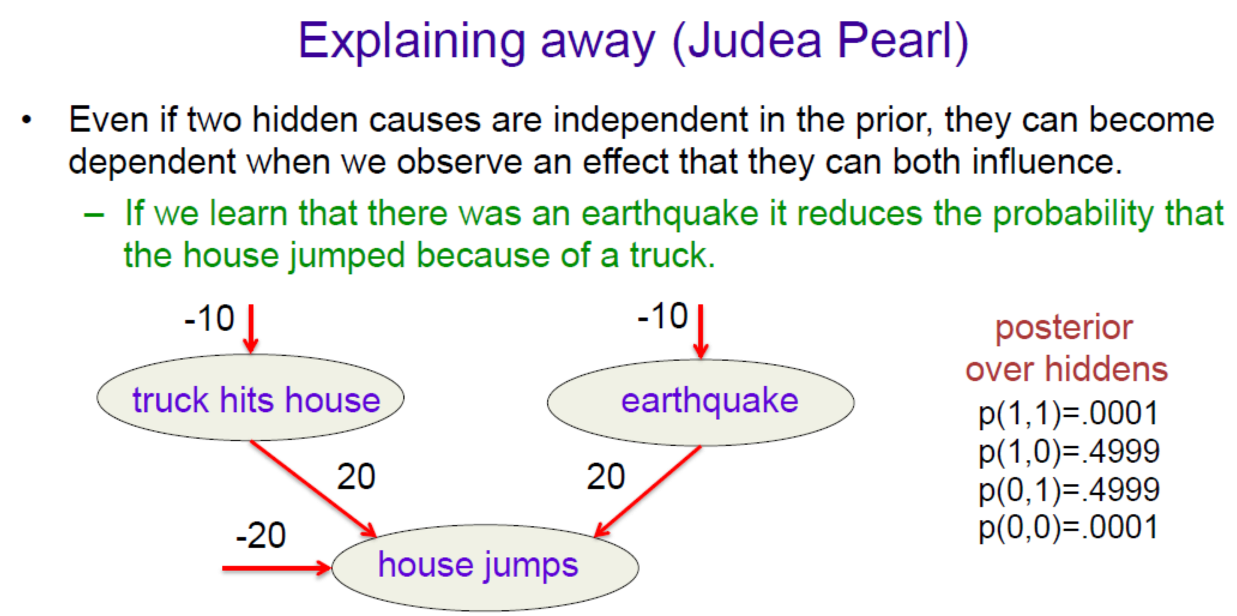
Explaining Away Makes Belief Network Hard to Train

Lei Mao

7/7/2017

The following examples are taken from Hinton's Coursera course "Neural Networks for Machine Learning" and the explanations are based on his introductions to belief network in the related course materials.



$P(\text{truck} = \text{True}) = 1/(1+e^{10}) = 0.00004539786$

$P(\text{earthquake} = \text{True}) = 1/(1+e^{10}) = 0.00004539786$

$P(\text{house} = \text{True} \mid \text{truck} = \text{True}, \text{earthquake} = \text{True}) = 1/(1+e^{(-20)}) = 0.99999999793$

$P(\text{house} = \text{True}, \text{truck} = \text{True}, \text{earthquake} = \text{True}) = P(\text{house} = \text{True} \mid \text{truck} = \text{True}, \text{earthquake} = \text{True}) * P(\text{truck} = \text{True}, \text{earthquake} = \text{True}) = P(\text{house} = \text{True} \mid \text{truck} = \text{True}, \text{earthquake} = \text{True}) * P(\text{truck} = \text{True}) * P(\text{earthquake} = \text{True}) = 2.06096569\text{e-}9$

$P(\text{house} = \text{True} \mid \text{truck} = \text{False}, \text{earthquake} = \text{True}) = 1/(1+e^{0}) = 0.5$

