{r} \frac{1}{n_{ij}!} p_{ij}^{n_{ij}}$$

We will use the likelihood ratio test. For its calculus we will need the maximum of this function, so let's calculate it.

Our only restriction is $\sum_{i,j} p_{ij} = 1$, so it is easy to calculate this maximum via the Lagrange multipliers. In order to apply this method, we construct

$$h(\vec{p}) = n! \prod_{i,j} \frac{1}{n_{ij}!} p_{ij}^{n_{ij}} + \lambda(\sum_{i,j} p_{ij} - 1)$$

[1] It is a non-parametric test, so, in general, we will not have problems with the needed hypotheses. These hypotheses can be consulted at [12], page 395.

and find the points where the gradient of h is null:

$$\vec{\nabla} h(\vec{p}) = 0 \Leftrightarrow n!\Big(\prod_{(i,j)\neq(r,s)} \frac{p_{ij}^{n_{ij}}}{n_{ij}!}\Big)\frac{p_{rs}^{n_{rs}-1}}{n_{rs}-1} - 1 = 0, \quad \forall r, s$$

That way, $\forall r, r', s, s'$ s.t. $r \neq r'$ and $s \neq s'$, it is true that

$$\frac{p_{r's'}^{n_{r's'}}}{n_{r's'}!} \cdot \frac{p_{rs}^{n_{rs}-1}}{(n_{rs}-1)!} = \frac{p_{rs}^{n_{rs}}}{n_{rs}!} \cdot \frac{p_{r's'}^{n_{r's'}-1}}{(n_{r's'}-1)!}$$

from where we can extract that

$$p_{r's'} = \frac{n_{r's'}}{n_{rs}} p_{rs}$$

what implies

$$1 = \sum_{r's'} p_{r's'} = \sum_{r's'} \frac{n_{r's'}}{n_{rs}} p_{rs} = \frac{p_{rs}}{n_{rs}} \sum_{r's'} n_{r's'}$$

Ergo we can finally conclude that the probabilities that maximize the probability function of our table is

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

On the other hand, our null hypotheses is

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j} \qquad \text{for every } i, j$$

And, under H_0, the probability of obtaining the previous contingency table is

$$n! \prod_{i=1}^{k} \prod_{j=1}^{r} \frac{1}{n_{ij}!} p_{i\cdot} p_{\cdot j}^{n_{ij}} = n! \frac{1}{\prod_{i,j} n_{ij}!} \prod_{i=1}^{k} p_{i\cdot} \prod_{j=1}^{r} p_{\cdot j}^{n_{ij}}$$

And following a similar reasoning we obtain that the maximum probability function of our table is reached at the points

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \ \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

.

That way, using the notation in 1.4, the likelihood ratio for the contrast of H_0 is

$$\Lambda := \frac{\max_{\theta \in \Theta_0} f_\theta(x_1, ..., x_n)}{\max_{\theta \in \Theta} f_\theta(x_1, ..., x_n)} = \frac{\prod_{i=1}^{k}(n_{i\cdot}/n)^{n_{i\cdot}} \prod_{j=1}^{r}(n_{\cdot j}/n)^{n_{\cdot j}}}{\prod_{i,j}(n_{ij}/n)^{n_{ij}}}$$

Ergo, using the theorem in 1.4,

$$D = -2 \log \Lambda = 2 \sum_{i,j} n_{ij} \Big(\log \hat{p}_{ij} - \log(\hat{p}_{i\cdot} \hat{p}_{\cdot j}) \Big) \xrightarrow{d_{\theta_0}} \chi^2_{k-q}$$

That way we have concluded the theoretical aspects of the χ^2 independence test.
Though, this last statistic might seem strange for people who are unfamiliar with statistics. This is because the traditional statistic, discovered by Pearson, as explained in [12], page 432, is another one that is very similar to this one obtained. Despite D shows to have all the features we would ask to a statistic to be able to define a critical set, most of basic books stick to the traditional Pearson statistic:

$$D' = \sum_{i,j} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n}$$

3.2.2. Use of χ^2 for conditional independence test

Conditional independence tests are a bit different to traditional independence test. In simple terms, these tests have a null Hypothesis H_0; this Hypotheses is defined by a set that contains all the cases in which the tested variables X, Y are, conditionally to Z, non-independent, i.e., $P(X, Y|Z) \neq P(X|Z) \cdot P(Y|Z)$, or, equivalently, $P(X|Z, Y) = P(X|Z)$.

