Artificial Intelligence

Week 10: Reasoning under Uncertainty

COMP30024

May 17, 2021

Bayes' Theorem

- Probability quantifies uncertainty when attempting to draw conclusions.
- *Bayes' Theorem* allows us to combine prior information with current information from data to update our uncertainty.
- Want to draw conclusions about unobserved θ based on observed data y.
- Want to find probability distribution of θ conditioned on observed data $\rightarrow p(\theta|y)$

Bayes' Theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- Likelihood $p(y|\theta)$: conditional probability of the data y given fixed θ.
- Prior $p(\theta)$: information we have, not part of the collected data y.
- Evidence $p(y)$: average value of likelihood under prior $p(\theta)$:

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta)$$

- $p(\theta|y)$ is the posterior.
  - Represents updated beliefs about θ now after observation of data y.

Bayes' Theorem

- Alternatively, observe the effect of some unknown cause. Wish to determine the cause:

$$p(\text{Cause}|\text{Effect}) = \frac{p(\text{Effect}|\text{Cause})p(\text{Cause})}{p(\text{Effect})}$$

  - The likelihood $p(\text{Effect}|\text{Cause})$ describes the relationship in the causal direction.
  - Computing the posterior $p(\text{Cause}|\text{Effect})$ allows us to *diagnose potential causes*.

Bayes' Theorem

- Given a set of hypotheses $\{H_1, \dots H_n\}$, corresponding to different values of θ - $\{\theta_1, \dots, \theta_n\}$.
- Want to find most likely hypothesis given observed data y.
- Compare pairs of hypotheses H, H' via ratio of posterior density at different points θ, θ'.

$$\frac{p(\theta|y)}{p(\theta'|y)} = \frac{p(\theta)p(y|\theta)}{p(\theta')p(y|\theta')}$$

- Using ratios avoids calculation of evidence $p(y)$ (difficult to do).

Hints

- Q2: You are given $p(\text{Test} = +|L)$ and want to find:

$$p(L|\text{Test} = +)$$

  Use Bayes' Theorem.

- Q3: You are given some likelihoods and want to find the value of the ratio of posteriors (the odds):

$$O = \frac{(\text{Actual} = {\color{blue}\bullet}|\text{Observed} = {\color{blue}\bullet})}{(\text{Actual} = {\color{green}\bullet}|\text{Observed} = {\color{blue}\bullet})}$$

  Use Bayes' Theorem for numerator and denominator.

Probability Notation

Consider possible states of the world.

- Ω: Sample space - all possible world states.

Probability Notation

Consider possible states of the world.

- Ω: Sample space - all possible world states.
- $\omega \in \Omega$ - event; one possible world state. ωs are disjoint.

Probability Notation

Consider possible states of the world.

- Ω: Sample space - all possible world states.
- $\omega \in \Omega$ - event; one possible world state. ωs are disjoint.
- Associate numerical probability $P(\omega)$ to each ω.

Probability Notation

Consider possible states of the world.

- Ω: Sample space - all possible world states.
- $\omega \in \Omega$ - event; one possible world state. ωs are disjoint.
- Associate numerical probability $P(\omega)$ to each ω.
- $\phi \subseteq \Omega$ - proposition/assertion. $\omega \in \phi$ if ω satisfies the conditions set by ϕ.

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Prob. of ϕ is sum of probabilities of world states where ϕ is true.

Probability Notation

Consider possible states of the world.

- Ω: Sample space - all possible world states.
- $\omega \in \Omega$ - event; one possible world state. ωs are disjoint.
- Associate numerical probability $P(\omega)$ to each ω.
- $\phi \subseteq \Omega$ - proposition/assertion. $\omega \in \phi$ if ω satisfies the conditions set by ϕ.

  $$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

  Prob. of ϕ is sum of probabilities of world states where ϕ is true.
- Assignment of random variables defines a world state:

$$\omega = \{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$

Probability Notation

Consider possible states of the world.