$P(\text{house} = \text{True}, \text{truck} = \text{False}, \text{earthquake} = \text{True}) = P(\text{house} = \text{True} \mid \text{truck} = \text{False}, \text{earthquake} = \text{True}) * P(\text{truck} = \text{False}, \text{earthquake} = \text{True}) = P(\text{house} = \text{True} \mid \text{truck} = \text{False}, \text{earthquake} = \text{True}) * P(\text{truck} = \text{False}) * P(\text{earthquake} = \text{True}) = 0.00002269789$

$P(\text{house} = \text{True} \mid \text{truck} = \text{True}, \text{earthquake} = \text{False}) = 1/(1+e^{0}) = 0.5$

$P(\text{house} = \text{True}, \text{truck} = \text{True}, \text{earthquake} = \text{False}) = 0.00002269789$

$P(\text{house} = \text{True} \mid \text{truck} = \text{False}, \text{earthquake} = \text{False}) = 1/(1+e^{20}) = 2.06115362\text{e-}9$

P(house = True, truck = False, earthquake = False) = P(house = True | truck = False, earthquake = False) * P(truck = False, earthquake = False) = P(house = True | truck = False, earthquake = False) * P(truck = False) * P(earthquake = False) = 2.06096648e-9

To calculate the marginal probability, P(house = True):

P(house = True) = P(house = True, truck = True, earthquake = True) + P(house = True, truck = False, earthquake = True) + P(house = True, truck = True, earthquake = False) + P(house = True, truck = False, earthquake = False) = 0.0000453999

To calculate the posteriors (for sampling purpose), apply Bayes rule:

P(truck = True, earthquake = True | house = True) = P(house = True | truck = True, earthquake = True) * P(truck = True, earthquake = True) / P(house = True) = 0.00004539582

P(truck = True, earthquake = False | house = True) = P(house = True | truck = True, earthquake = False) * P(truck = True, earthquake = False) / P(house = True) = 0.49995462545

P(truck = False, earthquake = True | house = True) = P(house = True | truck = False, earthquake = True) * P(truck = False, earthquake = True) / P(house = True) = 0.49995462545

P(truck = False, earthquake = False | house = False) = P(house = True | truck = False, earthquake = True) * P(truck = True, earthquake = False) / P(house = True) = 0.00004539583

With these samples, we can now sample the hidden node. However, what if the number of nodes in the hidden layer get larger to a big number of n. Say, given house = True, the number of posteriors we need to calculate is 2^n. This makes the inference intractable and actually NP-hard.

P(truck = True | house = True) = P(truck = True, earthquake = True | house = True) + P(truck = True, earthquake = False | house = True) = 0.5

P(earthquake = True | house = True) = P(truck = True, earthquake = True | house = True) + P(truck = False, earthquake = True | house = True) = 0.5

P(truck = True | house = True, earthquake = True) = P(truck = True, earthquake = True | house = True) / P(earthquake = True | house = True) = 0.00009079164

(Apply P(B|C)*P(A|C,B) = P(A,B|C), i.e., P(A|C,B) = P(A,B|C)/P(B|C))

Given house = True, knowing earthquake = True, P(truck = True | house = True, earthquake = True) is significantly smaller than P(truck = True | house = True). This is called “explaining away”. This causes truck and earthquake not conditionally independent of house and makes the calculation of posteriors no easy. Note that here P(truck = True | house = True) * P(earthquake = True | house = True) = 0.25. P(truck

= True, earthquake = True | house = True) = 0.00004539582. P(truck = True, earthquake = True | house = True) != P(truck = True | house = True) * P(earthquake = True | house = True).

Let me say it in another way. If A and B are conditionally dependent of C, one can calculate the posteriors easily using equation $P(A,B|C) = P(B|C)*P(A|C)$ (the posterior is factorial). The problem might not be NP-hard because we can re-use terms to calculate the posteriors. However, this is not true given the diagram here (truck and earthquake are not conditionally dependent of house, although they are independent to each other). $P(A|C,B) \neq P(A|C)$ ($P(A|C,B) << P(A|C)$). Therefore, “explaining away” is one of the reasons that cause belief network extremely hard to learn.