BITS F464: Machine Learning

22 Graphical Model: Bayesian Belief Networks

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

March 15, 2021 ONLINE (Campus @ BITS-Pilani Jan-May 2021)

http://ktiwari.in/ml

Bayesian Learning

Addresses most probable classification of new instance instead of best hypothesis for data. Searching a possibility to do better then MAP

- **Bayes optimal classification:** $\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$
  Outperforms on an average but, quite costly to apply
- **GIBBS Algorithm:** Choose a hypothesis $h \in H$ at random, according to the posterior probability distribution over H.
  (Expected misclassification error is bounded to the twice of the Bayes optimal classifier)
- **Naive Bayes Classifier:** assumes independence given the target value $\text{argmax}_{v_j \in V} P(a_1, a_2, ..., a_n|v_j)P(v_j)$

$$\underset{v_j \in V}{\text{argmax}}\, P(v_j)\Pi_i P(a_i|v_j)$$

Highly practical method

Probability

- $P(x, y) = P(x) \times P(y|x)$
- **Independence** of x and y implies $P(y|x) = P(y)$
  Then $\boxed{P(x, y) = P(x) \times P(y)}$
- Bayes Rule

$$P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(y|x) \times P(x)}{P(y)}$$

- **Marginal:** distribution of a single variable x can be obtained from a given joint distribution $p(x, y)$ by

$$p(x) = \sum_y p(x, y)$$

- The process of computing a marginal from a joint distribution is called marginalisation.

$$p(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = \sum_{x_i} p(x_1, x_2, ..., x_n)$$

Marginalisation

Conditional Independence when two variable are independent of each other, provided that we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: C as soft XOR

A	B	$p(C{=}1\|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0) = \sum_B p(A = 1, B, C = 0)$
  $= \sum_B p(C = 0|A = 1, B)p(A = 1)p(B)$
  $= p(C = 0|A = 1, B = 0)p(A = 1)p(B = 0)$
  $+p(C = 0|A = 1, B = 1)p(A = 1)p(B = 1)$
  $= 0.2 \times 0.65 \times 0.23 + 0.75 \times 0.65 \times 0.77$
  $= \boxed{0.405}$
- Similarly $p(A{=}0, C{=}0) = 0.075$
- $p(A{=}1|C{=}0) = \frac{p(A=1, C=0)}{p(C=0)} = \frac{p(A=1, C=0)}{p(A=1, C=0)+p(A=0, C=0)}$
  $= \boxed{0.843}$

Let's see this

Statement:

$$\sum_J (p(J|R) \times f(R)) = f(R)$$

Proof:

$$\begin{aligned}
\sum_J (p(J|R) \times f(R)) &= \sum_J \left(\frac{p(J, R)}{p(R)} \times f(R)\right) \\
&= \frac{\sum_J (p(J, R) \times f(R))}{p(R)} \\
&= \frac{f(R) \times \sum_J p(J, R)}{p(R)} \\
&= \frac{f(R) \times p(R)}{p(R)} \\
&= f(R)
\end{aligned}$$

Belief network (a graphical model)

- **Belief network** uses graphs to represent independence among the variables in probabilistic model
- Independently specifying all the attributed is overkill
- With distribution of n attributes, marginal for one takes $O(2^{n-1})$
- By constraining variable interaction (specifying independence) one can get the form like

$$p(x_1, x_2, ..., x_{100}) = \Pi_{i=1}^{99} \phi(x_i, x_{i+1})$$

- Belief networks are a convenient framework for representing such independence assumptions
- Belief networks are also called as *Bayes' Networks* or *Bayesian Belief Networks*

Modeling Independencies

One morning Tracey leaves her house and realities that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night? Next she notices that the grass of her neighbor, Jack, is also wet.

(R=1) → rain last night,
(S=1) → sprinkler on last night,
(J=1) → Jack's grass is wet,
(T=1) → Traceya's Grass is wet

- Model of Traceya's world involves probability distribution on T, J, R, S that has $2^4 = 16$ states

Modeling Independencies (contd..)

- However we know

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned}$$

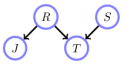
- Computation of $p(T|J, R, S)$ requires us to specify $2^3 = 8$ values
- With $p(T = 1|J, R, S)$, one can use normalization to compute $p(T = 0|J, R, S)$ as $1 - p(T = 1|J, R, S)$
- Computation of other factors would also need 4+2+1 values
- Total we need 8+4+2+1=15 values

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J, R, S) = p(T|R, S)$
- Assume that Jack's grass is wet is influenced only directly by whether or not it has been raining $p(J|R, S) = p(J|R)$
- Furthermore, we assume the rain is not directly influenced by the sprinkler $p(R|S) = p(R)$
- Therefore, our model becomes

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \\ &= p(T|R, S)p(J|R)p(R)p(S) \end{aligned}$$

- Number of values we need to specify is 4+2+1+1=8
- We can represent these conditional independence as



Belief network

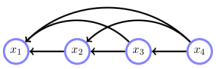
Answers "how to represent these conditional independence?"

