Conditional Probability Approach for Causal Effect Inference with Deep Latent-Variable Models

Mehmet Enis İşgören

Abstract

Discovering causal relations from observational data is an important aspect of current machine learning research. With availability of big data researchers are able to model much more complex problems but still limited in terms of handling confounders, factors that that affect both researched method and output. To build the best model one should model all factors that affects the research but in real life it may not be possible to get all confounders, but confounders may be found via proxies which may have some noise. In this paper a previous method which is build on Variational Autoencoders (VAE) and showed that their method matches the state-of-the-art on benchmarks focused on individual treatment effects.

1. Introduction

Understanding the change's effect to the outcome is important for researching new methodologies and common in different fields of research such as medicine and psychology. In both fields subject is humans and there are huge amount of details affects person's psychical and mental capacities. With the availability of big data, researchers are trying to come up with new drugs and new therapies for patients but just giving the drug to everybody and just seeing whether they are cured or not isn't sufficient way to do since it may just be that people are recovering from the disease without help of the drug, it is needed to have a control group to see whether an intervention is causing changes. When people are grouped to test the affects of the drug, it is needed to know the factors other than researched drug that affects the treatment that can be such as socio-economic status or genetic factors, these factors are important aspects of causal relationship research named confounders. Confounders affect both intervention and the outcome so if confounders can be measured such as genetic factors assuming all the details of the disease is known, it is needed to be modeled. However, if confounders can't be measured, such as socio-economic status which affects the general health of a person. These factors are hard to model even when it is assumed that genetic factors are well known since the affect of wealth on health isn't certain or how well genetic factors affect person's health. So directly modeling them in real life cases is somewhat impossible instead of directly modelling them they can be modelled via proxy variables for example person's socio-economic status can be estimated by the place they live and it can be observed, availability of big data is an opportunity to find possible interventions and also good for discovering proxy confounders.

Another important factor that is needed to be reminded is that proxy confounders cause bias since they are like outcomes of confounders, again there may be factors that can't be fully known, see the Figure 1. So directly using them as confounders isn't feasible. New method of using available proxies as fruitful knowledge carriers of confounder by using latent variable models to discover hidden confounders and how they affect the intervention and outcome specifically using maximum-likelihood methods (Louizos et al., 2017). Their paper builds on VAEs and VAEs are shown to be successful in different fields to capture latent structure. In this paper their method is going to be used to identify causal effect with changes and in the next chapter what changes are going to be applied will be discussed.

2. Identification of Causal Effect

In their paper(Louizos et al., 2017) they assume the causal model in Figure and also they assume that intervention t is binary. They further assume that the joint distribution p (Z, X, t, y) of the latent confounders Z and the observed confounders X can be approximately recovered solely from the observations (X, t, y). Their goal is to recover individual treatment effect(ITE) also known as the conditional average treatment effect (CATE), of a treatment t, as well as the average treatment effect (ATE):

$$ITE(x) := E[y|X = x, do(t = 1)] - E[y|X = x, do(t = 0)] \quad (1)$$

$$ATE := E[ITE(x)] \quad (2)$$

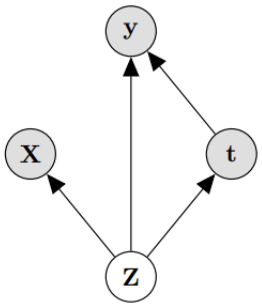

Figure 1. Example of a proxy variable. t is a treatment, e.g. drug; y is an outcome, e.g. whether a person is cured or not. Z is an unobserved confounder, e.g. socio-economic status; and X is noisy views on the confounder Z, say income in the last year and place of residence.

Their theorem is using joint distribution and they are proving that observing (X,t,y) is enough to recover ITE. However, they use joint distribution but they never interfere on X so instead of using joint distribution they can and should use a conditional distribution it will also be sufficient and simpler in terms of modeling. In the next subsection, a new theorem will be build and proofed.

2.1. Identifying individual treatment effect

Theorem 1: If we recover p (Z, t, y | X) then we recover the ITE under the causal model in Figure 1.
So to prove the theorem do-calculus will be applied to t as t=1 but t=0 case is identical.

