# Auto Learn Data Analysis
# HW3
**Rishabh Sinha (rsinha2)**
**CSC 522**

**Q1**

| Department | Age | Salary | Status |
|---|---|---|---|
| Sales | 31-40 | Medium | Senior |
| Sales | 21-30 | Low | Junior |
| 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 | Low | Junior |
| Marketing | 21-30 | Medium | Junior |

**Using data given in Table 1 as training data, answer the following question:**
**a. (6 points) Construct decision tree using GINI index. Show all work and draw the resulting tree (no pruning) .**

GINI = $1 - \Sigma\ p(i/t)^2$
AGE:
The GINI index value for the age attribute is :-
GINI(21-30) = $1 - (5/5)^2 - (0/5)^2 = 0$

GINI(31-40) = $1 - (3/5)^2 - (2/5)^2 = 0.48$

GINI(41-50) = $1 - (2/2)^2 - (0/2)^2 = 0$

GINI(Age) = $(5/12)*0.48 = 0.2$

Salary:
The GINI index value for the Salary attribute is :-
GINI(Medium) = 1 – $(4/6)^2$ – $(2/6)^2$ = 0.444

GINI(Low) = 1 – $(3/3)^2$ – $(0/3)^2$ = 0

GINI(High) = 1 – $(3/3)^2$ – $(0/3)^2$ = 0

GINI(Salary) = (6/12)*0.4444 = 0.2222

Department:
The GINI index value for the Department attribute is :-
GINI(Sales) = 1 – $(1/3)^2$ – $(2/3)^2$ = 0.444

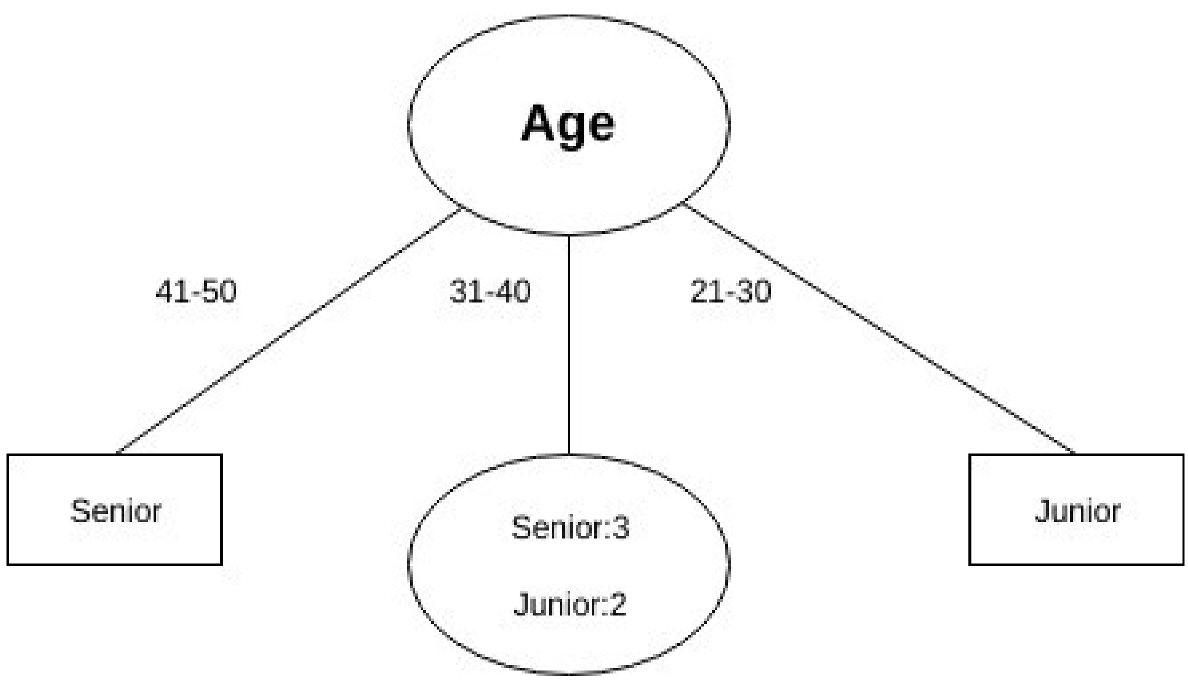
GINI(Systems) = 1 – $(2/4)^2$ – $(2/4)^2$ = 0.5

GINI(Marketing) = 1 – $(2/5)^2$ – $(3/5)^2$ = 0.48

GINI(Department) = (5/12)*0.48 + (3/12)*0.4444 + (4/12)*0.5 = 0.477

The minimum value for the GINI index is for the attribute Age.
Thus we use age as the root node.