- Ω: Sample space - all possible world states.
- $\omega \in \Omega$ - event; one possible world state. ωs are disjoint.
- Associate numerical probability $P(\omega)$ to each ω.
- $\phi \subseteq \Omega$ - proposition/assertion. $\omega \in \phi$ if ω satisfies the conditions set by ϕ.

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

  Prob. of ϕ is sum of probabilities of world states where ϕ is true.
- Assignment of random variables defines a world state:

$$\omega = \{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$

  - Usually interested in relationships between random variables.

Marginalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ:

$$P(\phi) = \sum_{\{\omega:\phi(\omega)=\text{True}\}} P(\omega)$$

- More generally, find the distribution of ϕ by averaging all possible values of $P(\phi|x)$:

$$P(\phi) = \sum_{x} P(\phi|x)P(x)$$

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.
- Probability that the test is positive, given patient has Leckieitis: $p(\text{Test} = +|L) = 0.99$

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.
- Probability that the test is positive, given patient has Leckieitis: $p(\text{Test} = +|L) = 0.99$
- Probability that the test is positive:

$$\begin{aligned} p(\text{Test} = +) &= p(\text{Test} = +|L)p(L) + p(\text{Test} = +|\neg L)p(\neg L) \\ &= 0.99 \times 10^{-4} + 0.01 \times (1 - 10^{-4}) \\ &= 0.0098 \end{aligned}$$

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.
- Probability that the test is positive, given patient has Leckieitis: $p(\text{Test} = +|L) = 0.99$
- Probability that the test is positive:

$$\begin{aligned} p(\text{Test} = +) &= p(\text{Test} = +|L)p(L) + p(\text{Test} = +|\neg L)p(\neg L) \\ &= 0.99 \times 10^{-4} + 0.01 \times (1 - 10^{-4}) \\ &= 0.0098 \end{aligned}$$

- Probability of having Leckieitis given the test is positive:

$$\begin{aligned} p(L|\text{Test} = +) &= \frac{p(\text{Test} = +|L)p(L)}{p(\text{Test} = +)} \\ &= 9.8 \times 10^{-3} \end{aligned}$$

Bayes' Rule

- Discrimination between blue/green taxis is 75% reliable, and you observed a blue taxi.
- Probability that the actual color is blue, given you observed blue:

$$p(\text{Actual} = \textcolor{blue}{\bullet} | \text{Observed} = \textcolor{blue}{\bullet}) = \frac{p(\text{Observed} = \textcolor{blue}{\bullet} | \text{Actual} = \textcolor{blue}{\bullet}) p(\text{Actual} = \textcolor{blue}{\bullet})}{p(\text{Observed} = \textcolor{blue}{\bullet})}$$

- Probability that the actual color is green, given you observed blue:

$$p(\text{Actual} = \textcolor{green}{\bullet} | \text{Observed} = \textcolor{blue}{\bullet}) = \frac{p(\text{Observed} = \textcolor{blue}{\bullet} | \text{Actual} = \textcolor{green}{\bullet}) p(\text{Actual} = \textcolor{green}{\bullet})}{p(\text{Observed} = \textcolor{blue}{\bullet})}$$

Bayes' Rule

- Want to find the ratio betwween posteriors:

$$O = \frac{p(\text{Observed} = \textcolor{blue}{\bullet} | \text{Actual} = \textcolor{blue}{\bullet}) p(\text{Actual} = \textcolor{blue}{\bullet})}{p(\text{Observed} = \textcolor{blue}{\bullet} | \text{Actual} = \textcolor{green}{\bullet}) p(\text{Actual} = \textcolor{green}{\bullet})}$$

- As $p(\text{Actual} = \textcolor{blue}{\bullet} | \text{Observed} = \textcolor{blue}{\bullet}) = 0.75$:

$$O = \frac{3p(\text{Actual} = \textcolor{blue}{\bullet})}{p(\text{Actual} = \textcolor{green}{\bullet})}$$

- If we know the *prior probabilities*: $p(\text{Actual} = \textcolor{green}{\bullet}) = 9p(\text{Actual} = \textcolor{blue}{\bullet})$, then incorporate this into the posterior ratio to find that, while you swear that the taxi is blue, being struck by a green taxi is still 3 times more likely.
  - The prior heavily influences your final uncertainty.

