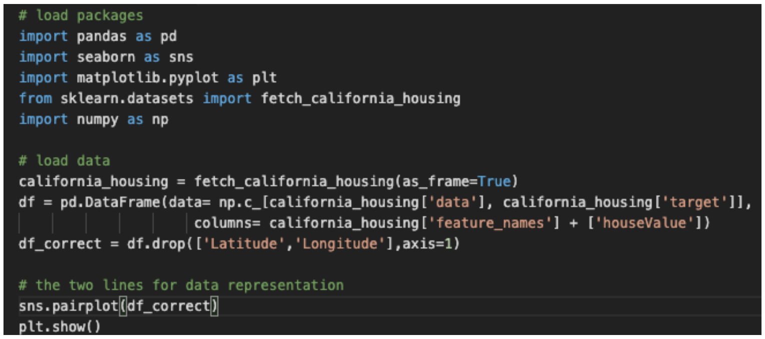
Introduction to DAGs, their applicability and how they come about

Posted November 28, 2021 by Anders Bast Olsen – 5 min read

Correlation does not imply causation, but what does imply causation then?

In most machine learning, the analyst's starting point will be to get a sense of the underlying relationships in a dataset, formally known as *exploratory data analysis* (EDA). During EDA, especially a scatterplot is a popular tool to visualize correlation between any given variables in two-dimensional space. The popularity may in part be attributed to the ease of use, that is, in just two lines of codes the underlying relationships in the data, in the form of scatterplots, are generated across the dataset. See for instance the case of California housing data:



```
# load packages
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import fetch_california_housing
import numpy as np

# load data
california_housing = fetch_california_housing(as_frame=True)
df = pd.DataFrame(data= np.c_[california_housing['data'], california_housing['target']],
                     columns= california_housing['feature_names'] + ['houseValue'])
df_correct = df.drop(['Latitude','Longitude'],axis=1)

# the two lines for data representation
sns.pairplot(df_correct)
plt.show()
```

Fig 1 – python code for generating scatterplots

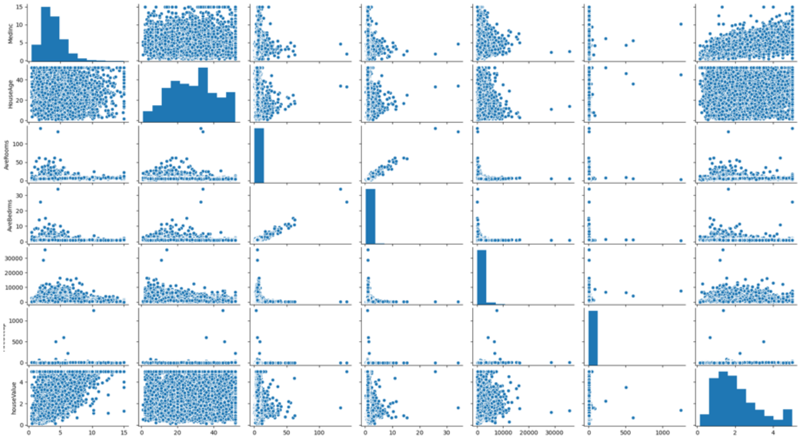


Fig 2 – scatterplots for all variables in the California housing

However, the ease of use does not necessarily make it the best representation of the data at hand. Clearly, the best representation depends on the purpose. For the sake of the argument, let's define the purpose: to find the effect of x, number of bedrooms, on y, market value of the house. Here, we are searching for a causal effect, for which an understanding of the underlying causal relationships would be of more help than one of the correlational relationships. To this regard, the best representation would be the *directed acyclic graph* (DAG), which expresses the causal relationship between each of the observed variables:

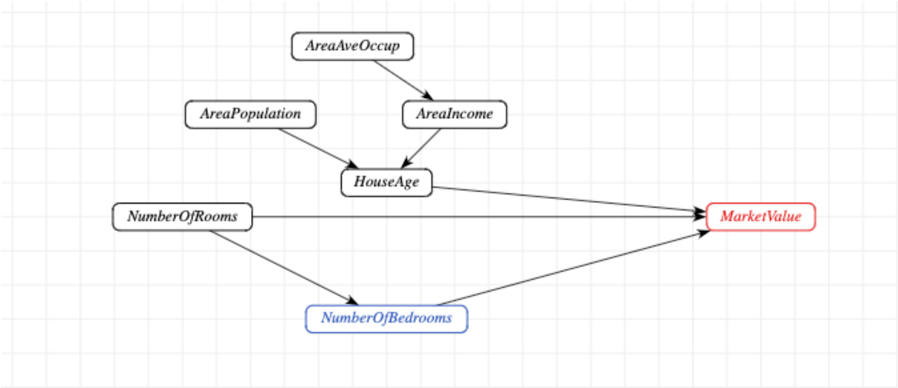


Fig 3 – dummy DAG made with causalfusion

From the DAG in fig 3, an intuitive and very useful interpretation of the causal relationships can be obtained, merely upon inspection of the directionality. Hence, to find the effect of x on y, the DAG's directionality shows that it would require controlling for another variable, number of rooms. To make the life of the analyst even easier, using software packages like www.causalfusion.org, answering the query, effect of x on y, is fully automated when the DAG is drawn. Thus, if the analyst gives the software the query relevant for the purpose, such as, $P(MarketValue|do(NumberOfBedrooms)$, an estimable equation for that query, if one exists, is returned:

$$\sum_{NumberOfRooms} P(MarketValue|NumberOfBedrooms, NumberOfRooms)P(NumberOfRooms)$$

Nothing is that easy, where is the catch?

Obviously, no matter how beautiful this automated computation is; it is not magic. The computation is based on directional and mathematical interpretation of the DAG and several reformats of that DAG, all of which is summed up by three rules of do-calculus. However, since these rules have already been automated and readily available for everyone to use, the "catch" is not the underlying use of math. Rather, the "catch" is the process in which the DAG comes about; after all, the DAG in fig 3 is based purely on guesswork.

From raw data to DAG using causal discovery

Since the application of do-calculus, automated or not, assumes a credible underlying causal model, pure guesswork is far from sufficient. In fact, most of the analyst's effort is likely spend extrapolating a credible causal model, a process formally known as causal discovery. Unlike do-calculus, causal discovery does not operate in a finite problem space in which completeness has been proven. As such, the discovery tends to be an iterative process, wherein methodologies even range from pure social science to pure math, or a combination of the two. Irrespective of the method, four assumptions are commonly shared across different causal models: Acyclicity, Markov Property, Faithfulness, Sufficiency. In short, these assumptions can be summarized respectively as the ability to represent the causal model as a DAG (G), that each node is independent of their non-descendants when conditioned on their parent nodes, conditional independences in true underlying distribution p are represented in G, and any pair of nodes in G has no common external cause.

Approaching the causal discovery using social science, may come down to having domain and statistical experts collaborate, such that the human expertise is gets embedded in the causal model, e.g., "California houses the most one-percenters in the nation since great tax reductions are offered for people building new houses." Clearly, this is a made-up example, but assuming it was not, the causal model would embed this domain expertise by specifying, AreaIncome → HouseAge.

Though beyond the scope of this paper, a popular mathematical approach worth highlighting is the conditional independence testing. Here, the analyst starts from fully connected and undirected graph, and would then, based on the argument that two statistically independent variables are not causally linked to one another, remove connected edges, and orient the directions amongst the remaining edges.

So where does that leave the potential of the causal model?

As shown earlier, a causal query may be just as easy to estimate using DAGs and fusion, as fitting a regression-plane to the independent variables showing the highest correlation to the dependent variable (see fig 2). And as a bonus, uncovering causality and not correlation, should yield a better performing and more credible result. However, the realm of causal modelling prerequisites a causal model, which the realm of machine learning does not. Still, with the do-calculus being fully automated and the superiority in knowing causation, compared to correlation, causal discovery may, if not fully automated, at least become much easier over time, making the future potential tremendous.