One simple, yet effective, way to check this, when Z is discrete, is to test whether it is true that

$$P(X, Y|\{Z = z\}) = P(X|\{Z = z\}) \cdot P(Y|\{Z = z\})$$

for every possible value of Z.
That way we can use the same χ^2 test that we were using before, but comparing it with all the possible values of Z. The correct, statistical, way to do is by applying to the significance level, α, a Bonferroni correction. This has been done in our implementation

of *PC* with the following pseudocode:

```
p_values = []
for z in self.__allValuesOf(Z, data):
    conditionalData = data.loc[(data[Z]==z).all(axis=1)]

    res = chi2_contingency(pd.crosstab(conditionalData[X],
                                       conditionalData[Y]))
    p_values.append(res[1])

# Bonferroni correction
adjusted_alpha = self.alpha / len(p_values)
```

3.3. External knowledge

Usually, we do not have a dataset without any information. In the common case there are causal restrictions; if, for example, an event happened earlier than another, it is evident that the later can not be a cause of the first. Our PC algorithm will cover 2 of these external knowledge cases:

- **Endogenous** variables, i.e., those variables that are not the effect of any other. An example would be the age of a person, when studying some of his/her characteristics. Computationally, the only addition is changing the direction of all edges related with these variables, so that they start from them, and end in the adjacent variable.

- **Exogenous** variables, i.e., those variables that are not the cause of any other. An example would be the effect caused by a sequence of actions, that happened temporally later. Computationally, the only addition is changing the direction of all edges related with these variables, so that they end in them, and start from the adjacent variable.

4. Bayesian Classifiers

Given that our PC algorithm is able to obtain causal relations, it might be interesting to use those relations via the Bayes Theorem, which allows extracting information somehow similar to an intervention.

There are more advanced models that allow using more rigorous interventions, but, for simplicity, here we will introduce 2 bayesian estimators; one based on the obtained causal relations and the other without considering them, in order to see if the information extracted in the causal discovery was useful.

4.1. Bayesian Networks

We will be using [17] as reference for the current chapter, and, since Bayesian Networks are not the finality of this study, the theoretical issues will be left more apart than in the previous sections.

Once we have obtained a DAG that explains the relations between our problem's variables, we need to use this information for some purpose. Usually, this purpose is related with obtaining a prediction over certain variable. To do so, one of the most basic models (that, though, might get really complex with addition of features) could be the **Bayesian Network**, that allows creating a full SCM.

This is a model that, taking as basis a DAG, uses the implications and causalities between variables to obtain the probabilities of possible outputs according to the values of the inputs.

Definition 4.1. *A Bayesian Network is a pair $\mathcal{B} = (G, \Theta)$, where G is the DAG, whose nodes $X_1, ..., X_n$ are the random variables as defined in previous chapters, and Θ is the set of parameters of the network. Each element of this set is a function $\theta_{x_i|\vec{\pi}_i} : \mathbb{R}^{k_i} \to [0, 1]$*

that defines the probability of X_i conditioned on its parents $\vec{\pi}_i = (X_{r_1}, ..., X_{r_{k_i}})$,

$$\theta_{x_i|\vec{\pi}_i}(x_{r_1}, ..., x_{r_{k_i}}) := P\{X_i = x_i|(X_{r_1}, ..., X_{r_{k_i}}) = (x_{r_1}, ..., x_{r_{k_i}})\}$$

This Network defines a unique Joint Probability (or density) Distribution,

$$P_{\mathcal{B}}(x_1, ..., x_n) = \prod_{i=1}^{n} P_{\mathcal{B}}(x_i|\pi_i) = \prod_{i=1}^{n} \theta_{x_i|\pi_i}$$

It is easy to note that a Bayesian Network defines unequivocally a Structural Causal Model. If we want to predict a set of variables, V, the transformation from Bayesian Network to Structural Causal Model simply consists on assigning to the set of causal relations $\mathbf{F}$, the functions obtained from the Network parameters Θ, in any direction.