Naive Bayes

- For n possible boolean evidence variables there are 2^n possible combinations of conditional probabilities we need to know.
- Conditional independence of two variables X, Y given a third Z allows us to use only a reasonable number of combinations.

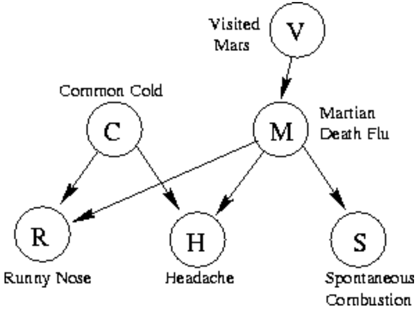
$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- For n effects that are all conditionally independent given the cause, the representation is $\mathcal{O}(n)$ instead of $\mathcal{O}(2^n)$.
- If a single cause is the direct cause of a number of effects, all of which are conditionally independent, then the full joint distribution is:

$$P(\text{Cause}, \text{Effect}_1, \ldots \text{Effect}_n) = P(\text{Cause}) \prod_i^n P(\text{Effect}_i|\text{Cause})$$

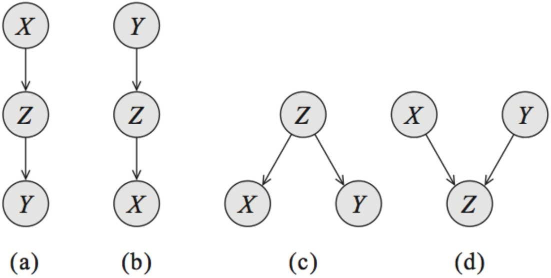
Bayesian Networks

- Full joint probability distribution specifies probability of each assignment of values to random variables. For n variables there are 2^n entries.
- Conditional independence between effect variables, given a cause variable, allows factorization of the full joint distribution into smaller conditional distributions.
- Bayesian Networks are a compact representation of the full joint distribution that shows dependencies between variables graphically.



Bayesian Networks

- Vertices correspond to random variables.
- Edges between vertices, e.g. $X \rightarrow Y$ indicates X has a direct influence on Y. Causes should be parents of effects.
- Each vertex has a conditional probability distribution summarizing effects of parents on the random variable $P(X_i | \text{Parents}(X_i))$.



Bayesian Networks

- Chain rule allows decomposition of joint into conditionals:

$$P(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)\ldots p(x_n|x_{n-1}, \ldots, x_2, x_1)$$

- Via conditional independence, each random variable x_i only directly depends on a small number of variables: Parents(x_i).

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(x_i))$$

- If each variable has d possible values and at most k parents, then the joint distribution has $\mathcal{O}(nd^k)$ entries (versus $\mathcal{O}(d^n)$).
- e.g. 20 random variables, each with 5 parents, then the Bayesian network approach uses 640 random variables versus over 10^6 for the full joint.

