# Counterfactuals

Eric Rosen

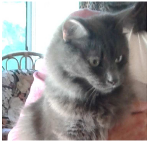
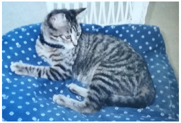
Machine Learning Reading Group, UBC, August 22, 2017

# What are counterfactuals and how do we calculate counterfactual probabilities?

1. An example from Pearl's Causality book, chapter 7, with the characters changed.
2. An example from Balke and Pearl (1994) *Probabilistic evaluation of counterfactual queries* (characters changed again)
3. Johansson et al (2016) *Learning representations for counterfactual inference*

# Example from Pearl's book

Suppose that we have:

- a cat, Oscar 
- another cat: Bastet 
- a bird feeder outside that we assume is always populated with birds unless at least one cat is outside
- and a window.

If the door is open the cats will always go outside. The door is normally closed but if the temperature outside goes above 20C, 🌡 someone will open it. There are always birds at the feeder unless there is at least one cat outside in which case they will all leave the feeder. We have the following propositions:

- T = The temperature is above 20C.
- D = Someone opens the door.
- O = Oscar goes outside.
- B = Bastet goes outside.
- L = All the birds leave the feeding station.

# Some sentences:

1. *Prediction* If Bastet did not go outside then there are birds at the feeding station.

   $\neg B \Rightarrow \neg L$.

2. *Abduction* If there are birds at the feeder then no one opened the door.

   $\neg L \Rightarrow \neg D$. (Given $D \Rightarrow O \wedge B$, and $B \vee O \Rightarrow L$, then $D \Rightarrow L$, so its contrapositive is true.)

3. *Transduction* If Oscar went outside then so did Bastet.

   $O \Rightarrow B$ (Given $D \Leftrightarrow B$ and $D \Leftrightarrow O$, then $O \Rightarrow D$. So $O \Rightarrow B$.)

4. *Action* If no one opened the door and Bastet snuck outside through a window then all the birds will leave the feeder and Oscar will remain inside.

   $\neg D \Rightarrow L_B \ \& \ \neg O_B$

5. *Counterfactual* If the birds have left the feeder then they still would have left the feeder even if Bastet had not gone outside.

   $L \Rightarrow L_{\neg B}$

# Sentences 1 - 3 can be handled by standard logical deduction.

Pearl p.209.:

> *"The feature that renders S1 - S3 manageable in standard logic is that they all deal with epistemic inference – that is, inference from beliefs to beliefs about a static world."*
>
> ⋮
>
> *"From our discussion of actions . . . , any such action must violate some premises, or mechanisms, in the initial theory of the story. To formally identify what remains invariant under the action, we must incorporate causal relationships into the theory; logical relationships alone are not sufficient."*

# Equality to show two-way inference

Pearl uses equality rather than implication in order to permit two-way inference. The independent variable is given in brackets in the second column below, to demonstrate the causal asymmetry.

Here is the causal model so far:

Model $M$

$$\begin{array}{lll}
 & (T) & \\
D = T & (D) & \text{(Door opens iff temp > 20C.)} \\
O = D & (O) & \text{(Oscar goes out iff door opens.)} \\
B = D & (B) & \text{(Bastet goes out iff door opens.)} \\
L = O \vee B & (L) & \text{(Birds leave iff Oscar or Bastet goes out)}
\end{array}$$

# A submodel

To evaluate S4, ($\neg D \Rightarrow L_B$ & $\neg O_B$) we form submodel $M_B$ in which the equation $B = D$ is replaced by $B$.

If no one opened the door and Bastet snuck outside through a window then all the birds will leave the feeder and Oscar will remain inside.

Model $M_B$

$$\begin{array}{ll}
 & (T) \\
D = T & (D) \\
O = D & (O) \\
B & (B) \\
L = O \vee B & (L) \\
\hline
\text{Facts: } \neg D & \\
\hline\hline
\text{Conclusions: } B, L, \neg O, \neg T, \neg D &
\end{array}$$

$\neg D \Rightarrow \neg O$ by contrapositive but $L$ is still true since $B \Rightarrow L$.

# Pearl's view

Pearl, pp. 209-210:

> *"It is important to note that problematic sentences like S4, whose antecedent violates one of the basic premises in the story, [in this case, that Bastet got outside without the door being opened] are handled naturally in the same deterministic setting in which the story is told. Traditional logicians and probabilists tend to reject sentences like S4 as contradictory and insist on reformulating the problem probabilistically so as to tolerate exceptions to [a] law. . . . Such reformulations are unnecessary; the structural approach permits us to process commonplace causal statements in their natural deterministic habitat without first immersing them in nondeterministic decor. In this framework, all laws are understood to represent defeasible default expressions  subject to breakdown by deliberate intervention."*

# Evaluating counterfactuals: step 1

If the birds have left the feeder then they still would have left the feeder even if Bastet had not gone outside.

The counterfactual $L_{\neg B}$ stands for the value of $L$ in submodel $M_{\neg B}$ below. The value $L$ depends on the value of $T$, which is not specified in $M_{\neg B}$ The observation $L$ removes the ambiguity: if we see the birds have left the feeder, we can infer that the temperature rose above 20C and thus the door was opened. If Bastet had not gone outside then Oscar would have, scaring the birds away from the feeder. We can derive $L_{\neg B}$ as follows.