Now the Decision tree looks like:




Now When the value of Age is 31-40 we get don't get 100% accuracy, thus we can split this node.

Thus the value for GINI coefficient for Salary, when the Age is 31-40 is:
GINI(Salary|Age = 31-40) =

GINI(Medium) = 1 – $(2/3)^2$ – $(1/3)^2$ = 0.444

GINI(Low) = 1 – $(1/1)^2$ – $(0/1)^2$ = 0

GINI(High) = 1 – $(1/1)^2$ – $(0/1)^2$ = 0

GINI(Salary|Age = 31-40) = (3/5)*0.4444 = 0.2664

GINI coefficient for Department, when the Age is 31-40 is:
GINI(Department|Age = 31-40) =

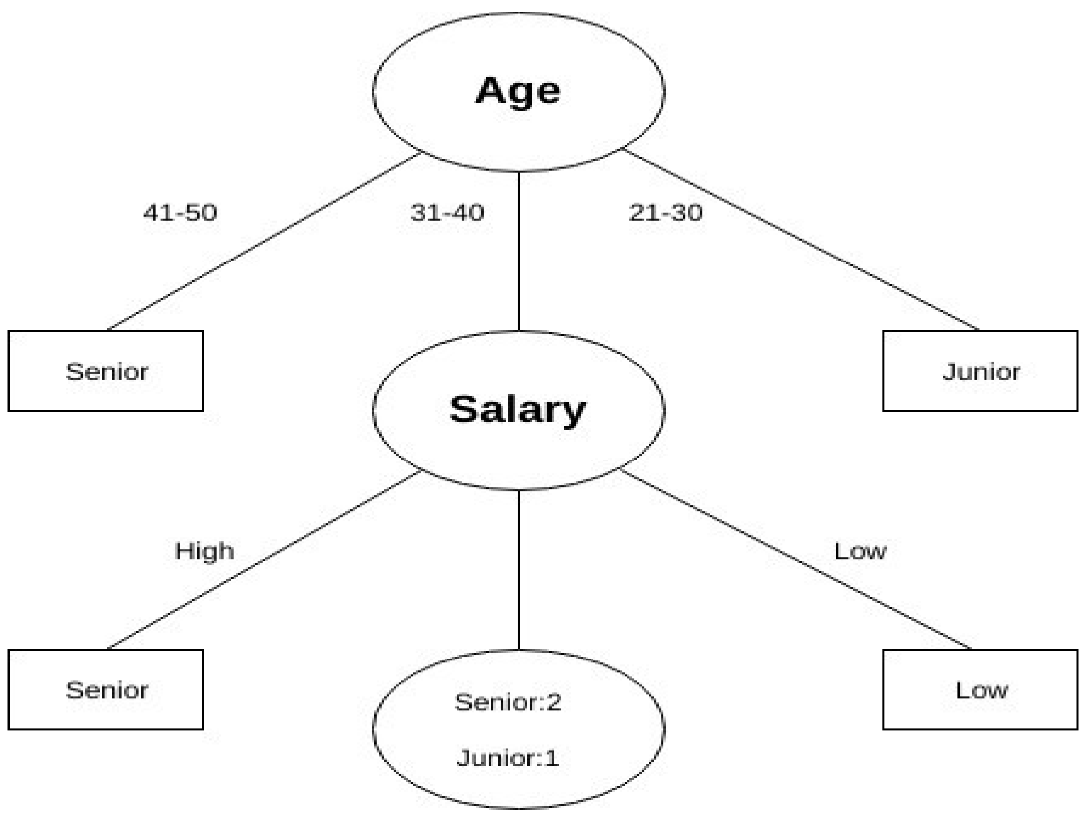
GINI(Sales) = 1 – $(1/2)^2$ – $(1/2)^2$ = 0.5

