Zhuo Liu
CSCI-630-02 Foundations of FIS
Homework3.5

1 (20 pts) For each of the following machine learning scenarios, several possible input variables are listed. For each variable, briefly justify whether you think it would be useful, theoretically useful but impractical to obtain (for training and/or testing), or not useful. Also, suggest how accurate you think such a system could be, given the input variables you have chosen and a feasible amount of training data, and why.

a. Predicting whether it will rain in Rochester tomorrow: whether it rained in Rochester today; whether it rained in Cleveland yesterday; the locations in the US where it rained yesterday; the wind speed in Rochester yesterday; the day of the week; the month of the year
b. Predicting whether a song will become a hit: previous sales figures of the artist; length of the song; company releasing the song; lyrics of the song
c. Predicting whether you will like a particular restaurant: the opinions of the last hundred people who ate there; the Yelp review of the restaurant; the type of food; the number of insects in the kitchen
d. Predicting the final score of a given Premier League football (soccer) match: the scores of each team's last 5 games; the predicted weather at game time; the number of fans in attendance at the match; the hours of practice by each member of each team in the previous week

a Predicting whether it will rain in Rochester tomorrow

In my opinion, this system would not be accurate, because the weather conditions are always unpredictable and there are various factors which can influence this forecast system

The amount of training data can be 200 which collect last two year statistic data

variable: whether it rained in Rochester today

reason: it will be useful, because two consecutive days of rainfall are associated with each other

variable: whether it rained in Cleveland yesterday

reason: it is useful, because Cleveland and Rochester are close to each other, we can use the weather condition of Cleveland for reference, the weather conditions are nearly the same

variable: the locations in the US where it rained yesterday

reason: it is not useful, because the locations we choose might be too far away from Rochester, the accuracy of the data can not be guaranteed

variable: the wind speed in Rochester yesterday

reason: it is useful, because wind speed is a key factor which can influence on whether it will rain in Rochester, because wind can bring the rainfall

variable: the day of the week

reason: it is not useful, because the day of the week can not affect whether it will rain or not in Rochester

variable: the month of the year

reason: it is not useful, because the month of the year and whether it will rain in Rochester are irrelevant factors

so the feature and type of input variables are shown as followed:

whether it rained in Rochester today;	variable:Yes,No
whether it rained in Cleveland yesterday;	variable: Yes,No
the wind speed in Rochester yesterday	variable: level of strong,medium,weak

b Predicting whether a song will become a hit

In my opinion, this system will be inaccurate, because the variables are fewer, people's taste is hard to tell

The amount of training data could be 500 of past 5 years, because this system tends to be inaccurate

variable: previous sales figures of the artist

reason: it is useful, if the artist is already famous, people tend to see whether his/her next song is good or not

variable: length of the song

reason: it is not useful, whether a song is popular or not does not depend on the length of the song,

variable: company releasing the song

reason: it is useful, if the company is powerful enough, it will help propagate the release of the song

variable : lyrics of the song

reason: it is theoretically useful but impractical to obtain, because it is hard to give a criteria of lyrics of the song

so the feature and type of input variables are shown as followed:

previous sales figures of the artist;	variable:1,100,1K,10K etc
company releasing the song;	variable:a number of company's name

c Predicting whether you will like a particular restaurant

I think this system would be accurate, because personal preference of specific food can be inferred from data in software such as Yelp

The amount of training data should be 100, because of the location of people's house, they can not go to as many as restaurant variable: the opinions of the last hundred people who ate there

reason: it is useful, because it is highly possible that my preference is among the last hundred people, so their opinions are useful to me

variable: the Yelp review of the restaurant

reason: it is useful, because many people tend to use Yelp software before they go to a restaurant, so the review of the Yelp can affect people's opinion

variable: the type of the food

reason: it is useful, if there are limited types of food in a restaurant, people will have limited options, so they may not want to visit

variable: the number of insects in the kitchen

reason: it is useful but impractical to obtain, it is hard to calculate indeed how many insects are in the kitchen

so the feature and type of input variables are shown as followed:

the opinions of the last hundred people who ate there; variable:5 representatives of 100 people

the Yelp review of the restaurant; variable:top three reviews which are listed in the software

the type of food; variable:2 or 3 main dishes of the restaurant

d Predicting the final score of a given Premier League football (soccer) match

From my perspective, this system can not be accurate, because there are only limited options of variables in this system

The amount of training data should be 38, because one team will play 38 matches per season

variable: the scores of each team's last 5 games

reason: it is useful, because we can get a statistic figure of the scores of each team's last 5 games, and then we can use this figure to predict the final score

variable: the predicted weather at game time

reason: it is useful but impractical to obtain, because we can predict the weather condition in one day or in a week, but it is hard to predict weather condition in a short period, it will be inaccurate

variable: the number of fans in attendance at the match

reason: it is not useful, because the performance of a team does not depend on the audience, it depends on the hard work of each player in the team