We add the fact $L$ to the original model and evaluate $T$. Then we form submodel $M_{\neg B}$ and reevaluate $L$ in $M_{\neg B}$ using the value of $T$ found in the first step.

Model $M$ (Step 1)

| | | |
|---|---|---|
| | $(T)$ | |
| $D = T$ | $(D)$ | (Door opens iff temp > 20C.) |
| $O = D$ | $(B)$ | (Oscar goes out iff door opens.) |
| $B = D$ | $(O)$ | (Bastet goes out iff door opens.) |
| $L = O \vee B$ | $(L)$ | (Birds leave iff Oscar or Bastet go out) |
| Facts: $L$ | | (Birds leave the feeder.) |
| Conclusions: $T, B, O, D, L$ | | |

# Evaluating counterfactuals: step 2

Step 2
Model $M_{\neg B}$

| | | |
|---|---|---|
| | $(T)$ | |
| $D = T$ | $(D)$ | (Door opens iff temp > 20C.) |
| $O = D$ | $(O)$ | (Oscar goes out iff door opens.) |
| $\neg B$ | $(B)$ | (Bastet does not go out.) |
| $L = O \vee B$ | $(L)$ | (Birds leave iff Oscar or Bastet go out) |

Facts: $T$

Conclusions: $T, \neg B, O, D, L$

Pearl remarks that it is only the value of $T$ which he refers to as a ‘background variable’ that is carried over to Step 2. Everything else must be re-evaluated.

Pearl’s next step is to combine steps 1 and 2 into one by using an asterisk to denote variables whose truth value pertains to the hypothetical world created by the modification – in this case $\neg B$. So we rewrite S5 as follows:

# Combined theory

$$\begin{array}{lll}
 & & (T) \\
D^* = T & D = T & (D) \\
\neg B^* & B = D & (B) \\
O^* = D^* & O = D & (O) \\
L^* = O^* \vee B^* & L = O \vee B & (L) \\
\hline
\multicolumn{3}{l}{\text{Facts: } L} \\
\hline\hline
\multicolumn{3}{l}{\text{Conclusions: } T, B, O, D, L, \neg B^*, O^*, D^*, L^*}
\end{array}$$

Given $L$, we have $O \vee B$. Since at least one of $O$ or $B$ is true, we must have $D$ and therefore $T$, which exists in both worlds. Therefore $D^*$. Therefore $L^*$, therefore $L^*$ in spite of $\neg B^*$

# Why is S4 ‘action’ and S5 ‘counterfactual’?

- In S4, the fact given (no one opened the door) is not affected by the antecedent (Bastet snuck outside through a window.)
- In S5 we were asking if changing $B$ to $\neg B$ would affect the outcome $L$ vs. $\neg L$. To determine this we had to calculate the potential impact of $\neg B$ on $L$ and route the impact of $\neg B$ through $T$.

# Probabilistic evaluation of counterfactuals

Suppose that . . .

1. There is a probability $P(T) = p$ that the temperature goes above 20C.
2. Bastet has a probability $q$ of sneaking out through a window.
3. Bastet's inclination to sneak out a window is independent of $T$.

We want to compute the probability $P(\neg L_{\neg B}|L)$, the probability that the birds would not have left the feeder if Bastet had not gone outside, given that the birds have in fact left the feeder.

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$
- This comes to the probability that the birds all left under the circumstances that the temperature did not rise above 20C divided by the probability that the birds all left.

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$
- This comes to the probability that the birds all left under the circumstances that the temperature did not rise above 20C divided by the probability that the birds all left.
- The only way that the birds would have all left if the temperature had not gone above 20C is that Bastet snuck out a window. (We are assuming that Oscar is not capable of doing so.)

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$
- This comes to the probability that the birds all left under the circumstances that the temperature did not rise above 20C divided by the probability that the birds all left.
- The only way that the birds would have all left if the temperature had not gone above 20C is that Bastet snuck out a window. (We are assuming that Oscar is not capable of doing so.)
- So the numerator is the probability that Bastet snuck out a window times the probability that the temperature did not rise above 20C (the two events independent) which is $q(1-p)$.

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$
- This comes to the probability that the birds all left under the circumstances that the temperature did not rise above 20C divided by the probability that the birds all left.
- The only way that the birds would have all left if the temperature had not gone above 20C is that Bastet snuck out a window. (We are assuming that Oscar is not capable of doing so.)
- So the numerator is the probability that Bastet snuck out a window times the probability that the temperature did not rise above 20C (the two events independent) which is $q(1-p)$.
- The denominator is 1 minus the probability that the birds are still there, the latter only being possible if the temperature did not rise above 20C and Bastet did not sneak out a window – i.e. $1-(1-p)(1-q)$

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$
- This comes to the probability that the birds all left under the circumstances that the temperature did not rise above 20C divided by the probability that the birds all left.
- The only way that the birds would have all left if the temperature had not gone above 20C is that Bastet snuck out a window. (We are assuming that Oscar is not capable of doing so.)
- So the numerator is the probability that Bastet snuck out a window times the probability that the temperature did not rise above 20C (the two events independent) which is $q(1-p)$.
- The denominator is 1 minus the probability that the birds are still there, the latter only being possible if the temperature did not rise above 20C and Bastet did not sneak out a window – i.e. $1-(1-p)(1-q)$
- So $P(\neg L_{\neg B}|L) = \frac{q(1-p)}{1-(1-p)(1-q)}$.