GINI(Systems) = 1 – $(1/1)^2$ – $(0/1)^2$ = 0

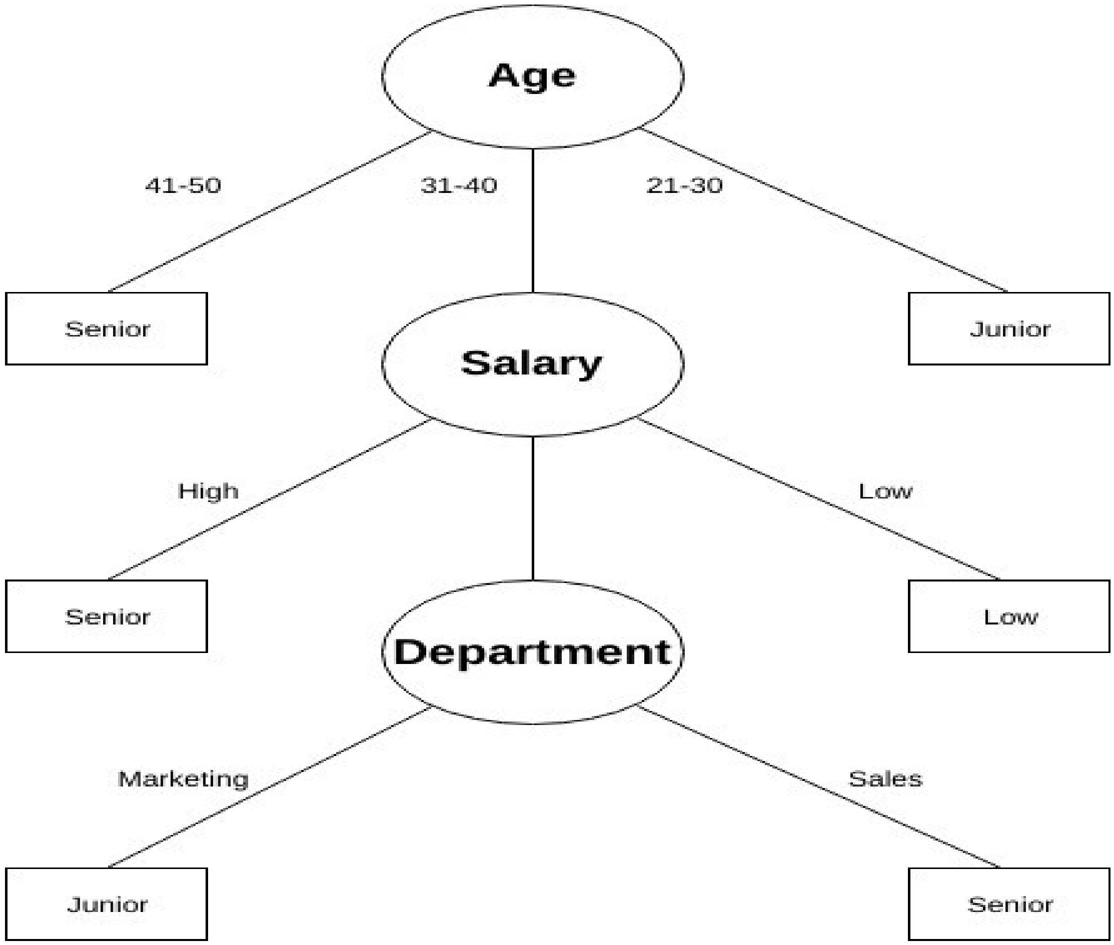
GINI(Marketing) = 1 – $(1/2)^2$ – $(1/2)^2$ = 0.5

GINI(Salary) = (4/5)*0.5 = 0.4

Since lower value for the GINI index is for Salary, we use it as the next node.
The Intermediate Decision tree is as follows:



Now we use the department attribute to split the medium salary level. Again, for Department = Marketing we are unable to get a pure split, that is we get 1 Junior and 1 Senior. Thus we arbitrarily assign it the value Junior(Because of overall majority). The final Decision tree is given below:




