**Question 1 – Decision trees: Consider the training examples shown below for a binary classification task (i.e. two classes, C0 and C1).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer | Gender | Car Type | Shirt Size | Class |
| 1 | M | Family | Small | C0 |
| 2 | M | Sports | Medium | C0 |
| 3 | M | Sports | Medium | C0 |
| 4 | M | Sports | Large | C0 |
| 5 | M | Sports | Extra Large | C0 |
| 6 | M | Sports | Extra Large | C0 |
| 7 | F | Sports | Small | C0 |
| 8 | F | Sports | Small | C0 |
| 9 | F | Sports | Medium | C0 |
| 10 | F | Luxury | Large | C0 |
| 11 | M | Family | Large | C1 |
| 12 | M | Family | Extra Large | C1 |
| 13 | M | Family | Medium | C1 |
| 14 | M | Luxury | Extra Large | C1 |
| 15 | F | Luxury | Small | C1 |
| 16 | F | Luxury | Small | C1 |
| 17 | F | Luxury | Medium | C1 |
| 18 | F | Luxury | Medium | C1 |
| 19 | F | Luxury | Medium | C1 |
| 20 | F | Luxury | Large | C1 |

1. Looking at each attrribute in turn, draw the branches that would arise if it was chosen for the root node, showing the breakdown of classes in each leaf node.
2. Without doing calculations, which attribute looks the most promising?
3. Using the calculations given below, what is the entropy for each branch? Which seems to give the best information gain and does this concur with the attribute you selected in part b) above?

-6/10 Log (6/10) - 4/10 log (4/10) = 0.970950594

-6/10 Log (6/10) - 4/10 log (4/10) = 0.970950594

-7/8\* LOG(7/8) - 1/8\*LOG(1/8) = 0.543564443

-1/4\*LOG(1/4) - 3/4\*LOG(3/4)= 0.811278124

-8/8\*LOG(8/8) - 0/0 LOG(0/0)= 0

-3/5\*log(3/5) - 2/5\*LOG(2/5)= 0.970950594

-3/7\*LOG(3/7) - 4/7\*LOG(4/7)= 0.985228136

-2/4\*LOG(2/4))-2/4\*LOG(2/4)= 1

-2/4\*LOG(2/4) -2/4\*LOG(2/4)= 1

1. The branch for Car Type=Family matched four rows in the dataset, one row in C0, and three rows in C1. Grow the tree one more level from this branch, trying gender first, and then shirt size. Does gender or shirt size give the best information gain for these four rows?

**Question 2 – *k*-NN: Consider the training examples shown below for a binary classification task (i.e. two classes, C0 and C1).**

|  |  |  |  |
| --- | --- | --- | --- |
| **Gender** | **Car Type** | **Shirt Size** | **Class** |
| M | Family | Small | C0 |
| M | Sports | Medium | C0 |
| M | Sports | Medium | C0 |
| F | Luxury | Large | C0 |
| M | Family | Large | C1 |
| F | Luxury | Medium | C1 |
| F | Luxury | Medium | C1 |
| F | Luxury | Large | C1 |

The following row of data needs to be classified:

|  |  |  |  |
| --- | --- | --- | --- |
| M | Luxury | Medium | ?? |

1. Calculate the distance between the row above, and each row in the dataset.
2. Rank the rows in order of how similar they are to the row you are classifying
3. If k=3, what is the class label?