# Intuitive calculation

- Intuitively, $\neg L_{\neg B}$ is true, given $\neg B$ iff the temperature did not go above 20C. So we want to compute $P(\neg T|L) = \frac{P(\neg T \wedge L)}{P(L)}$
- This comes to the probability that the birds all left under the circumstances that the temperature did not rise above 20C divided by the probability that the birds all left.
- The only way that the birds would have all left if the temperature had not gone above 20C is that Bastet snuck out a window. (We are assuming that Oscar is not capable of doing so.)
- So the numerator is the probability that Bastet snuck out a window times the probability that the temperature did not rise above 20C (the two events independent) which is $q(1-p)$.
- The denominator is 1 minus the probability that the birds are still there, the latter only being possible if the temperature did not rise above 20C and Bastet did not sneak out a window – i.e. $1-(1-p)(1-q)$
- So $P(\neg L_{\neg B}|L) = \frac{q(1-p)}{1-(1-p)(1-q)}$.

# A probabilistic causal model

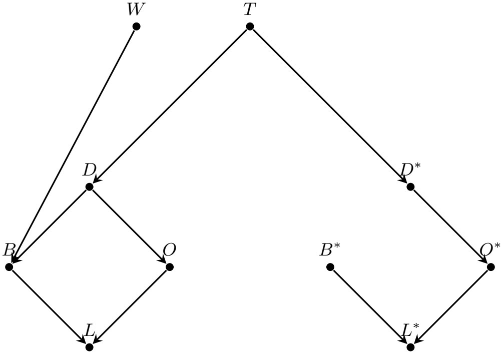
Pearl comments that we can calculate this using a probabilistic causal model using two background variables $T$ (temperature rises above 20C) and $W$ (Bastet decides to go out through a window.)

$$P(t,w) = \begin{cases} pq & \Longleftrightarrow t = 1, w = 1, \\ p(1-q) & \Longleftrightarrow t = 1, w = 0, \\ (1-p)q & \Longleftrightarrow t = 0, w = 1, \\ (1-p)(1-q) & \Longleftrightarrow t = 0, w = 0 \end{cases}$$

We need to first compute the posterior probability $P(t,w|L)$. This can become a problem computationally to compute and store if there are a lot of background variables. And conditioning on some variable $e$ normally destroys the mutual independence of the background variables so that we have to maintain the joint distribution of all the background variables.

# Solution: Balke and Pearl 1994: Twin network graphical model

Two networks: one to represent the actual world and one to represent the hypothetical world.

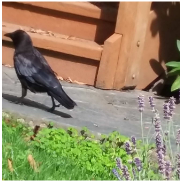


Since we are conditioning on Bastet not going outside, there is no path from $D^*$ to $B^*$.

# A different example

We now look at a different example from Balke and Pearl that illustrates the calculations in more detail.

- There is a crow that sometimes comes to the yard to look for worms but only if it is raining.



- Bastet goes outside if the crow is out there but otherwise almost never goes outside if it is raining.
- Oscar likes going outside as much as possible even if it is raining but, strangely, is afraid of the crow, so avoids going outside if the crow is there.
- If Bastet and Oscar are both outside, one will likely chase the other away. There is also a slight chance that if both are inside, one will chase the other.

# Variables for this example

We have the following variables:

- $C$ The crow is outside or not outside.
- $B$ Bastet is outside or not outside.
- $O$ Oscar is outside or not outside.
- $W$ One of the cats chases the other away.

$$c \in \left\{ \begin{array}{lcl} c_0 & \equiv & \text{The crow is not outside.} \\ c_1 & \equiv & \text{The crow is outside.} \end{array} \right\}$$

$$b \in \left\{ \begin{array}{lcl} b_0 & \equiv & \text{Bastet is not outside.} \\ b_1 & \equiv & \text{Bastet is outside.} \end{array} \right\}$$

$$o \in \left\{ \begin{array}{lcl} o_0 & \equiv & \text{Oscar is not outside.} \\ o_1 & \equiv & \text{Oscar is outside.} \end{array} \right\}$$

$$w \in \left\{ \begin{array}{lcl} w_0 & \equiv & \text{There is no cat chase.} \\ w_1 & \equiv & \text{There is a cat chase.} \end{array} \right\}$$

# A conversation by observers

Imagine the following conversation by observers who notice that Bastet is inside even though it is raining.

A: The crow must not be outside, or Bastet would be there instead of inside.

B: That must mean that Oscar is outside!

A: If Bastet were outside, then Bastet and Oscar would surely chase each other.

B: No. If Bastet was there, then Oscar would not be there, because the crow would have been outside.

A: True. But if Bastet were outside even though the crow was not, then Bastet and Oscar would be chasing each other.

B: I agree.