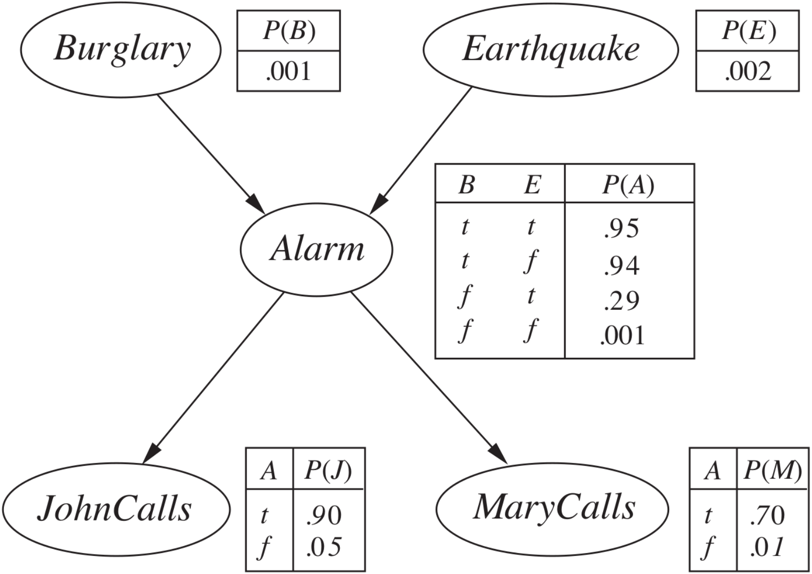
Bayesian Inference

- In the context of Bayesian networks, compute posterior $P(X|\mathrm{e})$ for a query X (some assignment of random variables) given observed event e (assignment to a set of evidence variables).
- 'Find probability of X, given we know e has occurred.'
  - Let H denote all variables outside X, e (call H hidden variables), let Z be the evidence (i.e. normalizing constant). From Bayes' Theorem:

$$\begin{aligned} P(X|\mathrm{e}) &= \frac{1}{Z} P(X, \mathrm{e}) \\ &= \frac{1}{Z} \sum_{\mathrm{H}} P(X, \mathrm{e}|\mathrm{H}) p(\mathrm{H}) \\ &= \frac{1}{Z} \sum_{\mathrm{H}} P(X, \mathrm{e}, \mathrm{H}) \end{aligned}$$

- To solve a Bayesian inference problem:
  - Identify query, evidence and hidden variables.
  - Decompose joint distribution using Bayesian network structure into product of simpler conditional distributions.
  - Fix evidence variables to observed values.
  - Sum (marginalize) over remaining hidden variables H.

Bayesian Inference



$P(B)$
.001

$P(E)$
.002

B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Bayesian Inference

- If A observed, B and E are no longer independent! Knowledge of A couples the parent variables. This is an example of a V-structure.
- Parents are independent if child is unobserved, but coupled when child is observed.
- Simpler example: suppose your lawn is wet in the morning (C). A (rain) and B (sprinkler) are two possible causes for it being wet. If we know C is true and A is false, then B must be true. i.e. A and B are not conditionally independent given C.

Sequential Bayesian Updates

- Bayes' Theorem allows us to update our uncertainty as new information is acquired.
- Hypothesis H; observe a series of independent measurements $\{x_1, x_2, \ldots x_T\}$.
- How does our uncertainty about H evolve given these observations?

Sequential Bayesian Updates

- Given sequential measurements $\{x_1, x_2, \ldots x_T\}$, our likelihood at time t summarizes the probability of the data given the hypothesis H:

$$p(x_1, \ldots x_t | H) = p(x_1|H)p(x_2|x_1, H) \ldots p(x_t|\mathrm{x}_{t-1}, H)$$

  Where we let $\mathrm{x}_n = (x_1, x_2, \ldots, x_n)$.

- Bayes' Theorem:

$$p(H|\mathrm{x}_t) \propto p(\mathrm{x}_t|H)p(H)$$

- At time $t+1$, the posterior is:

$$p(H|\mathrm{x}_{t+1}) \propto p(\mathrm{x}_{t+1}|H)p(H)$$

- How to get from $P(H|\mathrm{x}_t)$ to $P(H|\mathrm{x}_{t+1})$?

Sequential Bayesian Updates

- Use the chain rule:

$$
\begin{aligned}
p(H|\mathrm{x}_{t+1}) &\propto p(\mathrm{x}_{t+1}|H)p(H) \\
&= p(\mathrm{x}_{t+1}, \mathrm{x}_t|H)p(H) \\
&= p(\mathrm{x}_{t+1}|\mathrm{x}_t, H)p(\mathrm{x}_t|H)p(H) \\
&\propto p(\mathrm{x}_{t+1}|H)p(H|\mathrm{x}_t)
\end{aligned}
$$

$$
\text{New posterior} = \text{Likelihood of new measurement} \times \text{Current posterior} \tag{1}
$$

- How does our uncertainty about H evolve given these observations?
  - Answer: Reuse the current posterior distribution as the prior distribution in the next time step, and normalize appropriately.

