CMPT459 Spring 2020
Data Mining
Martin Ester
TAs: Ruijia Mao and Ruchita Rozario

Solution of Assignment 1

Solution posted and discussed in class: January 30, 2020

Assignment 1.1

Given the following dataset of bank clients with categorical attributes Age, Salary, City, and Creditworthiness:

Age	Salary	City	Creditworthiness
young	high	Vancouver	bad
medium	medium	Burnaby	bad
young	low	Vancouver	bad
old	medium	Coquitlam	good
old	high	Richmond	good
medium	low	Richmond	bad
young	medium	Vancouver	bad
old	low	Burnaby	bad
young	medium	Coquitlam	good
young	medium	Vancouver	good

We want to use this dataset to train a decision tree classifier to predict the creditworthiness of a bank client, i.e. we are using Creditworthiness as the class label and the other attributes as features. Suppose that the gini index is used to determine the split attributes.

a) Which split attribute does the decision tree algorithm select for the root of the decision tree? Show all the relevant gini indexes that you have to compute.

The algorithm chooses City for the root, since City has the smallest gini index of 0.25.

gini(D) = $1 - \Sigma\, p_i^2$ for i from 1 to 2, i.e. for the classes "good" and "bad"

gini(D|age) = 5/10*gini(D|age=young)
+ 2/10* gini(D| age=medium)
+ 3/10* gini(D| age=high)
= 0.5*0.48 + 0.2*0.0 + 0.3*0.44 = 0.24 + 0.0 + 0.132
= 0.372

gini(D|age=young) = $1-(3/5)^2-(2/5)^2 = 1-13/25=12/25=0.48$
gini(D| age=medium) = $1-(0/2)^2-(2/2)^2 = 1-1 = 0.0$
gini(D| age=old) = $1-(1/3)^2-(2/3)^2 = 1-5/9=0.44$

gini(D|salary) = 3/10*gini(D|salary=low)

+ 5/10* gini(D|salary=medium)
+ 2/10* gini(D|salary=high)
= 0.3*0.0 + 0.5*0.48 + 0.2*0.5 = 0.0 + 0.24 + 0.1
= 0.34

gini(D|salary=low) = $1-(3/3)^2-(0/3)^2 = 1-1= 0.0$
gini(D|salary=medium) = $1-(2/5)^2-(3/5)^2 = 1-13/25 – 12/25= 8/25=0.48$
gini(D|salary=high) = $1-(1/2)^2-(1/2)^2 = 1-1/2=0.5$

gini(D|city) = 4/10*gini(D|city=Vancouver)
+ 2/10* gini(D|city=Burnaby)
+ 2/10* gini(D|city=Coquitlam)
+ 2/10* gini(D|city=Richmond)
= 0.4*0.375 + 0.2*0.0 + 0.2*0.0 +0.2*0.5= 0.15 + 0.0 + 0.0 + 0.1
= 0.25

gini(D| city=Vancouver) = $1-(1/4)^2-(3/4)^2 = 1-106/16= 0.375$
gini(D| city=Burnaby) = $1-(2/2)^2-(0/2)^2 = 1-1–0= 0.0$
gini(D| city=Coquitlam) = $1-(0/2)^2-(2/2)^2 = 1-1= 0.0$
gini(D| city=Richmond) = $1-(1/2)^2-(1/2)^2 = 1-1/4 –1/4= 0.5$

b) Does the gini index favor attributes with few or with many distinct values, or in other words, is an attribute with few or with many distinct values more likely to have a low gini index? Explain your answer. What rule does this suggest to break ties, i.e. to select one attribute among multiple attributes that all have the same smallest gini index?

The gini index favors attributes with many distinct values. The more distinct values, the smaller the partitions of the training dataset corresponding to the individual attribute values, and the more likely the partitions are pure with respect to the class label. In the extreme case, the partitions contain only one training record and achieve the minimum gini index of 0. This suggests to break ties by selecting, among the attributes that all have the same smallest gini index, the attribute with the smallest number of attribute values.

Assignment 1.2

Given the above dataset of bank clients with categorical attributes Age, Salary, City, and Creditworthiness. We have trained a Naïve Bayes classifier to predict the creditworthiness of a bank client, i.e. we are using Creditworthiness as the class label. Here are the parameters of the classifier (computed without using Laplacian smoothing, and using C as abbreviation for Creditworthiness):

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

P(Age=young|C=good) = 0.5 P(Age=medium|C=good) = 0.0 P(Age=old|C=good) = 0.5
P(Age=young|C=bad) = 0.5 P(Age=medium|C=bad) = 0.34 P(Age=old|C=bad) = 0.16

P(Salary=low|C=good) = 0.0 P(Salary=medium|C=good) = 0.75
P(Salary=high|C=good) = 0.25
P(Salary=low|C=bad) = 0.5 P(Salary=medium|C=bad) = 0.34
P(Salary=high|C=bad) = 0.16

P(City=Vancouver|C=good) = 0.25 P(City=Burnaby|C=good) = 0.0
P(City=Coquitlam|C=good) = 0.5 P(City=Richmond|C=good) = 0.25

P(City=Vancouver|C=bad) = 0.5 P(City=Burnaby|C=bad) = 0.34
P(City=Coquitlam|C=bad) = 0.0 P(City=Richmond|C=bad) = 0.16

a) What class label (Creditworthiness) does the Naïve Bayes classifier predict for an unseen client with Age = young, Salary = low and City = Vancouver? Show the necessary computations.

The classifier predicts the Creditworthiness to be bad.

For good:
P(C=good) * P(Age =young|C=good) * P(Salary = low|C=good) * P(City = Vancouver|C=good)
= 0.4 * 0.5 * 0.0 * 0.25 = 0.0

For bad:
P(C=bad) * P(Age = young|C=bad) * P(Salary = low|C=bad) * P(City = Vancouver|C=bad)
= 0.6 * 0.5 * 0.5 * 0.5 = 0.075

b) Given the Naïve Bayes assumption and the parameters of the above Naïve Bayes classifier, what is the probability of observing a client with Age = old, Salary = medium and City = Vancouver? Show the necessary computations.

The probability of a client with these attribute values is the marginal probability aggregating over the conditional probabilities of all classes, i.e.

P(Age =old, Salary = medium, City = Vancouver) =
P(Age =old, Salary = medium, City = Vancouver|good) * P(good)
+ P(Age =old, Salary = medium, City = Vancouver|bad) * P(bad)

According to the Naïve Bayes assumption,
P(Age =old, Salary = medium, City = Vancouver|good)=
P(Age =old|good) * P(Salary = medium|good) * P(City = Vancouver|good) =
0.5 * 0.75 * 0.25 = 0.09375
P(Age =old, Salary = medium, City = Vancouver|bad)=
P(Age =old|bad) * P(Salary = medium|bad) * P(City = Vancouver|bad) =
0.16 * 0.34 * 0.5 = 0.0272

In conclusion,
P(Age =old, Salary = medium, City = Vancouver)
= 0.09375 * 0.4 + 0.0272 * 0.6 = 0.0375 + 0.01632 = 0.05382

The probability of such a client is 0.05382.