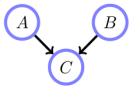
- **Belief network** is a distribution of the form

$$p(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} p(x_i|pa(x_i))$$

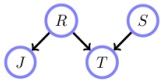
where $pa(x_i)$ represent the parental variables of variable x_i

- Represented as a directed graph, with an arrow pointing from a parent variable to child variable, a belief network corresponds to a Directed Acyclic Graph (DAG)





$p(A, B, C) = p(C|A, B)p(A)p(B)$

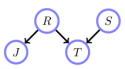


Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be

$p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$, $p(T{=}1|R{=}0, S{=}0) = 0$

Using following Beleif Network; calculate



1. Probability that sprinkler was *ON* overnight, given that Traceya's grass is wet. $p(S{=}1|T{=}1) = 0.3382$ How? on next slide
2. Probability that sprinkler was *ON* overnight, given that Traceya's grass is wet and Jack's grass is also wet. $p(S{=}1|T{=}1, J{=}1) = 0.1604$

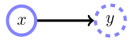
Example: Probability of $p(S{=}1|T{=}1)$

$$\begin{aligned} p(S = 1|T = 1) &= \frac{p(S = 1, T = 1)}{p(T = 1)} \\ &= \frac{\sum_{J,R} p(S = 1, J, R, T = 1)}{\sum_{J,R,S} p(T = 1, J, R, S)} \quad (1) \\ &= \frac{\sum_{J,R} p(J|R)p(T = 1|R, S = 1)p(R)p(S = 1)}{\sum_{J,R,S} p(J|R)p(T = 1|R, S)p(R)p(S)} \\ &= \frac{\sum_{R} p(T = 1|R, S = 1)p(R)p(S = 1)}{\sum_{R,S} p(T = 1|R, S)p(R)p(S)} \quad (2) \\ &= \frac{0.9 \times 0.8 \times 0.1 + 1 \times 0.2 \times 0.1}{.9 \times .8 \times .1 + 1 \times .2 \times .1 + 0 \times .8 \times .9 + 1 \times .2 \times .9} \\ &= \boxed{0.3382} \end{aligned}$$

Uses given belief network in (1) and proof in (2)

Uncertain evidence

- **Soft or uncertain evidence**, let $dom(x) = \{red, blue, green\}$ and the vector $\tilde{y} = (0.6, 0.1, 0.3)$ represents the belief in the respective states. **Hard evidence** are like (0, 0, 1).
- Assumption is that $p(x|y, \tilde{y}) = p(x|y)$
- $p(x|\tilde{y}) = \sum_y p(x, y|\tilde{y}) = \sum_y p(x|y, \tilde{y})p(y|\tilde{y}) = \sum_y p(x|y)p(y|\tilde{y})$ where $p(y = i|\tilde{y})$ represents the probability that y is in state i
- Dashed circle is used to represent a variable in soft-evidence state

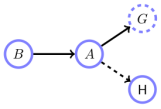


Example: Let probability of fire, when there is a fire alarm is 0.9. Monu said he is 70% confident that he had heard a fire alarm. What is probability of fire.

$p(F = 1|\tilde{A}) = \sum_A p(F = 1|A)p(A|\tilde{A}) = p(F = 1|A = 0)p(A = 0|\tilde{A}) + p(F = 1|A = 1)p(A = 1|\tilde{A}) = 0.1 \times 0.3 + 0.9 \times 0.7 = 0.66$

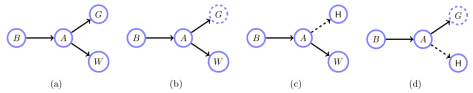
Unreliable Evidence

- **Unreliable:** (continued from previous example) I asked Raju about the alarm. It is believed that if alarm had sound, there is 80% chance that Raju would tell it sound. If alarm had NOT sounded, there is 70% chance that he would tell NOT sound.
- Unreliable evidences are modeled using dashed lines as below



Example

- Let B=1, means Holmes house has been burgled. A=1, means alarm went off. W=1, means Watson heard alarm. G=1, means Gibbon heard alarm.



(a) BN for the environment, (b) If Gibbon is a little def and is only 80% sure about the alarm sound being heard, (c) Replacement of evidence, (d) Holmes feels Watson's observation is unreliable

Thank You!

Thank you very much for your attention!

Queries ?

(Reference[1])

[1] [1] Text Book: ch:1/2/3 *Bayesian Reasoning and Machine Learning*, by *David Barber*