NYU | STERN

[redacted]

Homework #2

Due: turned in on NYU Classes by 11:59pm Mon 4/3/2019

__Zhengyi Chen__

(put your name above)

_______zc969______

(put your Net ID above)

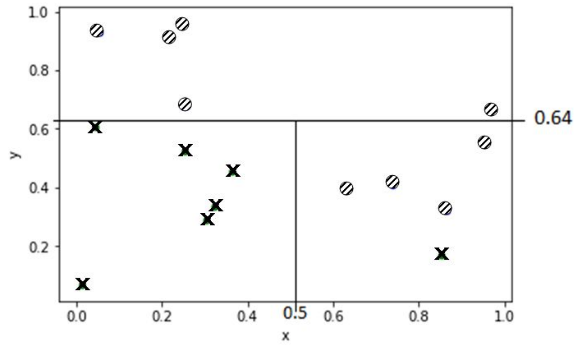
Total grade: ________ out of ________ points

Please answer all questions/follow all directions. ***<u>Put your name above, and include your last name in the filename of your homework submission</u>. Include all of the requested material in a compressed file (.zip, .tar.gz, .rar, etc)***

1) Label each case as describing either data mining process (DM), or the use of the results of data mining (Use). To help you answer this question, revisit the definition of data mining.

a) _Use _ Choose customers who are most likely to respond to an on-line ad.
b) _DM_ Discover rules that indicate when an account has been defrauded.
c) _DM_ Find patterns indicating what customer behavior is more likely to lead to response to an on-line ad.
d) _Use__ Estimate probability of default for a new credit application.
e) _Use_ Predict whether a customer is pregnant.

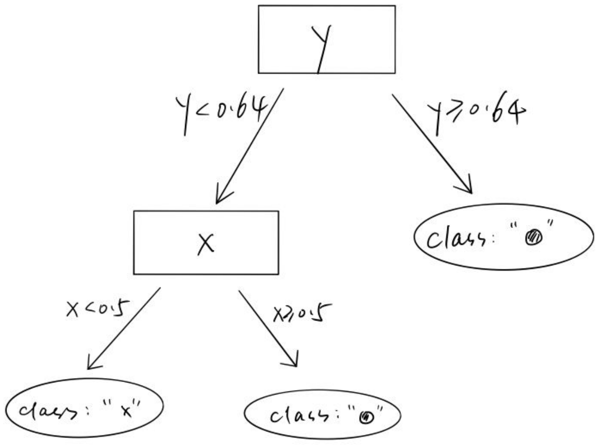
2) You are given a 2D plot, with instances that belong to 2 classes: circles and x's. After learning a Decision Tree classifier on that dataset, you get the decision boundaries shown in the image below.



a) Present the resulting Decision Tree model. You can show it as a tree structure *or* as the rules that you'll get out of them.

For the tree structure, include an image (here or w/ your submission) that you drew by hand, via MS / Mac paint, or any other tool you like. For each node in your tree, include the split condition (if any). Report the <u>class value</u> only in the leaf nodes. You do not need to report other values for any of the nodes.

***Alternatively*, write the *rules* that correspond to the model. For each rule, the conditions need to be in the order as you would follow them in the tree model, i.e., the 1st condition you write must be at the root, the 2nd condition must be at layer 1, and so on.**



b) Briefly (2-3 sentences) explain why the first feature was selected.

From the plot of data, all "x" class data are below the line of y = 0.64, which means this feature partitions the current group into two subgroups which are as pure as possible.

Also, splitting at y = 0.64 gives the maximum of information gain(IG).

c) What is the probability that a new instance with (x,y) = (0.7, 0.2) belongs to class “x” ? What is the probability for the same instance to belong to class “x” with Laplace correction?

The new instance (x,y) = (0.7, 0.2) belongs to the region in the second leaf node in the decision tree model above (count from left to the right), since this the region of this leaf node corresponds to the rectangular in the bottom right corner in the 2D plot.

There are 4 circles and 1 “x“ in this region, so the probability that the new instance belongs to “x” is: $\#\ of\ "x"/\#\ of\ data\ =\ 1/5$

The Laplace correction is: $p("x")\ =\ (\#\ of\ "x"+1)/(\#\ of\ data\ points\ +2)\ =\ (1+1)/(5+2)\ =2/7$

3) Soc-Marketeer Inc. is a social media marketing agency that can help you “place your ad-content in front of 100K accounts” on a popular social network. They have a varied *clientele who deal in sports, fashion, tourism, entertainment, education, domestic retail, technology and other domains.* They charge (effectively) a flat rate for their 100K outreach, but you’ve heard that their customers very often see a large discrepancy between the advertised (100K) and real traffic that they receive from the social network, leaving them thinking they’ve overpaid.

Assume that these 100K are all real people (no bots) with verified accounts (no duplicates). Soc-Marketeer really has access to them, and all users see all ads. The users are knowingly targeted with ads and all legal paperwork is in order. Focus on data science-related problems.