# Balke & Pearl p. 232

*‘In the fourth sentence, B tries to explain away A’s conclusion by claiming that Bastet’s presence would be evidence that the crow was outside which would imply that Oscar was not outside. B, though, analyzes A’s counterfactual statement as an indicative sentence by imagining that she had observed Bastet’s presence outside; this allows A to use the observation for abductive reasoning. But A’s subjunctive (counterfactual) statement should be interpreted as leaving everything in the past as it was [e.g. that Bastet is inside] (including conclusions obtained from abductive reasoning from real observations [e.g. that the crow must be outside and therefore Oscar must be inside]) while forcing variables to their counterfactual values. This is the gist of A’s last statement.*

# Unknown factors

Suppose that we have the following probabilities:

$p(b_1|c_1) = 0.9$

$p(b_0|c_0) = 0.9$

We observe that neither Bastet nor the crow is outside and ask whether Bastet would be there if the crow were there: $p(b_1^*|\hat{c}_1^*, c_0, b_0)$. The answer depends on what causes Bastet not to go outside even when the crow is there.

We model the influence of $A$ on $B$ by a function: $b = F_b(a, \epsilon_b)$ where $\epsilon$ represents all the unknown factors that could influence $B$ as quantified by the prior distribution $P(\epsilon_b)$. For example, possible components of $\epsilon_b$ could be Bastet being sick or Bastet being sulky about never being able to catch the crow.

# Response function variables

Each value in $\epsilon_b$'s domain specifies a response function that maps each value of A to some value in B's domain.

$r_b$ : domain$(\epsilon_b) \rightarrow$ N:

$$r_b(\epsilon_b) = \begin{cases} 0 \text{ if } F_b(a_0, \epsilon_b) = 0 \;\&\; F_b(a_1, \epsilon_b) = 0 \; (b = b_0 \text{ regardless of a}) \\ 1 \text{ if } F_b(a_0, \epsilon_b) = 0 \;\&\; F_b(a_1, \epsilon_b) = 1 \; (b = b_1 \Longleftrightarrow a = a_1) \\ 2 \text{ if } F_b(a_0, \epsilon_b) = 1 \;\&\; F_b(a_1, \epsilon_b) = 0 \; (b \text{ has opposite value of } a) \\ 3 \text{ if } F_b(a_0, \epsilon_b) = 1 \;\&\; F_b(a_1, \epsilon_b) = 1 \; (b = b_1 \text{ regardless of a}) \end{cases}$$

$r_b$ is a random variable that can take on as many values as there are functions between $a$ and $b$. Balke and Pearl call this a *response function variable.*

# Response functions for this example

Specifically for this example:

$b = f_b(c, r_b) = h_{b,r_b}(c)$

Whether Bastet goes outside or not is a function of whether the crow is there and of the response function that accounts for other factors that can influence Bastet's behaviour. We can also think of a function $h$ of $c$ that returns a value of $b$ given the value of $c$ and the value of the response variable:

| | | |
|---|---|---|
| $h_{b,0}(c) = b_0$ | | Bastet doesn't go outside<br>regardless of whether the crow is there.<br>e.g. Bastet is ill. |
| $h_{b,1}(c) = \begin{cases} b_0 \text{ if } c = c_0 \\ b_1 \text{ if } c = c_1 \end{cases}$ | | Bastet goes outside only if the crow is there. |
| $h_{b,2}(c) = \begin{cases} b_1 \text{ if } c = c_0 \\ b_0 \text{ if } c = c_1 \end{cases}$ | | Bastet goes outside only if the crow isn't there. |
| $h_{b,3}(c) = b_1$ | | Bastet goes outside regardless. |

# An example counterfactual

If we have the prior probability $P(r_b)$ we can calculate $P(b_1^*|\hat{c}_1^*, c_0, b_0)$: i.e. ‘Given that the crow is not outside and Bastet is not outside, if the crow were outside, what is the probability that Bastet would be outside?’

We crucially assume that:

> *‘…the disturbance $\epsilon_b$, and hence the response-function $r_b$, is unaffected by the actions that force the counterfactual values; therefore, what we learn about the response-function from the observed evidence is applicable to the evaluation of belief in the counterfactual consequent.’*

If we observe $(c_0, b_0)$ (neither Bastet nor the crow is outside), then it must be that $r_b \in \{0, 1\}$, an event with prior probability $P(r_b = 0) + P(r_b = 1)$. This updates the posterior probability of $r_b$ as follows, letting $\vec{P}(r_b) = \langle P(r_b = 0), P(r_b = 1), P(r_b = 2), P(r_b = 3)\rangle$:

$$
\begin{aligned}
\vec{P}(r_b) &= \vec{P}(r_b|c_0, b_0) \\
&= \left\langle \frac{P(r_b = 0)}{P(r_b = 0) + P(r_b) = 1}, \frac{P(r_b = 1)}{P(r_b = 0) + P(r_b) = 1}, 0, 0 \right\rangle
\end{aligned}
$$

# Calculating this counterfactual

From the definition of $r_b(\epsilon_b)$ above, if C were forced to $c_1$ (the crow is outside), then B would have been $b_1$ (Bastet would have also been outside iff $r_b \in \{1, 3\}$, whose probability is $P'(r_b = 1) + P'(r_b = 3) = P'(r_b = 1)$. ($P'(r_b = 3)$ must be zero since we have determined that $r_b \in \{0, 1\}$.) This gives the solution to the counterfactual query:

$P(b_1^* | \hat{c}_1^*, c_0, b_0) = P'(r_b = 1) = \frac{P(r_b=1)}{P(r_b=0)+P(r_b=1)}$

The probability of external influence 1 that causes Bastet to go outside if the crow is there divided by the probability of external influence 1 plus exernal influence 0, the latter causing Bastet to stay inside regardless.

