MTAT.03.227 Machine Learning
Practice session 6
Decision Trees - ID3 Algorithm

October 21-25, 2019

Exercise 1.

To make further computations easier, implement a function that calculates the value of entropy E for one attribute T that has 2 possible values:
$E(T) = E(N_0, N_1) = -plog_2(p) - (1-p)log_2(1-p)$, where $p = \frac{N_1}{N_0+N_1}$
N_0 is the count of instances with negative class and N_1 is the count of instances with positive class.

(a) In what situation is entropy zero?

(b) What is the value of entropy if classes are equally divided?

Exercise 2.

Given the Tennis dataset in the figure below, simulate the ID3 algorithm:

Outlook	Temp	Humidity	Windy	Play
Overcast	Hot	High	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

(a) Calculate the entropy of the target "Play".

(b) Information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

When the dataset is split on different attributes, entropy for each branch is calculated. Then they are added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information gain (IG).

Calculate information gains for all possible splits in the first level. Choose the best split based on these values. Information gain after split:

$$IG = E(T) - \sum_v P(v)E(N_0^v, N_1^v)$$

(c) Divide the dataset by the branches of chosen attribute and repeat the same process on every branch. Note that a branch with entropy of 0 is a leaf node. Branches with entropy > 0 need further splitting.

Exercise 3.

Draw the final Decision Tree.

Exercise 4.

Based on this decision tree, predict the decision on a new day with features:
{Outlook=RAINY, temp=MILD, humidity=HIGH, windy=Weak}.
DNF(Disjunctive Normal Form)

(a) Write out the decision tree as DNF.

(b) Can you guess why decision trees are called logical models?

(c) Can any logical expression be written out as a decision tree?

Entropy for two attributes is calculated as the following proportional sum of entropies:

$$E(T,X) = \sum_{v \in X} P(v)E(v),$$

where v = possible values for the attribute X