The study about propagation of information is not trivial, but, in our case, in order to do not focus on this peripheral theme, we will be using the Bayesian Network implemented in the Probabilistic Graphical Models library for Python, [18], and translate the information obtained via Causal Discovery to this model.

4.2. Naive Bayes

In class we saw in more detail what a Naive Bayes model is. Here it is introduced as a particular case of a Bayesian Network:

Definition 4.2. *A* ***Naive Bayes model*** *is a Bayesian Network where all of the edges have the same, unique, node destination, y.*

This means that we are assuming that all variables are independent, because we assume that there is no causal relation between them, and we are assuming that all of them have implications on the output Y:

Theorem 4.1. *Given a discrete Naive Bayes model, $\mathbb{M}$, with destination node Y, and whose features, or exogenous variables, are $X_1, ..., X_n$, then, for all $y_0 \in Image(Y)$:*

$$P_{\mathbb{M}}\{Y = y_0|X_1 = x_1, ..., X_n = x_n\} = \frac{P_{\mathbb{M}}\{Y = y_0\} \prod_{i=1}^{n} P_{\mathbb{M}}\{X_i = x_i|Y = y_0\}}{\sum_{y \in Image(Y)} \left(P_{\mathbb{M}}\{Y = y\} \prod_{i=1}^{n} P_{\mathbb{M}}\{X_i = x_i|Y = y\}\right)}$$

Proof

Trivial, using the Bayes Theorem. ■

It is easy to check that, effectively, this equation for the conditioned probability is equivalent to the formula in slide 15 of "*tema07 - Clasificador Bayesiano.pdf*", in Moodle.

5. Practical Study

To test the results of the last sections, it might be interesting to apply our algorithm to a dataset that is relatively understandable and simple, in order to get a wider perspective of how could our algorithm be used. This isn't an extensive work about why causal discovery gets good results (despite be will be actually be getting relatively good results), but an insight about the discovery of relations.

With this purpose, we have used the classical Titanic dataset [19], and we find ourselves interested in finding a classifier that is able to predict whether a passenger would have died. That way, as mentioned previously, the first step is to construct the fully connected graph, shown in Figure 5.1.

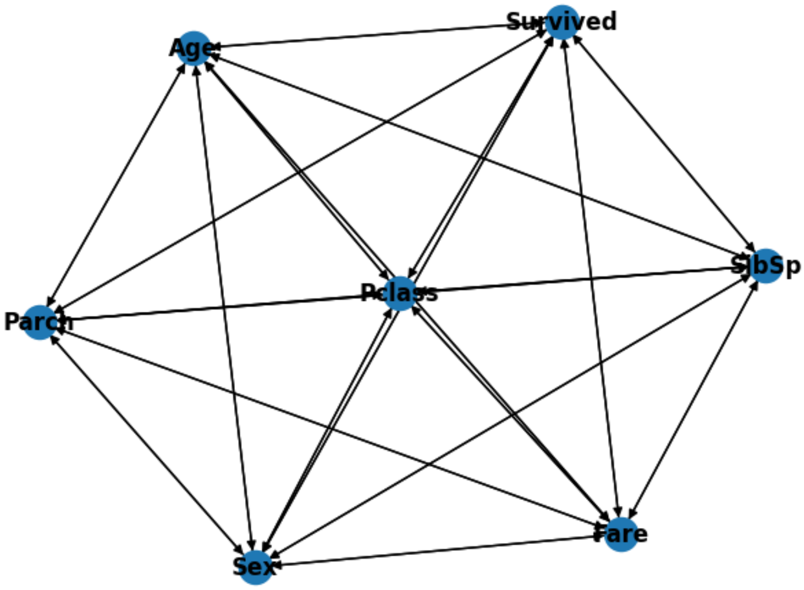

Figure 5.1: Fully connected graph with the Titanic dataset's variables.

