Decision Trees Example Problem

Sakib Ashraf Zargar

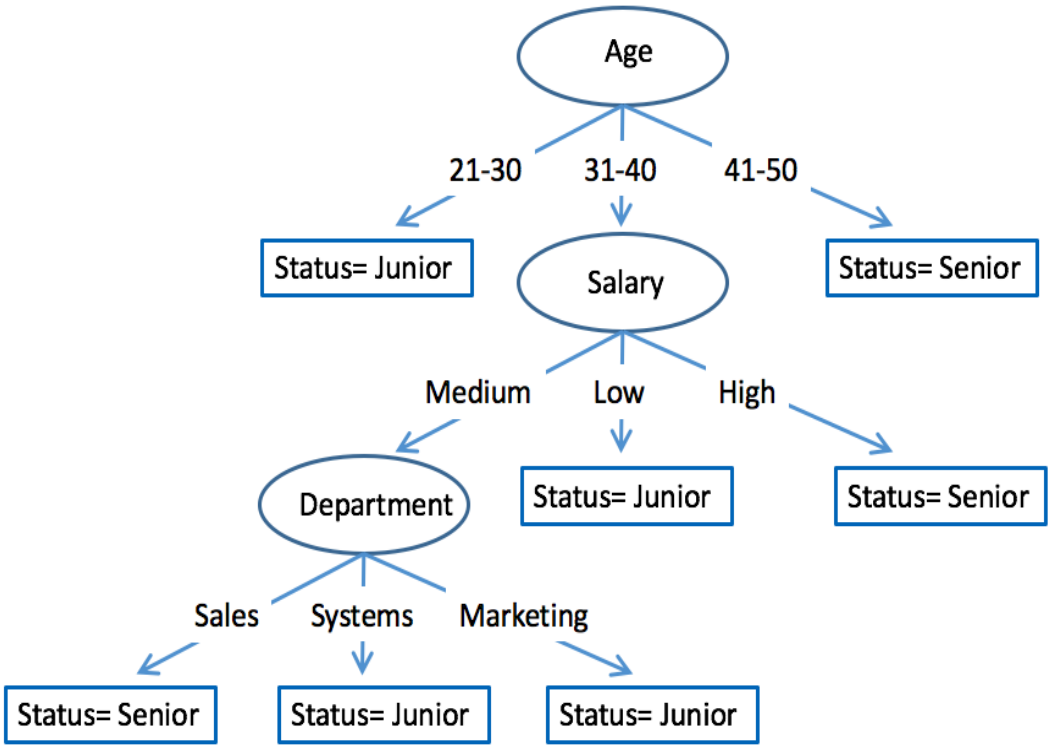
Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, North Carolina 27606, USA

Using data given in the table below as training data, answer the following questions:

1. Construct decision tree (no pruning) using Entropy.
2. Compute the following on training data:(i) individual class accuracy (ii) overall class accuracy
3. For the following test data, predict the class label for each instance using the tree

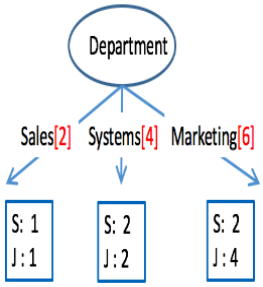
Department	Age	Salary	Status
Sales	31-40	Medium	Senior
Sales	31-40	Low	Junior
Systems	21-30	Medium	Junior
Systems	31-40	High	Senior
Systems	21-30	Medium	Junior
Systems	41-50	High	Senior
Marketing	31-40	Medium	Senior
Marketing	31-40	Medium	Junior
Marketing	41-50	High	Senior
Marketing	21-30	High	Junior
Marketing	31-40	Medium	Junior
Marketing	31-40	Medium	Junior

(a) The decision tree and the step wise procedure is as follows:



Step 1: Choosing the root node

Before splitting: $\rightarrow E(p) = -[(5/12)\log_2(5/12) + (7/12)\log_2(7/12)] = 0.9799$



$$E(D = \text{Sales}) = -[(1/2)\log_2(1/2) + (1/2)\log_2(1/2)] = 1.0000$$

$$E(D = \text{Syst.}) = -[(2/4)\log_2(2/4) + (2/4)\log_2(2/4)] = 1.0000$$

$$E(D = \text{Mark}) = -[(2/6)\log_2(2/6) + (4/6)\log_2(4/6)] = 0.9183$$

$$\rightarrow \bar{E} = (2/12)1 + (4/12)1 + (6/12)0.9183 = 0.9591$$

$$\rightarrow G(p, \text{Department}) = 0.9799 - 0.9591 = 0.0208$$

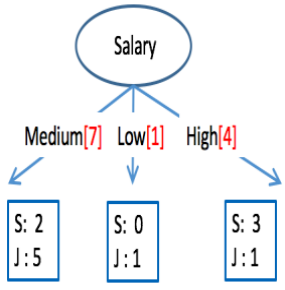
$$E(A = 31\_40) = -[(3/7)\log_2(3/7) + (4/7)\log_2(4/7)] = 0.9858$$