## b. (2 points) Compute the following accuracy on training data:

## (i) individual class accuracy

For class Junior the Accuracy is True Junior/(True Junior + False Senior) = 7/7 = 1

For class Senior the Accuracy is True Senior/(True Senior+False Junior) = 4/5 = 0.8

## (ii) overall class accuracy

Overall Class accuracy is (True Junior + True Senior)/Total+11/12 = 0.9167

**c. (2 points) For the following test data, predict the class label for each instance using the tree constructed in (a)**

The class labels for the following are filled in the table:

| | | | |
|---|---|---|---|
| Sales | 21-30 | High | Junior |
| Systems | 21-30 | Medium | Junior |
| Marketing | 41-50 | High | Senior |
| Marketing | 31-40 | Low | Junior |

**2. (10 points) Naive Bayes Classification**

**a. (1 point) State the assumption(s) made by Naive Bayes Classifier.**

The Assumption made by Naive Bayes classifier is that all attributes are equally important as well as independent when class is given. That is:-

P(X1, X2, …, Xd |Yj ) = P(X1| Yj ) P(X2| Yj )… P(Xd| Yj )

for given value of class Yj.

**b. (3 points) Consider the following dataset given belowin Table 2.**

| A | B | C | Class |
|---|---|---|---|
| 0 | 0 | 0 | + |
| 0 | 0 | 1 | - |
| 0 | 1 | 1 | - |
| 0 | 1 | 1 | - |
| 0 | 0 | 1 | + |
| 1 | 0 | 1 | + |
| 1 | 0 | 1 | - |
| 1 | 0 | 1 | - |
| 1 | 1 | 1 | + |
| 1 | 0 | 1 | + |

**Estimate the conditional probabilities for**

Conditional Probability = P(Ai|C) = Nic/Nc
where
Nc: number of instances in the class
Nic: number of instances having attribute value Ai in class C

| | |
|---|---|
| P(A=0\|+) = 2/5 =0.4 | P(A=0\|-) = 3/5 = 0.6 |
| P(A=1\|+) = 3/5 =0.6 | P(A=1\|-) = 2/5 = 0.4 |
| P(B=0\|+) = 4/5 = 0.8 | P(B=0\|-) = 3/5 = 0.6 |
| P(B=1\|+) = 1/5 = 0.2 | P(B=1\|-) = 2/5 = 0.4 |
| P(C=0\|+) = 1/5 = 0.2 | P(C=0\|-) = 0/5 = 0 |
| P(C=1\|+) = 4/5 = 0.8 | P(C=1\|-) = 5/5 = 1 |

**c. (1 point) Predict class label when A=0, B=1,C=0 using the probabilities computed from (b).**

P(+|A=0,B=1,C=0) = (P(A=0,B=1,C=0|+)*P(+))/P(A=0,B=1,C=0)
=(P(A=0|+)*P(B=1|+)*P(C=0|+)*P(+))/P(A=0)P(B=1)P(C=0)
=(2/5 * 1/5 * 1/5 * 1/2)/(1/2 * 3/10 * 1/10) = 0.533.

Now
P(-|A=0,B=1,C=0) = (P(A=0,B=1,C=0|-)*P(-))/P(A=0,B=1,C=0)
=(P(A=0|-)*P(B=1|-)*P(C=0|-)*P(-))/P(A=0)P(B=1)P(C=0)
=(3/5 * 2/5 * 0 * 1/2)/(1/2 * 3/10 * 1/10)=0
since P(+|A=0,B=1,C=0)>P(-|A=0,B=1,C=0)

Thus we classify it as '+'