First thing to do is to choose what is the type of some special variables:

- **Exogeneous** variables are those that can not be causally influenced by another variables. A first attempt to set exogeneous variables would be to use temporal information, and conclude that neither the *Sex* of a person or it's *Age* can be influenced by his parch or his fare.
- **Endogeneous** variables are those that cannot be the cause of any other. A clear examples for this case the target variable *survived*.

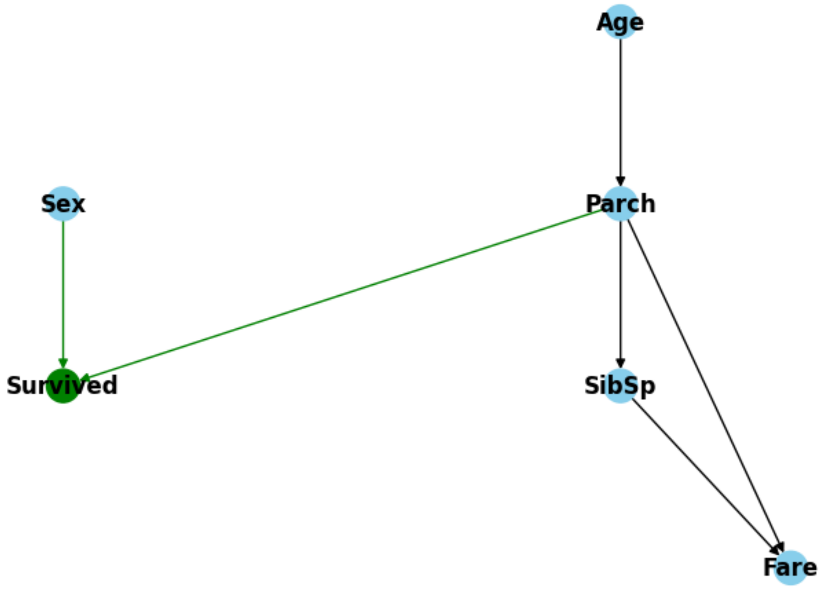

Figure 5.2: DAG obtained by our implementation of PC algorithm on the Titanic dataset.

Once this information is given to the model, and we use it to compute the DAG, we obtain the graph in 5.2. First interesting facts to see is that our assumptions were correctly used by the model, and both exogeneous and endogeneous variables work as they should. For instance, the target variable *Survived* is obtained from other variables, but it does not influence on another.

What this graph tells us is that *Sex* and *Parch* are the variables that have "direct implications" in the target variable. This doesn't mean that *Age* doesn't have information about the fact that a passenger survived or not, rather that the information it has about this is contained in the *Parch* variable. All this will be used when constructing a Bayesian Network in order to extract meaningful functions.

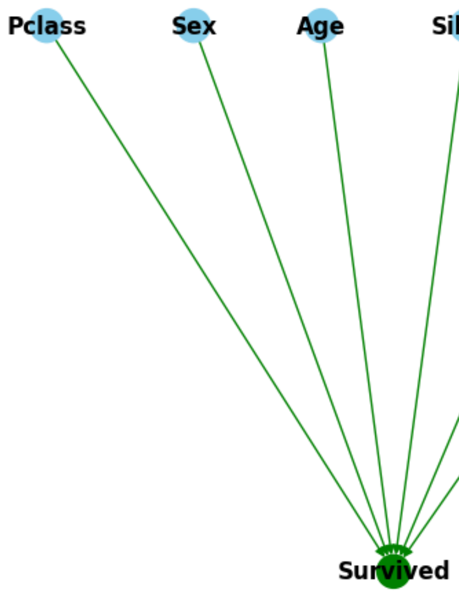

Figure 5.3: Causal relations that Naive Bayes supposes for the titanic dataset.

