Machine Learning 2 **Due: 21 May, 2016**

Homework 6

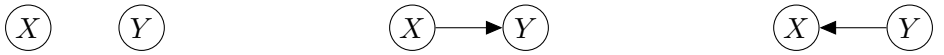
Instructor: Ke Tran Email: m.k.tran@uva.nl

Student: Andrea Jemmett UvA ID: 11162929
Collaborators: N/A Email: andrea.jemmett@student.uva.nl

Problem 2.

Solution:

(a)



(a) No cause-effect relation (b) X causes Y (c) Y causes X

Figure 1: Two nodes Causal Bayesian Networks

(b) For the Causal Bayesian Networks in Figures 1a, 1b and 1c respectively we have:

$$p(X,Y) = p(X)p(Y) \tag{1}$$
$$p(X,Y) = p(Y|X)p(X) \tag{2}$$
$$p(X,Y) = p(X|Y)p(Y) \tag{3}$$

(c) For the Causal Bayesian Networks in Figures 1a and 1c respectively we have:

$$p(Y|X) = p(X)p(Y) \tag{4}$$
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)} \tag{5}$$

while $p(Y|X)$ is already a term of the factorization for the graph in Figure 1b.

(d) For the Causal Bayesian Networks in Figures 1a, 1b and 1c respectively we have:

$$p(Y|do(X)) = p(Y) \tag{6}$$
$$p(Y|do(X)) = \frac{p(X,Y)}{p(X)} = p(Y|X) \tag{7}$$
$$p(Y|do(X)) = p(Y) \tag{8}$$

(e) $p(Y|X)$ is the probability of having lung cancer given the *observation* of smoking. $p(Y|do(X))$ instead represents the probability of having lung cancer given that we *force* that person to (not) smoke. Usual conditioning captures the correlation between two random variables whereas the do-operator represents an active intervention on the model and is capable of representing causation relationships.

■

Problem 3. Simpson's paradox

Solution:

1a. The recovery rate for *treatment* is 50%, while for *untreated* is 40%.

1b. I would advice to take the drug because the recovery rate is higher for the *treatment* group.

2a.

Recovery rates	Drug	No drug
Male	60%	70%
Female	20%	30%

2b. I would not advice to take the drug nor to male nor female patients because the recovery rate, given the patient's gender, does not support it.

3. With hindsight I would not advice a patient with unknown gender to take the drug because for both genders the recovery rate does not support it. This is in contradiction with the conclusion given in (1b).

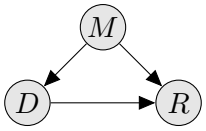


Figure 2: Causal model where M denotes the gender.

4a. By applying the back-door criterion on the causal model in Figure 2 we have:

$$p(R|do(D)) = \sum_M p(R|D,M)p(M) \tag{9}$$

because M is admissible for adjustment to find the causal effect of D on R.

4b. Using normal probability rules we have:

$$p(R|D) = \sum_M p(R,M|D) = \sum_M p(R|D,M)p(D|M) \tag{10}$$

which is generally not equal to $\sum_M p(R|D,M)p(M)$ and so $p(R|do(D)) \neq p(R|D)$ in this case.

4c. We have expressed $p(R|do(D))$ in terms of observables and we have the data to carry on the calculation:

$$p(R|do(D)) = p(R|D,M=m)p(M=m) + p(R|D,M=f)p(M=f) = 40\% \tag{11}$$
$$p(R|do(\neg D)) = p(R|\neg D,M=m)p(M=m) + p(R|\neg D,M=f)p(M=f) = 50\% \tag{12}$$

and so in this case I would not advice on taking the drug. The same conclusion could have been reached without computation by noting that $p(R|do(D)) \neq p(R|D)$.

5a. By applying the back-door criterion to the causal model in Figure 3 we have:

$$p(R|do(D)) = p(R|D) \tag{13}$$

because $\emptyset$ is admissible for adjustment to find the causal effect of D on R.

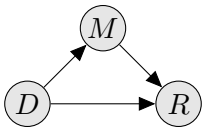


Figure 3: Causal model where M denotes for example "blood pressure".

5b. Yes, $p(R|do(D)) = p(R|D)$.

5c. From the data we have that:

$$p(R|do(D)) = p(R|D) = 50\% \tag{14}$$
$$p(R|do(\neg D)) = p(R|\neg D) = 40\% \tag{15}$$

so my advice in this case would be to take the drug. The same conclusion could have been reached without computation by noting that $p(R|do(D)) = p(R|D)$.

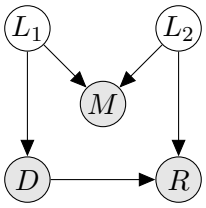


Figure 4: Causal model where M, L_1 and L_2 have meanings given in (6a).

6a. We can imagine a causal model where M denotes a certain kind of a specific disease (a certain kind of influenza for example), which is observed and so we can assume that it is diagnosable. The latent variables L_1 and L_2 might represent whether the patient has a congenital disorder that contributes as a cause to the observed disease and the effect of the drug (L_1) and if among his/her ancestors the same disease is present (L_2, so this affects the chances of recovery and the kind of disease the patient may have). We can suppose that the congenital disorder is not (easily) diagnosable and we do not have data for the patient's ancestors. The causal model is given in Figure 4.

6b. We can do this in one step by applying the second rule of do-calculus, because $(R \perp\!\!\!\perp D)_{\mathcal{G}_{\underline{D}}}$:

$$p(R|do(D)) = p(R|D) \tag{16}$$

6c. Yes, $p(R|do(D)) = p(R|D)$ in this case.

6d. My advice in this case would be to take the drug for the same reasons given in (5c).

■

Problem 5. Truncated factorization

Solution:

1. Assuming that both Y and nodes in $\mathbf{X}_{pa(X)}$ have no parents (this is just an assumption made to have a more uncluttered notation, but the result and derivation hold also without it) we have:

$$p(Y|do(X), \mathbf{X}_{pa(X)}) = \frac{p(Y, \mathbf{X}_{pa(X)}|do(X))}{p(\mathbf{X}_{pa(X)}|do(X))} \tag{17}$$

$$= \frac{p(Y)\prod_{X_i \in \mathbf{X}_{pa(X)}} p(X_i)}{\prod_{X_i \in \mathbf{X}_{pa(X)}} p(X_i)} \tag{18}$$

$$= p(Y) \tag{19}$$

$$= \frac{p(Y)p(X|\mathbf{X}_{pa(X)})}{p(X|\mathbf{X}_{pa(X)})} \tag{20}$$

$$= \frac{p(Y, X, \mathbf{X}_{pa(X)})}{p(X, \mathbf{X}_{pa(X)})} \tag{21}$$

$$= p(Y|X, \mathbf{X}_{pa(X)}) \tag{22}$$

2. Using the previous result:

$$p(Y|do(X)) = \int p(Y|do(X), \mathbf{X}_{pa(X)})p(\mathbf{X}_{pa(X)}|do(X))\mathrm{d}\mathbf{X}_{pa(X)} \tag{23}$$

$$= \int p(Y|X, \mathbf{X}_{pa(X)}) \prod_{X_i \in \mathbf{X}_{pa(X)}} p(X_i)\mathrm{d}\mathbf{X}_{pa(X)} \tag{24}$$

$$= \int p(Y|X, \mathbf{X}_{pa(X)})p(\mathbf{X}_{pa(X)})\mathrm{d}\mathbf{X}_{pa(X)} \tag{25}$$

■