Proff

$$p(Z, y, do(t = 1)|X) = \tag{3}$$

If Bayes' Theorem is applied to equation 3, equation 4 will be the result since $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

$$\frac{p(X|y, Z, do(t = 1)).p(y, Z, do(t = 1))}{p(X)} = \tag{4}$$

Then product rule is applied to the term($P(y, Z, do(t = 1))$ in equation 4 then acquired equation 5/

$$\frac{p(X|y, Z, do(t = 1)).p(y|Z, do(t = 1)).p(Z|do(t = 1))}{p(X)} \tag{5}$$

So by the rules of *do-calculus* for the fraction part $P(X|y, Z, do(t = 1))$ becomes $P(X|Z)$since X is independent from y,t when t is conditioned. $P(Z|do(t = 1))$ simply becomes $P(Z)$ since t is again conditioned and Z is now independent from.

$$\frac{p(y|Z, do(t = 1).p(X|Z).p(Z)}{p(X)} \tag{6}$$

So here to equation 6, once again Bayes' Theorem is applied to convert $\frac{P(X|Z)P(Z)}{P(X)}$ to $P(Z|X)$. Then to model the effect integral is applied in term of Z. Result will be equation 7

$$\int_Z p(y|Z, do(t = 1)).p(Z|X)dZ \tag{7}$$

As a result it is enough and sufficient to use conditional probability to model the encoder.

2.2. Individual Treatment Effect of Continuous Treatment

In this paper theorem and experiment data set will be different than deep latent variable model paper (Louizos et al., 2017). So their case of binary treatment isn't going to be sufficient enough since in this paper Boston (Harrison & Rubinfeld, 1978), Concrete (Yeh, 1998) and Energy(Tsanas & Xifara, 2012) data sets will be used for experiments. This data sets all have continuous values in their respective columns and treatment effect should be calculated for continuous values. Method to determine the treatment effect by using this data sets, for each column in a data set iteratively one of the columns will be pick as treatment and the last column *v* is the label column u = $|u_1, ..., u_D|t = u_d, x = u/u_d, y = v$.

So to determine scale of treatment effect, *min* and max values of the corresponding treatment column will be defined then model will be intervened by this values, difference between results of max intervened and min intervened models will be named as Causal Range, difference between *min* and *max* values will be named as *Effect Range*. To determine ITE, following formula will be applied. ITE = Causal Range/Effect Range. This will be ITE and after ITE is calculated for all samples, ATE will be the mean of collected values.

3. Causal Effect Autoencoders

Overall architecture of generative model is as following;
$p(z) = \mathcal{N}(z \mid 0, I_d)$

$p(y|z,t) = \mathcal{N}(y \mid g_2 \circ g_1([z,t]), g_3 \circ g_1([z,t]))$
$p(t|z) = \mathcal{N}(t \mid g_4(z), g_5(z))$
$p(x|z) = \mathcal{N}(x \mid g_6(z), g_7(z))$
where I_d is the identity matrix with dimensionality d$\times$ d, N is being Gaussian distribution and g_1 , g_2 , g_3 , g_4 , g_5 , g_6 , g_7 are neural networks with following structure $Linear \rightarrow ReLU \rightarrow Linear \rightarrow ReLU \rightarrow Linear$. This generative model only requires input of Z, but as discussed before it's not available. To find Z following inference network is constructed;

$q(z|x,t) = \mathcal{N}(z|g_9 \circ g_8([x,t]), g_{10} \circ g_8([x,t]))$

$p()$ is Model network , and $q()$ is being Inference Network, together they are implemented as Causal Effect Variational Autoencoder model.

3.1. Inference

Assuming $p(z \mid x,t) \approx q(z \mid x,t)$ which brings about an Evidence Lower Bound (ELBo)

$logp(y|x,t) \geq E_{q(z|x,t)}[logp(y|t,z)]DKL(q(z|x,t)||p(z))$
by implementing this ELBo and using following prediction process, experiment are done.

$p(y^*|x^*, do(t=t^*)) = p(y^*|z, do(t=t^*))q(z|x^*, do(t=t^*))dz$

4. Experiments

To evaluate performance of our causal inference model and modelling network 3 data sets will be used, Boston(Harrison & Rubinfeld, 1978), Concrete(Yeh, 1998) and Energy(Tsanas & Xifara, 2012). All of the data sets have one label column which is the last one which will be shown with v and rest are variable columns will be shown as $u = [u_1, ..., u_D]$, Boston data set has 13, Concrete has 9, Energy has 8 variable columns. First using this variable columns and splitting data the rows by %90 to training, a linear regressor is fit and root mean square error (RMSE) of test dataset is calculated by using trained model and labeled as $RMSE_{full}$ this experiment was repeated 10 times and results are included in graph COME HERE AND CONNECT THE GRAPH. Then for each dimesion of input u; $t = u_d, x = u/u_d, y = v$ is set and using Model network and inference network ATE effect is calculated for columns. ATE results of columns for each dataset as following for simplicity values are rounded;