Though, before performing this, we might think that it is interesting to compare the results for our case with the results in a similar case that doesn't use our extracted relations. This is exactly what the previously explained *Naive Bayes* classifier is aimed at!!. It's graph is shown in 5.3.

That way, after training, with the exact same train data, our two classifiers and obtained the F1 Score they get in the test data, these are the results:

```
F1 Score:
Bayesian Network -> 0.7037037037037038
Naive Bayes -> 0.608695652173913
```

We could start comparing more metrics, but, since our purpose is simply understanding the model, this will be left apart.
At first sight, it might be surprising that our model is able to extract more information, because it is not using, for example, the variable *Pclass*, and it does not let enter the "probability information" directly to the target node.

Though, as seen during the whole study, the causal model is able to extract important relations between the data, discard the useless implications, and create an "intelligent" prediction, knowing where does the information come from and where it goes to.

6. Possible applications in machine learning

After everything that has been commented, it seems obvious that Causal Inference is a really powerful tool, and it is able, by itself, to contribute a lot. Though, this study has been carried out within the framework of the subject "Introducción al Aprendizaje Automático", so it might be interesting to review some implications and applications that this tool might have for the theme at hand. There are 2 main aspects in which Causal Inference might help to machine learning tools.

6.1. Improve models and make them more robust

There are many implicit assumptions that we are making at the moment of creating a machine learning model to predict an outcome.
For example, as commented in Section 1.2, if we want to predict the type of medicament that is going to be applied on a patient using machine learning, we would, traditionally, be working with correlations and side effects that do not necessarily happen when we *force* the use of a medicament. When assigning a medicament, we are doing an *intervention*, and these results can be easily improved via the study of causal relations.

Also, models trained purely for prediction may capture correlations that do not necessarily imply causation. This can lead to misleading insights, especially when deployed in new contexts. Causal inference promotes *robustness* by discerning causal relationships from mere associations. Models grounded in causal understanding are more likely to generalize well to unseen data and diverse scenarios. Consequently, they provide more reliable predictions and mitigate the risk of erroneous decision-making.

There are many studies and papers [20] [21] that corroborate these strength and robustness conclusions.

6.2. Ability to solve new problems

By using Causal Inference, it is possible to solve new problems, similar to those studied in machine learning, mixing the tools of both fields.

Structural Causal Models are a clear example of how, using causal ideas, we are able to obtain a different model to those that we had before, and even previous machine learning algorithms and models are able to be adapted to these models (for example, as Bayesian Networks can be converted to SCM; explained in Section 4.1), so that, even using previous tools, we are able to answer new questions, such as counterfactuals or interventions. These questions can also be answered with the tools, as explained in the example in Section 2.4.

7. State of the art and future scope

Causal Inference is a continuously evolving field, and there are tons of new studies performed on it every year, but here we will explain the state of the art of a particular field: *Causal Discovery for Time Series analysis and forecasting*. We have already seen that temporal relations are a crucial factor at the time of extracting implications. A clear example is the use of sex and age between the Endogenous variables during the titanic dataset study. Though, as it is commonly known, time series have a very important modelization, using Stochastic Processes and tools such as Brownians or Gaussian Processes.

All this field is being mixed with the causal discovery in order to create better and more efficient models, on top of being able to extract meaningful information from the raw data.

For the future work in this, as explained in [22], the use of deep learning as basis of the causal discovery is an important option. Another one explained in the same work is the mixture of the obtained DAG with another machine learning models in order to allow them use the causalities.

It is also an open problem the improvement of the way in which we are extracting the Conditional Independence, firstly, because we need to discretize the variables over which we want to test the independence. Though, this fist problem isn't expected to cause many problems, in general, since it has been use as base for independence since Pearson's papers, and it works correctly in most fields.

The main objective one might have is to optimize the way in which the conditional is introduced in the test. Here we have implemented the easiest one; simply applying the definition to test whether the variables are independent for every single case of Z.

