Decision Theory Study Guide

Artificial Intelligence

1 Preferences

1. Consider the following preferences: an agent prefer puppies to kittens, kittens to babies, and babies to puppies. Are these preferences rational? Why or why not?

> **Solution:** These preferences are not rational because they violate the transitivity axiom of preference theory, namely, $(A \succ B) \wedge (B \succ C) \implies (A \succ C)$. Given the agent's first two preferences, it is irrational to prefer babies to puppies.

2 Utilities

1. I prefer iced sugar cookies to chocolate chip cookes. If I were to assign numerical utility values to each of these cookies I would assign a utility of 12 to chocolate chip cookies and 10 to iced sugar cookies. Are my utility values valid? Why or why not?

> **Solution:** Given my preferences, these utility values are not valid values of a utility function for a rational agent. If an agent's preferences obey the axioms of utility, then there exists a utility function U such that $U(A) > U(B)$ if and only if A is preferred to B, and $U(A) = U(B)$ if and only if the agent is indifferent between A and B. That is,
>
> $$U(A) > U(B) \iff A \succ B$$
> $$U(A) = U(B) \iff A \sim B$$

2. Given two attributes of cookies, *icing* and *nuts*, an agent has the following preferences:

$$icing \succ \neg icing$$
$$\neg nuts \succ nuts$$

What can you say about the agent's choice between cookie A with icing and no nuts, and cookie B with nuts and no icing?

> **Solution:** Cookie A *strictly dominates* Cookie B.