Concrete Data set	
Column Name	Average Treatment Effect
Cement	2.298
Blast Furnace Slag	0.188
Fly Ash	0.950
Water	0.0916
Superplasticizer	0.0475
Coarse Aggregate	1.711
Fine Aggregate	1.369
Age(day)	0.114

Boston Data set	
Column Name	Average Treatment Effect
CRIM	0.073
ZN	0.2432
INDUS	0.014
CHAS	0.006
NOX	0.0018
RM	0.0159
AGE	0.0365
DIS	0.0712
RAD	0.0447
TAX	0.6518
PTRATIO	0.0244
B - 1000	0.3484
LSTAT	0.0074

CRIM - per capita crime rate by town
ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS - proportion of non-retail business acres per town.
CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX - nitric oxides concentration (parts per 10 million)
RM - average number of rooms per dwelling
AGE - proportion of owner-occupied units built prior to 1940
DIS - weighted distances to five Boston employment centres
RAD - index of accessibility to radial highways
TAX - full-value property-tax rate per $10,000
PTRATIO - pupil-teacher ratio by town
B - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT - % lower status of the population

Energy Data set	
Column Name	Average Treatment Effect
Relative Compactness	0.0080
Surface Area	0.092
Wall Area	1.272
Roof Area	0.498
Overall Height	0.009
Orientation	0.009
Glazing Area	0.0044
Glazing Area Distribution	0.0044

From this results top performing 3 variables are selected and put in to a vector called $u^{'}$. By using $u^{'}$ another linear regression and evaluated the performance with root mean square error (RMSE) on test split, splitted same as before, result is named as $\text{RMSE}_{\text{causal}}$. Again this experiment with $u^{'}$ repeated 10 times with only 1 epoch, mean and standard error margin of results are put in to following the table. M is mean and E is error margin

	$\text{RMSE}_{\text{full}}$	$\text{RMSE}_{\text{causal}}$
Boston	M:**9.8602** E:1.220	M:**11.4896** E:0.879
Energy	M:**13.672** E:1.865	M:**13.99** E:1.77
Concrete	M:**20.16** E:2.559	M:**24.539** E:2.57

5. Evaluation of Results

Results for datasets are further investigated in the following sections.

5.1. Boston Dataset

In Boston dataset what model is trying to predict is Median value of owner-occupied homes in $1000's and the columns with highest average treatment effect values are TAX - full-value property-tax rate per $10,000, ZN - proportion of residential land zoned for lots over 25,000 sq.ft. and B - 1000(Bk0.63)2where Bk is the proportion of blacksby town. Basically higher the tax of a house more luxurious a house should be so direct causation between median value and tax is justifiable. ZN is how big of land a house is located, bigger the land bigger the house, B-1000 is the variable that is more hidden but makes sense, since the distribution of wealth in US isn't equally distributed and whites control the most wealth, lower the ration of black people more luxurious a house is. Other variables related to this properties such as average room per house is related to how big a house is. Other variables seems to be way lower than this ones in terms of average treatment effect, when variables with highest three ATE values are selected to train linear regression, linear regression still does a good job, average RMSE increases but the error decreases.

5.2. Energy Dataset

In energy dataset models are trying to predict amount of energy added to a space to maintain the temperature. Wall and Roof area seems to be the ones with highest ATE others are mostly little details. Furthermore, it is again justifiable that wall and roof area is mostly related aspects to the preserved heat of a room. In this case 3 variables with highest ATE is used but it can be further deducted to two variables. Since ATE values are this high when other variables are removed, RMSE doesn't change much but Error decreases a little.

5.3. Concrete Dataset

In concrete data set, what models are trying to predict is strength of the concrete. Variables with highest ATE values are cement, coarse aggregate and fine aggregate which are basic ingredients of concrete and most crucial ones. Other variables are location specific ones and Fly Ash seems to be somewhat important and Fly ash use in concrete improves the workability, strength and durability of concrete. In modelling RMSE increases to 24.539 from 20.16, this is most likely caused by eliminating the Fly Ash value when model is trained and tested.

References

Harrison, D. and Rubinfeld, D. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 03 1978. doi: 10.1016/0095-0696(78)90006-2.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems*, pp. 6446–6456, 2017.

Tsanas, A. and Xifara, A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49: 560–567, 06 2012. doi: 10.1016/j.enbuild.2012.03.003.

Yeh, I.-C. Modeling of strength of high-performance concrete using artificial neural networks." cement and concrete research, 28(12), 1797-1808. *Cement and Concrete Research*, 28:1797–1808, 12 1998. doi: 10.1016/S0008-8846(98)00165-3.