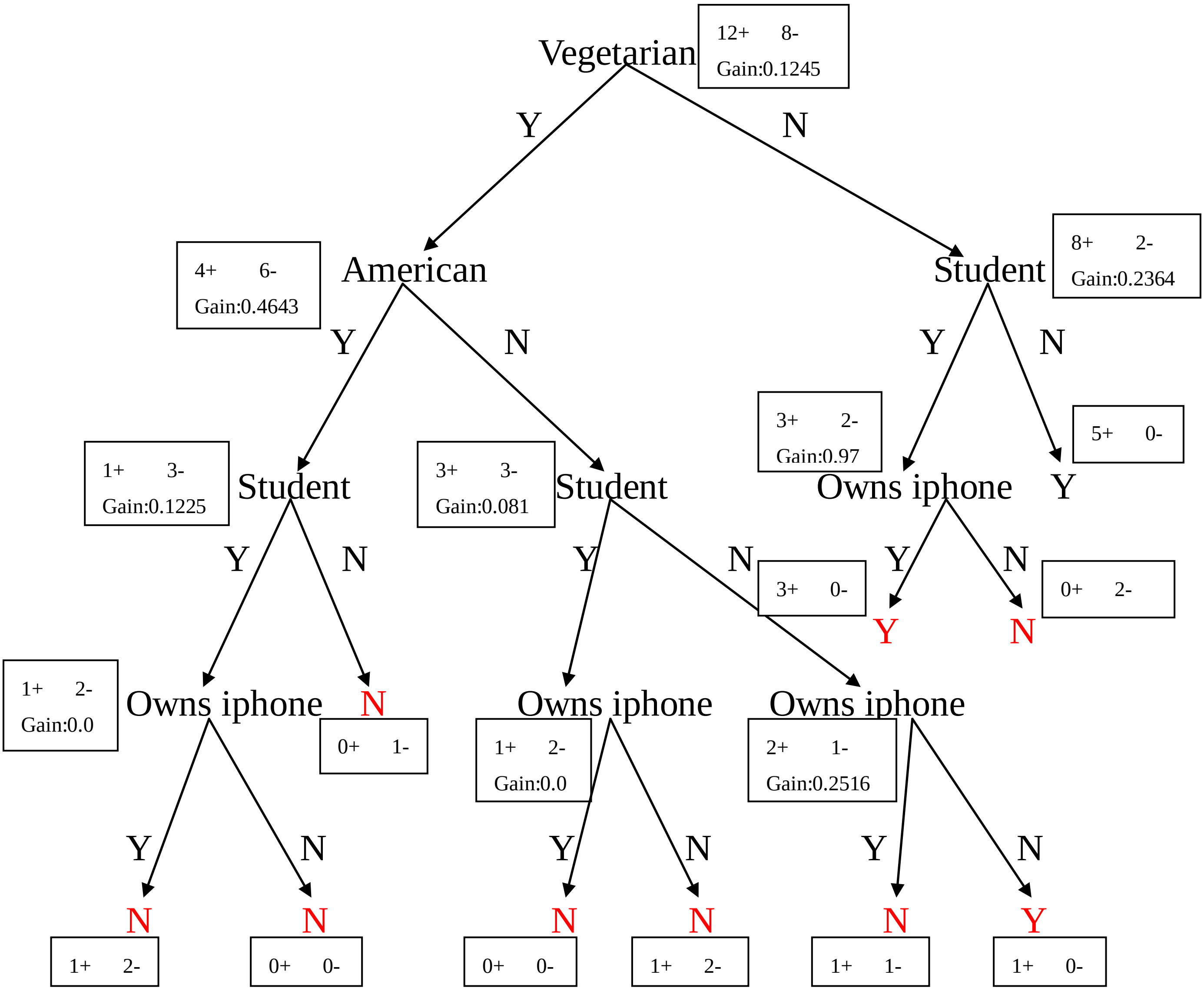
variable: the hours of practice by each member of each team in the previous week

reason: it is not useful, in my opinion, the performance of a team does not depend on the hours in the previous week, it relies on the years of practice

so the feature and type of input variables are shown as followed:
the scores of each team's last 5 games; variable:each score number of last 5 games

2 (30 pts) For the following data set, build a decision tree (using information gain) that best answers the question of whether a person drinks coffee based on the four given criteria. You should continue asking questions as long as the question gives some information gain up to a maximum tree depth of three; if you find more than one equally useful criteria at any point, pick any one. If you run out of useful questions to ask at any particular point, state that as well. You are free to use some code, or just do it by hand - you do not need to show all your entropy calculations (or code to do them), but do show them for the root node, and give at least the gain for each other node in your tree.

Student?	American?	Vegetarian?	Owns iPhone?	Drinks Coffee
N	Y	N	Y	Y
N	N	N	N	Y
Y	Y	Y	Y	Y
Y	N	Y	N	N
N	N	Y	Y	N
N	Y	N	N	Y
Y	Y	N	Y	Y
N	Y	Y	N	N
Y	N	N	Y	Y
Y	N	N	N	N
Y	Y	Y	Y	N
N	Y	N	N	Y
N	N	Y	Y	Y
Y	N	Y	N	N
Y	N	N	Y	Y
N	N	Y	N	Y
Y	Y	Y	Y	N
N	Y	N	N	Y
Y	N	Y	N	Y
Y	Y	N	N	N



$$gain = Entropy(root) - \frac{left}{total} * Entropy(leftnode) - \frac{right}{total} * Entropy(rightnode)$$