You are confident you can (help them) do better! In half to one page (max), using concepts we’ve introduced in class – in discussions and Jupyter Modules - and the corresponding chapters in the textbook, write a brief, *yet precise*, business proposal. Your answer should (at least) include the following:

a) Explain what the root cause of the traffic discrepancy is.

b) To reduce this discrepancy, several methods can be done.

c) What data science *task* would you use to reduce the discrepancy?

d) What information would you need for your task?

e) Who would provide the data that you’d need for your task? Should you involve others?

f) Present ***at least 5*** different features that you think i) you can get and ii) will help you to address the problem. Be (very!) precise (e.g. “frequency of” is not precise).

g) If the task is supervised, give a (precise) definition of your target variable.

h) Give an example of an instance of your data.

i) How would you evaluate your suggestion?

The advertisement are of various backgrounds since the clientele of Soc-Marketeer Inc involve in different industries, for example, sports, fashion, tourism, entertainment, education, domestic retail, and technology. Users react to advertisement differently based on their individual preferences. Hence, presenting advertisement over 100K accounts will not bring the same amount of traffic from social network, since users will not respond to the advertisement if they are not interested in that industry. This also explains why there is a traffic discrepancy between the advertised (100K) and real traffic that the clientele receive.

To reduce this discrepancy, several methods can be done. First, company should present users with the advertisements that are related to their interest field. This can largely increase the rate of successful advertisement spreading. For example, an athlete would be have a higher probability to click on the sports advertisements than an artist. Also, company should classify the percentage of time that a user respond to an advertisement. If a user never respond to an advertisement or has a very low percentage overall regardless of his/her interest, it would be a waste to send him/her advertisements again and again since we know the probability of responding is very low. Low percentage responding rate may be caused by advertisement blocking functions set individually on computer.

Hence, the data science task needed to reduce discrepancy is classification. Classification helps to split the accounts into different groups and can be used in predicting if a person would respond to a certain kind of advertisement. Similarity matching can be used to identify similar individuals based on their history data, since similar individuals have higher probability to react to advisements in a same way. Profiling (behavior description) can be applied to characterize the typical behavior of people who are tend to respond to a certain type of advertisement. In our situation, classification is the most important and efficient task and is focused in the following analysis.

The history data about whether a user ever responded to a certain type of advertisement and some background information about the user are needed for the classification.

The data needed can be collected from previous clientele regarding if a user responses to the advertisement. Individual background information can be provided by the users themselves while registering their accounts in social media.

The examples of features are as following:

1) The history probability that a user respond to certain type of advertisement. (User A has a 80% probability to respond to sports advertisements)
2) The age of the user (User A is 20 years old)
3) The occupation of the user (User A is an university student)
4) The gender of the user (User A is male)
5) The interest of the user (User A follows 10 famous sports accounts in the social media)

The task is supervised since our model studies the relationship between a set of selected features and a target variable. The target variable, in our case, is whether a user would respond to a certain type of advertisement.

The example of instance would be: User A is a 20 years old male university student, who following 10 sports accounts in social media, has a probability to respond to sports advertisements based on his history.

To evaluate the model, several methods can be employed. First, cross validation can be used to test the model performance and help to select a model which performs the best. Then, confusion matrix can also be applied to calculate accuracy and precision. The performance of decision tree can be improved by studying the learning curve and the performance of logistic regression can be improved by adding a regularization parameter to the objective function.

4) Due to time pressure, your firm is thinking about retaining the services of a consultant for your personalized advertising campaign, instead of building it in-house. You are asked to attend the meeting given your data science expertise, to help with the screening process.

During their presentation, they mention that they achieved an accuracy score of *91%* on a public, well-known dataset that you have worked on extensively. You find this result surprising, because following a rigorous evaluation methodology, you had only achieved accuracy around 74%, using the same modeling techniques that they presented: Decision Trees, Logistic Regression, SVMs.

Write below 2-3 different questions that you should ask them, to assess whether the reported 91% is truly reflective of their model's effectiveness. Your questions must be specific, in the sense that their respective answer should be short (few words, "yes" / "no" answers). For example, asking "*what was your evaluation methodology?*" (or similar) is not a valid question.

Questions:

Did you seperate your data set into training data and testing data, and using the test data to test the accuracy of your model?

Did you study the fitting graph for tree induction and chose the depth at the "sweet spot"?

5) Hands-on section: Log in to JupyterHub and start your server, if needed. Using the Jupyter Notebooks' controls, navigate to your class material folder (/notes/gv760/data_mining_spring2019/). In there, you should find a folder called "assignments" containing a Python notebook for the hands-on part. Follow the instructions in the notebook. Once you're done, download the notebook and include the notebook – and any other needed code – in your submission file.

<u>**Reminder:**</u> ***You will need to be connected to an NYU network to access JupyterHub, either by being on campus and connecting directly to one of NYU's networks, or via a VPN connection. Instructions for setting up a VPN conection, depending on your platform:***

For Mac Users: link

For Windows Users: link