$$E(A = 21\_30) = -[(0/3)\log_2(0/3) + (3/3)\log_2(3/3)] = 0$$

$$E(A = 41\_50) = -[(2/2)\log_2(2/2) + (0/2)\log_2(0/2)] = 0$$

$$\rightarrow \bar{E} = (7/12)0.9858 + (3/12)0 + (2/12)0 = 0.5747$$

$$\rightarrow G(p, \text{Age}) = 0.9799 - 0.5747 = 0.4052$$



$$E(S = \text{Med}) = -[(2/7)\log_2(2/7) + (5/7)\log_2(5/7)] = 0.8631$$

$$E(S = \text{Low}) = -[(0/1)\log_2(0/1) + (1/1)\log_2(1/1)] = 0$$

$$E(S = \text{High}) = -[(3/4)\log_2(3/4) + (1/4)\log_2(1/4)] = 0.8113$$

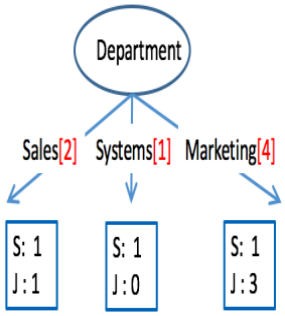
$$\rightarrow \bar{E} = (7/12)0.8631 + (1/12)0 + (4/12)0.8113 = 0.7739$$

$$\rightarrow G(p, \text{Salary}) = 0.9799 - 0.7739 = 0.2060$$

Winner → Age

Step 2: Choosing the next test attribute

$$\rightarrow E(A = 31-40) = 0.9858$$



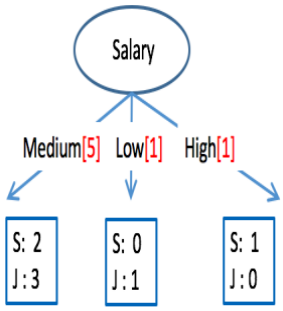
$$E(D = \text{Sales}) = -[(1/2)\log_2(1/2) + (1/2)\log_2(1/2)] = 1.0000$$

$$E(D = \text{Syst.}) = -[(1/1)\log_2(1/1) + (0/1)\log_2(0/1)] = 0$$

$$E(D = \text{Mark}) = -[(1/4)\log_2(1/4) + (3/4)\log_2(3/4)] = 0.8113$$

$$\rightarrow \bar{E} = (2/7)1 + (1/7)0 + (4/7)0.8113 = 0.7493$$

$$\rightarrow G(p, \text{Department}) = 0.9858 - 0.7493 = 0.2365$$



$$E(S = \text{Med}) = -[(2/5)\log_2(2/5) + (3/5)\log_2(3/5)] = 0.9710$$

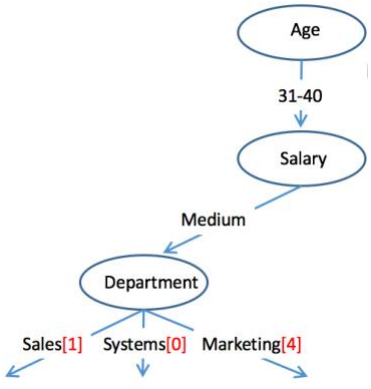
$$E(S = \text{Low}) = -[(0/1)\log_2(0/1) + (1/1)\log_2(1/1)] = 0$$

$$E(S = \text{High}) = -[(1/1)\log_2(1/1) + (0/1)\log_2(0/1)] = 0$$

$$\rightarrow \bar{E} = (5/7)0.9710 + (1/7)0 + (1/7)0 = 0.6935$$

$$\rightarrow G(p, \text{Salary}) = 0.9858 - 0.6935 = 0.2923$$

Winner → Salary



Now, for this branch, there appear to be 2 issues in the end which are handled as follows (as per the textbook):

1. There are no training examples associated with D=Systems, so it is declared as a leaf node with the same class as the majority class of the training examples associated with its Parent node. (S=2, J=3→ so we choose Junior).

2. Marketing node is still impure (S=1, J=3), so it is assigned the label of its majority class.

(b) Based on the decision tree constructed above, the following confusion matrix can be plotted:

	Predicted: Senior	**Predicted:** Junior
Actual: Senior	4 (TP)	1 (FN)
Actual: Junior	0 (FP)	7 (TN)

$\rightarrow$ Accuracy_Senior $= 4/5 = 0.8$ $\rightarrow$ Accuracy_Junior $= 7/7 = 1$ $\rightarrow$ Accuracy $= 11/12$

(c) Based on the decision tree constructed in (a):

Department	**Age**	**Salary**	**Status**
Sales	31-40	Low	Junior
Systems	31-40	Medium	Junior
Marketing	31-40	High	Senior
Marketing	21-30	Low	Junior