**d. (3 points) Estimate the following conditional probabilities using m-estimate approach, with p = 0.5, m = 4.**

| | |
|---|---|
| P(A=0\|+) = (2+2)/(5+4) = 4/9 | P(A=0\|-) = (3+2)/(5+4) = 5/9 |
| P(A=1\|+) = (3+2)/(5+4) = 5/9 | P(A=1\|-) = (2+2)/(5+4) = 4/9 |
| P(B=0\|+) = (4+2)/(5+4) = 2/3 | P(B=0\|-) = (3+2)/(5+4) = 5/9 |
| P(B=1\|+) = (1+2)/(5+4) = 1/3 | P(B=1\|-) = (2+2)/(5+4) = 4/9 |
| P(C=0\|+) = (1+2)/(5+4) = 1/3 | P(C=0\|-) = (0+2)/(5+4) = 2/9 |

| | |
|---|---|
| P(C=1\|+) = (4+2)/(5+4) = 2/3 | P(C=1\|-) = (5+2)/(5+4) = 7/9 |

**e. (1 point) Predict class label when A=0, B=1,C=0 using the probabilities computed from (d).**

P(+|A=0,B=1,C=0) = (P(A=0,B=1,C=0|+)*P(+))/P(A=0,B=1,C=0)
=(P(A=0|+)*P(B=1|+)*P(C=0|+)*P(+))/P(A=0)P(B=1)P(C=0)
Since P(A=0)P(B=1)P(C=0) is common to both it can be ignored for comparison
=(4/9 * 1/3 * 1/3 * 1/2) = 0.02469

Now
P(-|A=0,B=1,C=0) = (P(A=0,B=1,C=0|-)*P(-))/P(A=0,B=1,C=0)
=(P(A=0|-)*P(B=1|-)*P(C=0|-)*P(-))/P(A=0)P(B=1)P(C=0)
Again P(A=0)P(B=1)P(C=0) can be ignored.
=(5/9 * 4/9 * 2/9 * ½) = 0.02743

since P(+|A=0,B=1,C=0)<P(-|A=0,B=1,C=0)
Thus now the Class label must be '-'

**f. (1 point) Compare the two methods for estimating probabilities. Which method is better and why?**

The issue with Naive Bayes is that even If one of the conditional probabilities is zero, then the entire expression becomes zero. And since these conditional probability is calculated using a sample, It may or may not represent the actual chance in the population. Thus, m estimate technique helps in reducing this extreme effect of Naive Bayes, and by giving a chance to those expressions as well where one of the conditional variables of the sample comes out to be zero.

**3. (10 points) Holt's 1-Rule method is described as shown below:**

**For each attribute a, form a rule as follows:**
**For each value v from the domain of a,**
**Select the set of instances where a has value v.**
**Let c be the most frequent class in that set.**
**Add the following clause to the rule for a:**
**If a has value v then the class is c**
**Calculate the classification accuracy of this rule.**
**Use the rule with the highest classification accuracy.**

**a) (8 points) Apply Holt's 1-Rule for the following dataset. All attributes are categorical. Show the rules and accuracy for each attribute (A, B, C).**

| A | B | C | Class |
|---|---|---|---|
| 0 | 0 | 1 | - |
| 0 | 0 | 0 | - |
| 0 | 1 | 0 | - |
| 0 | 1 | 0 | - |
| 0 | 0 | 1 | + |
| 1 | 0 | 1 | + |
| 1 | 0 | 1 | - |
| 1 | 0 | 1 | - |
| 1 | 1 | 0 | + |
| 1 | 0 | 0 | + |

**Attribute A and value = 0**
Rule : If A has value 0, then class is '**-**';
Accuracy = 4/5 = 0.8;

**Attribute A and value = 1**
Rule : If A has value 1, then class is '**+**';
Accuracy = 3/5 = 0.6;

**Attribute B and value = 0**
Rule : If B has value 0, then class is '**-**';

Accuracy = 4/7 = 0.5714;

**Attribute B and value = 1**
Rule : If B has value 1, then class is **'-'**;
Accuracy = 2/3 = 0.667;

**Attribute C and value = 0**
Rule : If C has value 0, then class is **'-'**;
Accuracy = 3/5 = 0.6;

**Attribute C and value = 1**
Rule : If C has value 1, then class is **'-'**;
Accuracy = 3/5 = 0.6;

**b) (2 points) Name the best attribute (i.e., attribute for which the total error is minimum). If there are more than 1 attribute with same accuracy, name all of them.**

The Best attribute that is for which the total error is minimum is attribute A. A has total accuracy of (4+3)/10 = 0.7. while the accuracy for the other two attributes are 0.6 for both.
Thus A is the best attribute.