Roy Chan
301202770
chanroyc@sfu.ca

CMPT 459 Fall 2017

DataMining

Martin Ester

TA: Zhilin Zhang

Assignment 3

Assignment 3.1 (40 Marks)

a) Salary has the smallest gini index of 0.16 and is chosen as the split attribute for the root.

```
gini(Age) = 5/10*gini(age=young)
            + 2/10*gini(age=medium)
            + 3/10*gini(age=old)
            = 0.5*0.48 + 0.2*0.5 + 0.3*0.44
            = 0.472

            gini(age=young) = 1 - (3/5)^2 - (2/5)^2 = 0.48
            gini(age=medium) = 1 - (1/2)^2 - (1/2)^2 = 0.5
            gini(age=old) = 1 - (2/3)^2 - (1/3)^2 = 0.44

gini(salary) = 3/10*gini(salary=low)
            + 5/10*gini(salary=medium)
            + 2/10*gini(salary=high)
            = 0.3*0 + 0.5*0.32 + 0.2*0
            = 0.16

            gini(salary=low) = 1 - (3/3)^2 = 0
            gini(salary=medium) = 1 - (4/5)^2 - (1/5)^2 = 0.32
            gini(salary=high) = 1 - (2/2)^2 = 0

gini(city) = 4/10*gini(city=Vancouver)
            + 2/10*gini(city=Burnaby)
            + 2/10*gini(city=Coquitlam)
            + 2/10*gini(city=Richmond)
            = 0.4*0.5 + 0.2*0.5 + 0.2*0 + 0.2*0.5
            = 0.4

            gini(city=Vancouver) = 1 - (2/4)^2 - (2/4)^2 = 0.5
            gini(city=Burnaby) = 1 - (1/2)^2 - (1/2)^2 = 0.5
            gini(city=Coquitlam) = 1 - (2/2)^2 = 0
            gini(city=Richmond) = 1 - (1/2)^2 - (1/2)^2 = 0.5
```

b) The gini index favors attributes with few distinct values. An attribute with few distinct values is more likely to have a low gini index. By definition, the gini index is a measure of inequality, or statistical dispersion. So the fewer amount of distinct values will result in a lower gini index. A gini index of 0 would express perfect equality.

c) b) suggests to choose the attribute with fewer distinct values between two attributes that have the same smallest gini index.

Assignment 3.2 (40 marks)

a)

P(good) = 0.6
P(bad) = 0.4

P(Age=young|good) = 0.5
P(Age=medium|good) = 0.17
P(Age=old|good) = 0.33
P(Age=young|bad) = 0.5
P(Age=medium|bad) = 0.25
P(Age=old|bad) = 0.25

P(Salary=low|good) = 0.0
P(Salary=medium|good) = 0.66
P(Salary=high|good) = 0.3
P(Salary=low|bad) = 0.75
P(Salary=medium|bad) = 0.25
P(Salary=high|bad) = 0.0

P(City=Vancouver|good) = 0.33
P(City=Vancouver|bad) = 0.5
P(City=Burnaby|good) = 0.17
P(City=Burnaby|bad) = 0.25
P(City=Coquitlam|good) = 0.33
P(City=Coquitlam|bad) = 0.0
P(City=Richmond|good) = 0.17
P(City=Richmond|bad) = 0.25

b) Age = “Young”
Salary = “high”
City = “Richmond

Result of the decision function for class “good”:

P(good) * P(Age=Young|good) * P(Salary=high|good) * P(City=Richmond|good)
= 0.6 * 0.5 * 0.3 * 0.17 = 0.0153

Result of the decision function for class “bad”:

P(bad) * P(Age=young|bad) * P(Salary=high|bad) * P(City=Richmond|bad)
= 0.4 * 0.5 * 0 * 0.25 = 0.0

The classifier predicts good.

Roy Chan
301202770
chanroyc@sfu.ca

Assignment 3.3 (20 marks)