Discussion

CS 5/7320

Artificial Intelligence

Quantifying Uncertainty:

Probabilities & Bayesian Decision Making

AIMA Chapter 12

Slides by Michael Hahsler

based on slides by Svetlana Lazepnik

with figures from the AIMA textbook

Cover image: "Dice" by Steve A Johnson

Online Material

Making a Decision Under Uncertainty

Given outcome probabilities, which action should the agent choose?

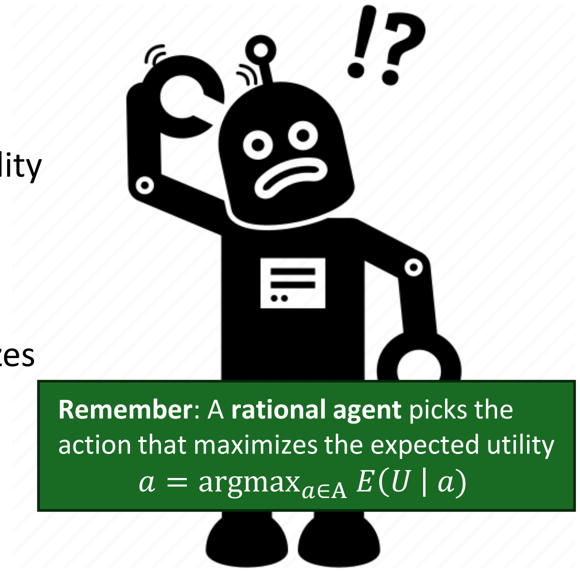
- Depends on **preferences** for missing a flight vs. time spent waiting.
- **Utility theory** represents preferences for different outcome using a utility function $U(outcome)$.

- **Decision Theory = Probability Theory + Utility Theory**
- The agent should choose actions that lead to an outcome that maximizes the **expected utility**.

$$a = \text{argmax}_{A_t}\, E[U(A_t)]$$

- The outcome depends on the action taken:

$$U(A_t) = U(reached\ state) - Cost(action)$$



Remember: A **rational agent** picks the action that maximizes the expected utility $a = \text{argmax}_{a\in \text{A}}\, E(U \mid a)$

Example:

Belief states with probabilities

	A_{20}	A_{90}	A_{120}	A_{1444}
$P(on\ time)$	0.04	0.8	0.99	0.9999
$P(missed\ flight)$	0.96	0.2	0.01	1E-04
$P(at\ home)$	0	0	0	0

+

Utility structure

	Value
U$(on\ time)$	1000
U$(missed\ flight)$	-1000
Cost$(per\ minute)$	1

=

	Exp. Utility
A_{20}	-901.0
A_{90}	546.0
A_{120}	**862.2**
A_{1444}	-443.9

$$E[U(A_t)] = P(on\ time|A_t)\, U(on\ time \mid A_t) + P(missed\ flight|A_t) U(missed\ flight\ |A_t)$$

$$E[U(A_{120})] = 0.99 \times (1000 - 120 \times 1) + 0.01 \times (-1000 - 120 \times 1) = 862.4$$

Bayes' Theorem: The Bayesian Update Rule

The product rule gives us two ways to factor a joint distribution for events $X = x$ and $E = e$:

$$P(x, e) = P(x \mid e)P(e) = P(e \mid x)P(x)$$

Posterior Prob.

Prior Prob.

Therefore, $P(x \mid e) = \frac{P(e|x)\ P(x)}{P(e)}$

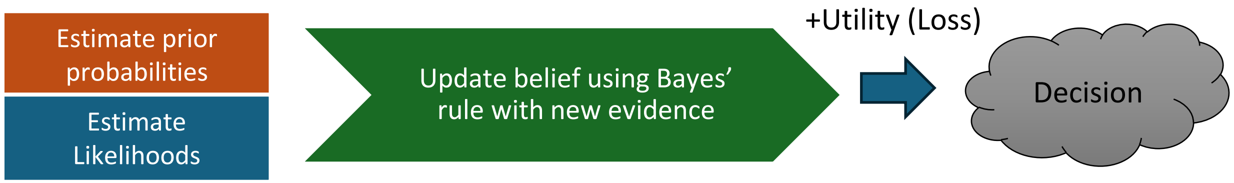
Add evidence

Why is this useful?