# Representation with a DAG

We can represent the causal influences over a set of variables in this example through a DAG. If the set of variables is $\{X_1, X_2, \dots X_n\}$, for each $x_i$ there is a functional mapping $x_i = f_i(\mathrm{pa}(x_i), \epsilon_i)$, where $\mathrm{pa}(x_i)$ is the value of $X_i$'s parents in the graph and there is a prior probability distribution $P(\epsilon_i)$ for each 'disturbance' $\epsilon_i$.

A counterfactual query will be of the form: 'What is $P(c_* | \hat{a}^*, obs)$, where $c^*$ is a set of counterfactual values for $C \subset X$, $\hat{a}^*$ is a set of forced values in the counterfactual antecedent and *obs* represents observed evidence.

For our example, we assume that Bastet is not outside ($b = b_0$) and want to ask 'what is $P(c_1^* | \hat{b}_1^*, b_0)$?' Or further, what is the probability that the cats will chase each other under those conditions?

# A possible causal theory with response variables

Suppose that we have the following of what Balke and Pearl call a ‘causal theory’:

$$\begin{array}{llcll}
c = & f_c(r_c) & = & h_{c,r_c}() & \text{(crow's presence depends only on } r_c\text{)} \\
b = & f_b(c, r_b) & = & h_{c,r_b}(c) & \text{(Bastet's presence depends on } r_b,\ \text{crow)} \\
o = & f_o(c, r_o) & = & h_{c,r_c}(c) & \text{(Oscar's presence depends on } r_o,\ \text{crow)} \\
w = & f_w(b, o, r_w) & = & h_{w,r_w}(b, o) & \text{(chase depends on } r_w,\ \text{Bastet and Oscar)}
\end{array}$$

$$P(r_c) = \begin{cases} 0.40 & \text{if} \quad r_c = 0 \\ 0.60 & \text{if} \quad r_c = 1 \end{cases} \qquad \text{(60\% chance crow is there)}$$

$$P(r_b) = \begin{cases} 0.07 & \text{if} \quad r_b = 0 \\ 0.90 & \text{if} \quad r_b = 1 \\ 0.03 & \text{if} \quad r_b = 2 \\ 0 & \text{if} \quad r_b = 3 \end{cases} \qquad \text{(90\% chance Bastet there if crow is)}$$

$$P(r_o) = \begin{cases} 0.05 & \text{if} \quad r_o = 0 \\ 0 & \text{if} \quad r_o = 1 \\ 0.85 & \text{if} \quad r_o = 2 \\ 0.10 & \text{if} \quad r_o = 3 \end{cases} \qquad \text{(85\% chance Oscar there if crow isn't)}$$

$$P(r_w) = \begin{cases} 0.05 & \text{if} \quad r_w = 0 \\ 0.90 & \text{if} \quad r_w = 8 \\ 0.05 & \text{if} \quad r_w = 9 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(90\% chance chase if B \& O there)}$$

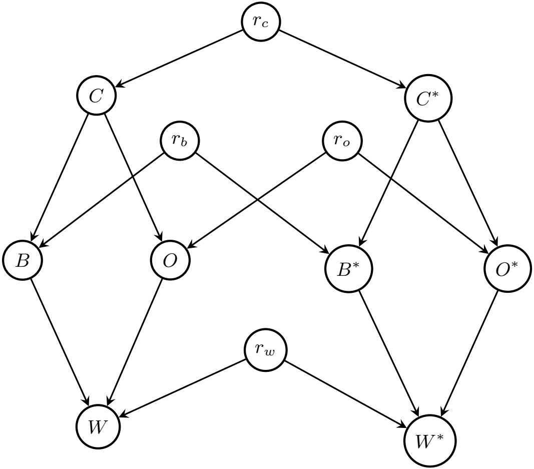
$h_{c,0}() = c_0$ (if $r_c = 0$ the crow is not there)
$h_{c,1}() = c_1$ (if $r_c = 1$ the crow is there)

$h_{w,0}(b,o) = s_0$ (if $r_s = 0$, there is no chase regardless)

$$h_{w,8}(b,o) = \begin{cases} s_0 \quad \text{if} \quad (b,o) \neq (b_1,o_1) & \text{no chase unless both B,O outside} \\ s_1 \quad \text{if} \quad (b,o) = (b_1,o_1) & \text{chase if both B,O outside} \end{cases}$$

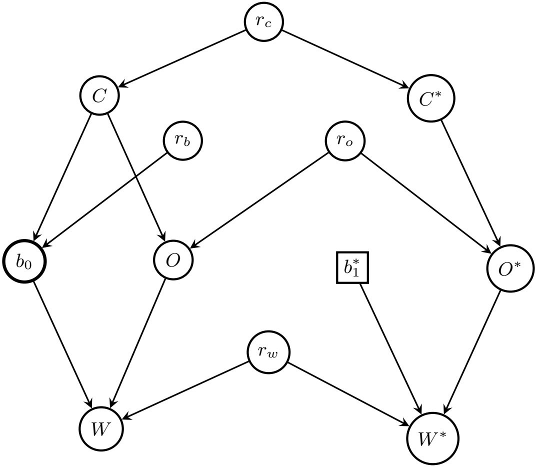
$$h_{w,9}(b,o) = \begin{cases} s_0 \text{ if } (b,o) \in \{(b_1,o_0),(b_0,o_1)\} & \text{no chase if 1 cat present} \\ s_1 \text{ if } (b,o) \in \{(b_0,o_0),(b_1,o_1)\} & \text{chase if B,O meet in or out} \end{cases}$$

# DAG



Variables marked with $*$ indicate the counterfactual world and those without the factual world. The $r$ variables are response functions.