Sequential Bayesian Updates

- Let $\pi_t(H)$ be the posterior at time t, then the recursive update reads:

$$\pi_{t+1}(H) \propto p(x_{t+1}|H)\pi_t(H)$$

- $\pi_t(H)$ summarizes entire history of the sequence.
    - Normalization factor Z is average over all possible values of H: $Z = \sum_{h'} p(x_{t+1}|H = h')\pi_t(H = h')$
- In summary, Bayesian inference provides an efficient way of sequentially updating our belief about a state that only depends on the current measurement and posterior.

Sequential Bayesian Updates in Robotics

- In robotics, your hypothesis can be e.g., your position or state θ, which evolves in time.
- Assume your dynamics are Markov. i.e. the state θ_{t+1} only depends on the current state θ_t:

$$p(\theta_0) = \pi(\theta_0), \quad p(\theta_{t+1}|\theta_0, \theta_1, \ldots, \theta_t) = p(\theta_{t+1}|\theta_t)$$

- State θ is hidden - only measurements x_t observed. To understand θ, look at joint density of all states $\boldsymbol{\theta}_t$ and measurements x_t:

$$p(\boldsymbol{\theta}_t, \mathrm{x}_t) = \pi(\theta_0) \prod_{i=0}^{t} p(\theta_i|\theta_{i-1}) p(x_i|\theta_i)$$

The Bayes Filter

- This Markovian + Bayesian model (HMM) is widely used in:
  - Speech recognition.
  - Robotics.
  - Particle physics.
  - GPS / target tracking.
  - Brain imaging.
- In robotics, use sensor data gathered to recursively update 'belief' of position/velocity estimate.
- Remember that the evolving state θ_t is unknown, typically we want to:
  - Filter: Compute $p(\theta_t|\mathrm{x}_t)$ to estimate current state.
  - Predict: Compute $p(\theta_{t+k}|\mathrm{x}_t)$ to predict future states.
  - Reconstruct: Compute $p(\theta_{t-k}|\mathrm{x}_t)$ to identify pre vious states.

The Bayes Filter

- Want posterior at $t+1$ given observations x_{t+1}:

$$p(\theta_{t+1}|\mathrm{x}_{t+1}) = p(\theta_{t+1}|\mathrm{x}_t, x_{t+1}) \tag{2}$$
$$\propto p(x_{t+1}|\theta_{t+1}, \mathrm{x}_t)p(\theta_{t+1}|\mathrm{x}_t) \tag{3}$$
$$= p(x_{t+1}|\theta_{t+1})p(\theta_{t+1}|\mathrm{x}_t) \tag{4}$$

- Compute $p(\theta_{t+1}|\mathrm{x}_t)$ by averaging over current state θ_t:

$$p(\theta_{t+1}|\mathrm{x}_t) = \sum_{\theta_t} p(\theta_{t+1}|\theta_t)p(\theta_t|\mathrm{x}_t)$$

The Bayes Filter

- We perform the prediction step by averaging over all possible values of the current state θ_t:

$$p(\theta_{t+1}|\mathsf{x}_t) = \int_{\theta_t} d\theta_t \; p(\theta_{t+1}|\theta_t)p(\theta_t|\mathsf{x}_t)$$

- Then perform the filter step by the New $\propto$ Current $\times$ Likelihood rule, combining the predictive distribution with the likelihood of the next measurement.

$$p(\theta_{t+1}|\mathsf{x}_{t+1}) \propto p(\theta_{t+1}|\mathsf{x}_t)p(\mathsf{x}_{t+1}|\theta_{t+1})$$

- So the overall process is:

Predict-Observe-Filter-Predict-Observe-Filter-. . .