# Machine Learning Homework # 2

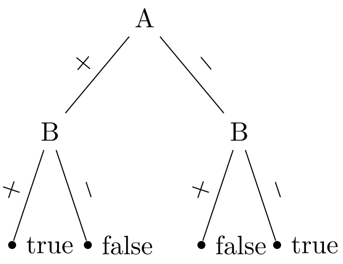
sakohl, milsen, jkirchner

May 2, 2015

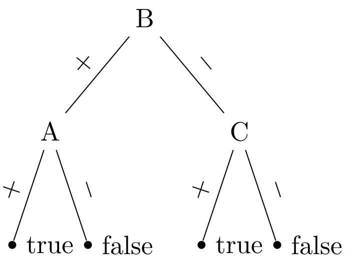
## 1 Exercise

Build/draw the decision trees for the following boolean functions:

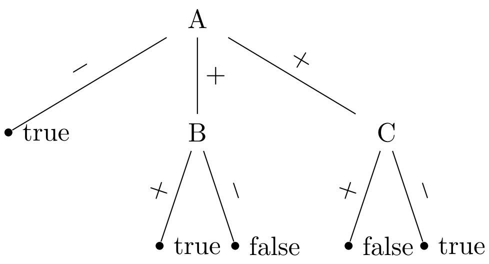
1. A⊕¬B (⊕ = xor)



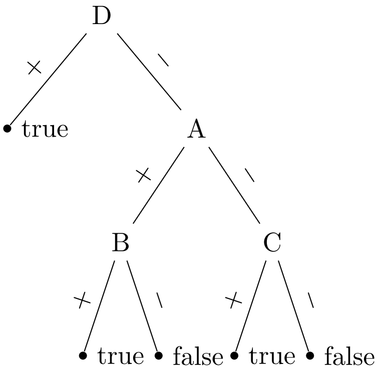
2. $(A \land B) \lor (\neg B \land C)$



3. $(A \to B) \lor (A \to \neg C)$



4. $(A \wedge B) \vee (\neg A \wedge C) \vee D$



## 2 Exercise

1. (4 p) Consider the five training examples from Table 2. Build the root node of a decision tree from these training examples. To do this, you calculate the information gain on all three distinct attributes (genre, main-character, has ninjas) to decide which one would be the best choice for the root node (the one with the largest gain).

   - $Entropy(S) = -\frac{3}{5}log_2(\frac{3}{5}) - \frac{2}{5}log_2(\frac{2}{5}) \approx 0.97095$
   - $Gain(S, Genre) = Entropy(S) - \sum_{v \in \{Action, Comedy, Romance\}} \frac{|S_v|}{|S|} Entropy(S_v)$
     $= Entropy(S) - (\frac{2}{5}Entropy(S_{action})) - (\frac{2}{5}Entropy(S_{romance})) - (\frac{1}{5}Entropy(S_{comedy}))$
     $= Entropy(S) - (\frac{2}{5} * 1) - (\frac{2}{5} * 1) - (\frac{1}{5} * 0)$
     $\approx 0.97095 - \frac{4}{5} = 0.17095$
   - $Gain(S, MainCharacter) = Entropy(S) - \sum_{v \in \{male, female\}} \frac{|S_v|}{|S|} Entropy(S_v)$
     $= Entropy(S) - (\frac{2}{5}Entropy(S_{male})) - (\frac{3}{5}Entropy(S_{female}))$
     $\approx Entropy(S) - (\frac{2}{5} * 1) - (\frac{3}{5} * 0.9183))$
     $\approx 0.97095 - 0.95098 = 0.01997$
   - $Gain(S, has\_ninjas) = Entropy(S) - \sum_{v \in \{true, false\}} \frac{|S_v|}{|S|} Entropy(S_v)$
     $= Entropy(S) - (\frac{4}{5}Entropy(S_{true})) - (\frac{1}{5}Entropy(S_{false}))$
     $\approx Entropy(S) - (\frac{4}{5} * 0.811) - (\frac{3}{5} * 0))$
     $\approx 0.97095 - 0.649 = 0.32193$

   → Therefore ninjas are the best choice for the root note

2. (2 p) Perform the same calculation as in a) but use the gain ratio instead of the information gain. Does the result for the root node change?

   - $SplitInformation(S, Genre) = -\frac{2}{5}log_2(\frac{2}{5}) - \frac{2}{5}log_2(\frac{2}{5}) - \frac{1}{5}log_2(\frac{1}{5}) = 1.52193$
   - $SplitInformation(S, main\_character) = -\frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5}) = 0.97095$

- $SplitInformation(S, has\_ninjas) = -\frac{4}{5}log_2(\frac{4}{5}) - \frac{1}{5}log_2(\frac{1}{5}) = 0.7219$
- $GainRatio(S, Genre) = \frac{Gain(S,Genre)}{SplitInformation(S,Genre)} \approx 0.11232$
- $GainRatio(S, main\_character) = \frac{Gain(S,main\_character)}{SplitInformation(S,main\_character)} \approx 0.0205$
- $GainRatio(S, has\_ninjas) = \frac{Gain(S,has\_ninjas)}{SplitInformation(S,has\_ninjas)} \approx 0.44595$

→ Therefore ninjas are still the best choice for the root note

3. (2 p) Let's assume the root node is a node which checks the value of the attribute has ninjas. Calculate the next level of the decision tree using the information gain.

   Decision for: $has\_ninjas = true$

   - $Entropy(has\_ninja) = -\frac{3}{4}log_2(\frac{3}{4}) - \frac{1}{4}log_2(\frac{1}{4}) \approx 0.811278124$
   - $Gain(has\_ninja, Genre) = Entropy(has\_ninja) - \sum_{v \in \{Action, Comedy, Romance\}} \frac{|S_v|}{|S|} Entropy(S_v)$
     $= Entropy(has\_ninja) - (\frac{2}{4}Entropy(S_{action})) - (\frac{1}{4}Entropy(S_{romance})) - (\frac{1}{4}Entropy(S_{comedy}))$
     $= Entropy(has\_ninja) - (\frac{2}{4} * 1) - (\frac{1}{4} * 0) - (\frac{1}{4} * 0)$
     $\approx 0.811278124 - 0.5 = 0.311278124$
   - $Gain(has\_ninja, MainCharacter) = Entropy(has\_ninja) - \sum_{v \in \{male, female\}} \frac{|S_v|}{|S|} Entropy(S_v)$
     $= Entropy(has\_ninja) - (\frac{2}{4}Entropy(S_{male})) - (\frac{2}{4}Entropy(S_{female}))$
     $= Entropy(has\_ninja) - (\frac{2}{4} * 1) - (\frac{2}{4} * 0)$
     $\approx 0.811278124 - 0.5 = 0.311278124$

   - → Therefore the values are not conclusive. The algorithm would probably choose a random attribute to split.

   Decision for: $has\_ninjas = false$

   - The examples are already classified perfectly, no further split.