# DAG for counterfactual evaluation



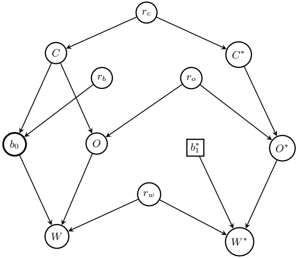
To evaluate $P(w_1^*|\hat{b}_1^*, b_0)$, instantiate $B$ as $b_0$ and $B^*$ as $b_1^*$. Sever links pointing to $b_1^*$

# Evaluating $W^*$

Balke and Pearl comment:

> *If a variable $X_j^*$ in the counterfactual world is not a causal descendant of any of the variables mentioned in the counterfactual antecedent $\hat{a}^*$, then $X_j$ and $X_j^*$ will always have identical distributions, because the causal influences that functionally determine $X_j$ and $X_j^*$ are identical.*

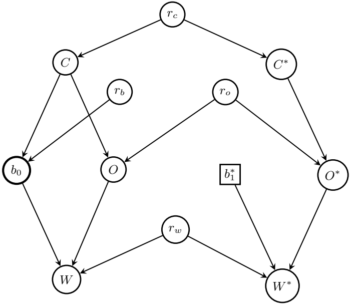
To evaluate $W^*$, we can start by looking at the graph in the factual world to see what values of parents of $b_0$ could lead to that value. We consider all the possible combinations of values of parents of $b_0$. The probability of each combination is the product of their probabilitites and the total prior probability of $b_0$ is the sum of probabilities of combinations that result in $B = b_0$.



# Evaluating $W^*$

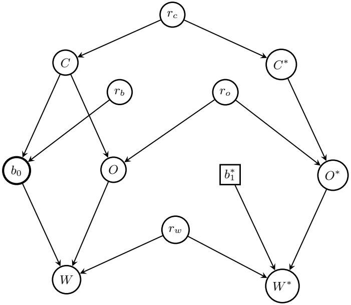
- $r_c = 0$ (0.4) and $r_b = 0$ (0.07) → $C = c_0$ and $B = b_0$ (0.028)
- $r_c = 0$ (0.4) and $r_b = 1$ (0.90) → $C = c_0$ and $B = b_0$ (0.36)
- $r_c = 0$ (0.4) and $r_b = 2$ (0.03) → $C = c_0$ and $B = b_1$ (0.012)
- $r_c = 1$ (0.6) and $r_b = 0$ (0.07) → $C = c_1$ and $B = b_0$ (0.042)
- $r_c = 1$ (0.6) and $r_b = 1$ (0.90) → $C = c_1$ and $B = b_1$ (0.54)
- $r_c = 1$ (0.6) and $r_b = 2$ (0.03) → $C = c_1$ and $B = b_0$ (0.018)

The prior probability $P(B = b_0) = 0.028 + 0.36 + 0.042 + 0.018 = 0.448$. So $p(C = c_0 | B = b_0) = \frac{0.028+0.36}{0.448} = 0.86607$



# Evaluating $W^*$

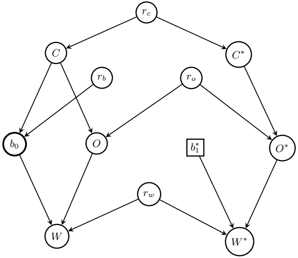
Given that $C^*$ is not a causal descendant of any $B$ variables, we can give counterfactual $C^*$ the same probability as $C$. We can now work down on the counterfactual side of the DAG and calculate $O^*$. We calculate the probability of each possible combination of values of $r_o$ and $C^*$ and determine for each the value of $O^*$ that results. We add to get the total probability of $o_1^*$.



# Evaluating $W^*$

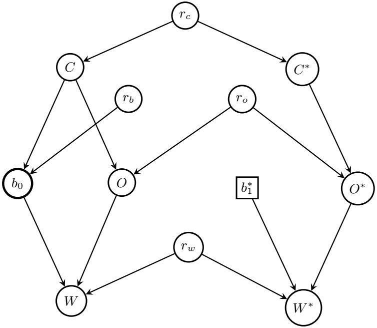
- $r_o = 0$ (0.05) and $C^* = c_0^*$ (0.86607) $\rightarrow O^* = o_0^*$ (0.043304)
- $r_o = 0$ (0.05) and $C^* = c_1^*$ (0.13393) $\rightarrow O^* = o_0^*$ (0.0066965)
- $r_o = 1$ (0) and $C^* = c_0^*$ (0.86607) $\rightarrow O^* = o_0^*$ (0)
- $r_o = 1$ (0) and $C^* = c_1^*$ (0.13393) $\rightarrow O^* = o_1^*$ (0)
- $r_o = 2$ (0.85) and $C^* = c_0^*$ (0.86607) $\rightarrow O^* = o_1^*$ (0.73616)
- $r_o = 2$ (0.85) and $C^* = c_1^*$ (0.13393) $\rightarrow O^* = o_0^*$ (0.1138405)
- $r_o = 3$ (0.10) and $C^* = c_0^*$ (0.86607) $\rightarrow O^* = o_1^*$ (0.086607)
- $r_o = 3$ (0.10) and $C^* = c_1^*$ (0.13393) $\rightarrow O^* = o_1^*$ (0.013393)