- We can update our beliefs about an event x based on new evidence e.
- Update rule $P(x) \leftarrow \frac{P(e|x)\ P(x)}{P(e)}$

Written a distribution over all values for X: $\boldsymbol{P}(X \mid e) = \frac{\boldsymbol{P}(e|X)\boldsymbol{P}(X)}{\boldsymbol{P}(e)}$

MAP: Maximum A Posteriori Decision



0-1 loss means we should use the value x that has the highest (maximum) posterior probability given the evidence e, i.e., the prediction that most likely leads to a loss of 0.

$$x^* = \operatorname{argmax}_x \overbrace{P(x|e)}^{\text{Posterior Prob.}} = \operatorname{argmax}_x \frac{P(e|x)\overbrace{P(x)}^{\text{Prior Prob.}}}{P(e)}$$

$$= \operatorname{argmax}_x P(e|x)P(x)$$

$P(e)$ is fixed for a given evidence.

This is the optimal decision for 0-1 loss.

MAP: Example

We observe: e = stripes

What is the animal? $x \in \{\text{zebra, dog, cat}\}$

$$x^* = \text{argmax}_x \overbrace{P(x|e)}^{\text{Posterior Prob.}} = \text{argmax}_x \frac{P(\text{stripes}|x)P(x)}{P(stripes)}$$

$$= \text{argmax}_x \underbrace{P(\text{stripes}|x)}_{\text{likelihood}} \underbrace{P(x)}_{\text{Prior Prob.}}$$

Zebra: The likelihood $P(\text{stripes} \mid \text{zebra})$ is the highest. But the decision also depends on the prior $P(\text{zebra})$, the chance that we see a zebra.

Cat: The likelihood for cats having stripes may be smaller, but the prior probability of seeing a cat is much higher. Cat may have a larger posterior probability!

Bayes Classifier

$F_1, F_2, \dots, F_n, H$

- Suppose we have many different types of observations (evidence, symptoms, features) $F_1, \dots, F_n$ that we want to use to decide on an underlying hypothesis H.

- The MAP decision involves estimating

$$h^* = \text{argmax}_{h \in H} \; P(f_1, \dots, \; f_n | h) P(h)$$

- How many entries does the tables $P(f_1, \dots, \; f_n | h)$ have?

Answer: If we assume that each feature can take on k values then the table has $\mathbf{O}(\boldsymbol{k^n})$ entries! What if we have 1000s of features?

Naïve Bayes Model

- We want to use the MAP decision which involves estimating

$$h^* = \text{argmax}_{h \in H} \; P(f_1, \dots, \; f_n | h) P(h)$$

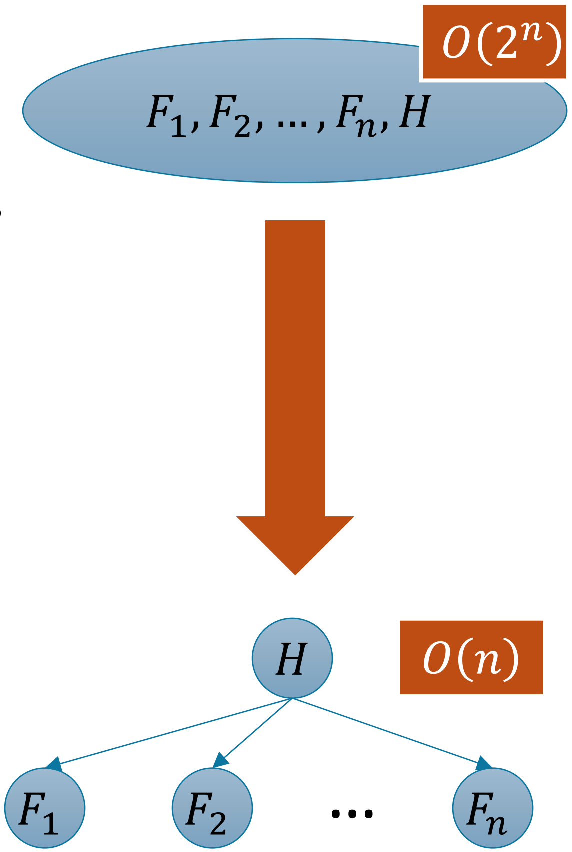
- **Issue**: The likelihood table size grows for n variables with k different values exponentially with $O(k^n)$

- The naïve Bayes model makes the **simplifying assumption** that the different **features are conditionally independent *given the hypothesis.***

  This reduces the needed number of probabilities to $O(k \times n)$:

$$\hat{h} = \text{argmax}_{h \in H} \; P(h) \prod_{i=1}^{n} P(f_i | h)$$

- The naïve Bayes decision is not optimal.



Conclusion

Bayesian Intelligent Agents: This is a type of utility-based agent is also called a decision-theoretic agent.

Approach