This implies obvious problems with computation time, and with the fact that it the decomposition is not easily performed when Z is not discrete; during the way to do so a lot of information is lost, and, depending on it, the final results might be radically different. Though, in causal discovery it is important to have robust results, that represent real information, and not the one that has been stochastically concluded. For that reason, the study and development of Conditional Independence tests that are able to consider continuous variables is an interesting and useful possible future work.

Bibliography

[1] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, " O'Reilly Media, Inc.", 2022.

[2] L. Serrano, Grokking machine learning, Simon and Schuster, 2021.

[3] S. M. Stigler, The history of statistics, Harvard University Press, 1990.

[4] V. Q. Paloma, A. G. Pérez, Lecciones de cálculo de probabilidades, Ediciones Díaz de Santos, 1988.

[5] R. Vélez Ibarrola, V. Hernández Morales, Cálculo de probabilidades 1, Universidad Nacional de Educación a Distancia, 1995.

[6] V. I. Ricardo, Cálculo de probabilidades 2, Universidad Nacional de Educación a Distancia, 2019.

[7] J. L. Abad, V. O. Prieto, Integral de Lebesgue, Sanz y Torres, S.L., 2024.

[8] J. Pearl, Causal inference in statistics: An overview, Wiley, 2009.

[9] J. Pearl, D. Mackenzie, The Book of Why, Basic Books, 2018.

[10] J. E. Marsden, A. J. Tromba, M. L. Mateos, Cálculo vectorial, Vol. 69, Addison-Wesley Iberoamericana México, 1991.

[11] C. R. Rao, C. R. Rao, M. Statistiker, C. R. Rao, C. R. Rao, Linear statistical inference and its applications, Vol. 2, Wiley New York, 1973.

[12] R. V. Ibarrola, A. G. Pérez, Principios de inferencia estadística, UNED, Universidad Nacional de Educación a Distancia, 2013.

[13] D. Q. Zha, statslab.cam.ac.uk, `https://www.statslab.cam.ac.uk//~qz280/teaching/causal-2023/notes-2021.pdf`, [Accessed 23-03-2024] (2021).

[14] J. Peters, D. Janzing, B. Schölkopf, Elements of causal inference: foundations and learning algorithms, The MIT Press, 2017.

[15] C. Hitchcock, M. Rédei, Reichenbachs common cause principle (Jan 2020).
URL `https://plato.stanford.edu/entries/physics-Rpcc/`

[16] T. D. Le, T. Hoang, J. Li, L. Liu, H. Liu, , S. Hu, arxiv.org, `https://arxiv.org/pdf/1502.02454.pdf#:~:text=It%20is%20a%20concep%2D%20tual,based%20on%20conditional%20independence%20decisions.`, [Accessed 24-03-2024] (2014).

[17] F. F. . K. R. Ruggeri F., Bayesian Networks — web.archive.org, `https://web.archive.org/web/20161123041500/http://www.eng.tau.ac.il/~bengal/BN.pdf`, [Accessed 19-04-2024] (2007).

[18] pgmpy, Bayesian Network; pgmpy 0.1.23 documentation — pgmpy.org, `https://pgmpy.org/models/bayesiannetwork.html`, [Accessed 19-04-2024] (2024).

[19] Kaggle, Titanic - Machine Learning from Disaster — kaggle.com, `https://www.kaggle.com/c/titanic/data`, [Accessed 16-05-2024] (2012).

[20] Y. Xin, N. Tagasovska, F. Perez-Cruz, M. Raubal, Vision paper: Causal inference for interpretable and robust machine learning in mobility analysis (2022). `arXiv:2210.10010`.

[21] V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, V. Syrgkanis, Applied causal inference powered by ml and ai (2024). `arXiv:2403.02467`.

[22] A. Nouri, diva-portal.org, `https://www.diva-portal.org/smash/get/diva2:1749596/FULLTEXT01.pdf`, [Accessed 16-05-2024] (2023).