$P(O^* = o_1^*) = 0 + 0.73616 + 0.086607 + 0.013393 = 0.83616$



# Evaluating $W^*$

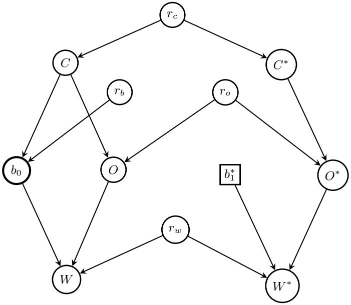
Given that $P(b_0^*) = 1$, we can now calculate $P(W^* = 1|b_0^*, O^*)$, moving further down the graph. We look at all the possible combinations of possible values of parents of $W^*$. Since $B^*$ is set at $b_1^*$, we need not include it in the set of combinations.



# Evaluating $W^*$

- $r_w = 0$ (0.05) and $O^* = o_0^*$ (0.16184) → $W^* = w_0^*$ (0.008092)
- $r_w = 0$ (0.05) and $O^* = o_1^*$ (0.83616) → $W^* = w_0^*$ (0.041808)
- $r_w = 8$ (0.90) and $O^* = o_0^*$ (0.16184) → $W^* = w_0^*$ (0.145656)
- $r_w = 8$ (0.90) and $O^* = o_1^*$ (0.83616) → $W^* = w_1^*$ (0.752544)
- $r_w = 9$ (0.05) and $O^* = o_0^*$ (0.16184) → $W^* = w_0^*$ (0.008092)
- $r_w = 9$ (0.05) and $O^* = o_1^*$ (0.83616) → $W^* = w_1^*$ (0.041808)

So $P(W^* = 1|b_0^*, O^*) = 0.75254 + 0.041808 = 0.79435$, which to two decimal places is the value given by Balke and Pearl.



# References

Pearl, Judea. 2000. *Causality* Chapter 7: The logic of structure-based counterfactuals. Cambridge University Press.

Balke, Alexander and Judea Pearl. 1994. Probabilistic evaluation of counterfactual queries. In proceedings of AAAI.

# Johansson et al (2016) *Learning representations for counterfactual inference*

Suppose that we have data on 1000 patients who either did or did not receive some medical treatment and we know some quantity that represents the results of the treatment such as blood sugar level or blood pressure. For each patient we know whether or not they received the treatment and what the result was.

What we do not know is what the characteristics were for each patient on which the decision to give the treatment or not was based. We also do not know what the results would have been if the opposite course of action had been taken for each patient – no treatment for those who received it and vice versa.

- For each patient $x$, $Y_1(x)$ is the measurable result of $x$ receiving the treatment and $Y_0(x)$ the result of $x$ not receiving the treatment.
- For each $x$ we know only one of the two.
- For each $x$ the individualized treatment effect ITE $= Y_1(x) - Y_0(x)$ is a quantity that would enable us to choose the best choice for each patient, but we do not know this quantity.
- If $p(x)$ is the distribution under which $x$ values occur, the average treatment effect ATE is defined as ATE$=\mathbb{E}_{x \sim p(x)}[ITE(x)]$.
- We also define the factual and counterfactual outcomes of the treatment given to individual $x$ as $y^F(x)$ and $y^{CF}(x)$ respectively.

What they call *direct modeling* for estimating ITE, works as follows.

- We have $n$ samples: $\{(x_i, t_i, y_i^F)\}_{i=1}^n$,
- where $x_i$ is the individual,
- $t_i$ is the binary value for treatment (1) or no treatment (1)
- and $y_i^F$ is the factual outcome.

- We can see that $y_i^F = t_i \cdot Y_1(x_i) + (1 - t_i) \cdot Y_0(x_i)$ where $Y_0$ is the function applied to $x$ of not giving the treatment and $Y_1$ is giving the treatment.
- Either $t_i$ or $1 - t_i$ must be zero, so one of the above two terms will be zero.
- If the treatment was given, $t_i = 1$ and the second term drops out;
- if not, $t_i = 0$ and the first term drops out.
- We want to learn a function $h$ that maps an individual $x_i$ paired with a treatment $t_i$ to an outcome $y_i^F$:

  $h : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{Y}$ such that $h(x_i, t_i) \approx y_i^F$

We can then use $h$ to represent the half of the ITE that we don't know. If the treatment was given, $t_i = 1$ and $\hat{\text{ITE}}(x) = y_i^F - h(x_i, 1 - t_i)$. If not, $t_i = 0$ and $\hat{\text{ITE}}(x) = h(x_i, 1 - t_i) - y_i^F$.

The observed sample is $\hat{P}^F = \{(x_i, t_i)\}_{i=1}^n$: ‘the factual distribution’.

The counterfactual distribution is $\hat{P}^{CF} = \{(x_i, 1 - t_i)\}_{i=1}^n$.

Crucially, the two distributions may be different – i.e. if the decision how to treat a given patient was not random, but based on some characteristics of that patient, $t$ and $x$ will not be independent. And we don’t have the information about how decisions about treatment assignment were made.

We have:
$P^F(x,t) = P(x) \cdot P(t|x)$
$P^{CF}(x,t) = P(x) \cdot P(\neg t|x)$

We assume that we have some information about the patients so that we can given them relevant features, even though we don't know if these were the same features that were used to decide on the treatment nor how the features might have been taken into account.

We want to learn two things:

A representation $\Phi$ that maps each patient onto a feature set and a function $h$ that gives a numerical outcome for each patient feature set paired with a treatment.

The represetnatation trades off three objectives:

1. low-error prediction of the observed outcomes over the factual representation
2. low-error prediction of unobserved counterfactuals by taking into account relevant factual outcomes
3. the distributions of treatment populations are similar or balanced

How do we accomplish each?

1. error-minimization over a training set using regularization to avoid overfitting
2. assign a penalty that encourages counterfactual predictions to be close to the nearest observed outcome from the respective treated or control set
3. minimize the *discrepancy distance* between treated and control populations

$$\begin{aligned} \text{Patients: } & X = \{x_i\}_{i=1}^{n} \\ \text{Treatments: } & T = \{t_i\}_{i==1}^{n} \\ \text{Factual outcomes: } & Y^F = \{y_i^F\}_{i=1}^{n} \end{aligned}$$

We define the nearest oppositely-treated neighbour of patient $x_i$, $x_j$, where $j \in \{1, \dots b\}$ and $t_j = 1 - t_i$ such that $j = \arg\min d(x_i, x_j)$ where $d(a, b)$ is the distance from $a$ to $b$ in the metric space $\mathcal{X}$.

We want to minimize the objective $B_{\mathcal{H},\alpha,\gamma}(\Phi, h)$ where $\mathcal{H}$ is the hypothesis space, $\alpha, \gamma > 0$ are hyperparameters that control the strength of imbalance penalties, $\Phi$ is the representation that maps each patient onto a feature set and $h$ is a function $h$ that gives a numerical outcome for each patient feature set paired with a treatment.

$B_{\mathcal{H},\alpha,\gamma}(\Phi, h)$ is the sum of three quantities:

- $\frac{1}{n}\sum_{i=1}^{n} |h(\Phi(x_i), t_i) - y_i^F|$ – the sum for all the patient-treatment pairs of the differences between each treatment result that our hypothesis predicts and the actual treament result
- $\alpha \text{disc}_{\mathcal{H}}(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF})$ – the discrepancy between the factual and counterfactual distributions, weighted by $\alpha$
- $\frac{\gamma}{n}\sum_{i=1}^{n} |h(\Phi(x_i), 1 - t_i) - y_i^F|$ – the sum, weighted by $\gamma$ for all the patient-treatment pairs of the differences between each treatment result that our hypothesis predicts and the result of the treatment that was *not* chosen for a given patient.

Once again:

- $\Phi$ is a representation of $\mathcal{X}$ that maps each patient onto a vector of features
- and $\mathcal{H}$ represents the set of possible hypothesis functions $h$ each of which maps each feature vector and treatment choice for a patient onto a real number that represents the hypothesized results of that treatment for that patient.
- The paper states that if the hypothesis class $\mathcal{H}$ is the set of linear functions, the term $\text{disc}_{\mathcal{H}}(\hat{P}_{\Phi}^{F}, \hat{P}_{\Phi}^{CF})$ has closed form $||\mu_2(P) - \mu_2(Q)||_2$ where $||A||_2$ is the spectral norm of $A$ and $\mu_2(P) = \mathbb{E}_{x \sim P}[xx^T]$ is the variance of the distribution $P$.
- If $x$ is a binary-valued matrix whose rows are patients and columns are features we have assigned to each patient,then $xx^T$ will represent the degree to which each pair of patients differs with respect to the binary values we have assigned to them. The more they differ, the lower the value in the cell of $xx^T_{i,j}$ for patients $i$ and $j$.

In the case here of counterfactual inference, $P$ and $Q$ differ only in what treatment was given.

Let $v$ be a vector that represents the difference, for each patient $x$, between the expected results of getting the treatment and the expected results of not getting the treatment, based on the representation $\Phi$ that we have given each patient. In other words:

$$v = \mathbb{E}_{(x,t)\sim \hat{P}^F_\Phi}[\Phi(x) \cdot t] - \mathbb{E}_{(x,t)\sim \hat{P}^F_\Phi}[\Phi(x) \cdot (1-t)]$$

And let $p$ represent the expected treatment for each patient.

So $\mathrm{disc}_{\mathcal{H}}(\hat{P}_{\Phi}^{F}, \hat{P}_{\Phi}^{CF})$ is the spectral norm of $\mathbb{E}_{x\sim\hat{P}_{\Phi}^{F}}[xx^T] - \mathbb{E}_{x\sim\hat{P}_{\Phi}^{CF}}[xx^T]$, that difference being the difference between the expected amount by which each pair of patients differs under the factual distribution and the expected amount by which each pair of patients differs under the counterfactual distribution.