EMORY UNIVERSITY

Machine Learning II

Homework #1 **Due: turned in by Monday 01/27/2020 before class**

Adil Ashish Kumar

(put your name above)

Total grade: _____ out of ___100___ points

There are 2 parts in this homework assignment with 17 numbered questions in total. Please answer them all and submit your assignment as a single PDF or Word file by uploading it to the HW1 drop-box on the course website.

Part I. Multiple Choice Questions (60 points; 4 points each)

1. **Which three are essential skills for a data scientist? (Choose three)**
   A. Software engineering Yes
   B. System administration
   C. Statistics and mathematics Yes
   D. Domain experience Yes
   E. Multidimensional data modeling
   F. Database administration

2. **What are three common practices of data scientist working with big data? (Choose three)**
   A. Applying machine learning techniques Yes
   B. Working with large and disparate data Yes
   C. Minimizing data redundancy through normalization
   D. Cleaning and transforming data Yes
   E. Publishing findings in academic journals

3. **What is the first step in the lifecycle of a typical data science project?**
   A. Identify the desired outcome
   B. Capture the data
   C. Determine which data is required
   D. Define the problem Yes

4. **Why should a data scientist iterate through the project lifecycle multiple times during one project? (Choose three)**
   A. To change techniques as the amount of data scales up Yes
   B. To benefit from lessons learned during the lifecycle Yes
   C. To recover from a cluster node failure
   D. To incorporate new problems and new data discovered during the lifecycle Yes
   E. Because the statistical significance of an inference did not reach the required threshold

5. **What are two reasons a data scientist conducts a preliminary analysis using a small amount of data? (Choose two)**
   A. Because simple, smaller analyses are superior to complex, larger analyses
   B. Because the results of large-scale analyses are difficult to communicate
   C. Because initial small-scale analyses can inform later large-scale analyses Yes
   D. Because a small-scale analysis can be used to validate an approach correct Yes

6. **When considering data quality and provenance, a data scientist should seek to acquire data that is: (Choose two)**
   A. Accurate Yes
   B. Structured
   C. Raw
   D. From a reputable source Yes

7. Which are three valid reasons for anonymizing data? (Choose three)

A. Anonymization simplifies the data transformation process
B. Laws, policies, and standards may require anonymization Yes
C. Anonymization mitigates the damage from intrusions Yes
D. Anonymization makes re-identification impossible
E. Anonymization protects against legal liability in the event of an attack Yes

8. Which model typically yields the best predictions of future data?

A. A model that fits a general pattern observed in the data correct Yes
B. A model that exactly fits the observed data

9. Which question can be answered using linear regression?

A. Based on its text content, what is the probability that a message is spam?
B. What is the relationship between the monthly fees charged to a customer and the likelihood they cancel their subscription?
C. What is the relationship between a customer's age and the number of minutes per day they spend using a service? Yes
D. What is the distribution of customer age?

10. What is a supervised learning algorithm?

A. An algorithm that discovers structure in data where no formal structure exists
B. An algorithm that must be monitored by the data scientist as it runs
C. An algorithm that automatically determines its ideal behavior to maximize a measure of performance
D. An algorithm that requires a training dataset with known labels correct Yes

11. Which should you consider when training supervised machine learning algorithms?

A. A binary classification algorithm must be trained on a dataset consisting exclusively of records with a positive value of the label
B. An algorithm typically performs best when trained on a dataset consisting of only the most relevant attributes Yes
C. The accuracy of an algorithm typically degrades as the volume of training data increases
D. A superior algorithm operating on a smaller training dataset is typically more accurate than a simpler algorithm operating on a larger training dataset

12. Data suggests that the average listening session for Earcloud users is longer for women than for men. You are designing an experiment to try to confirm this.
Your experiment confirms with high confidence that the average Earcloud user session is longer for women than for men. In the terminology of hypothesis testing, what decision is made?

A. Accept the null hypothesis
B. Reject the null hypothesis Yes
C. Accept the alternative hypothesis
D. Reject the alternative hypothesis

13. **Before running an A/B test on your recommender system to confirm that the cancellation rate for women decreases when they are recommended longer playlists, you will first run an A/A test. What should the A/A test do?**
    A. Recommend longer playlists to both women and men
    B. Recommend longer playlists to women and shorter playlists to men
    C. Divide women into two groups and make no change to the recommended playlists in either group yes
    D. Divide women into two groups and recommend longer playlists to one group

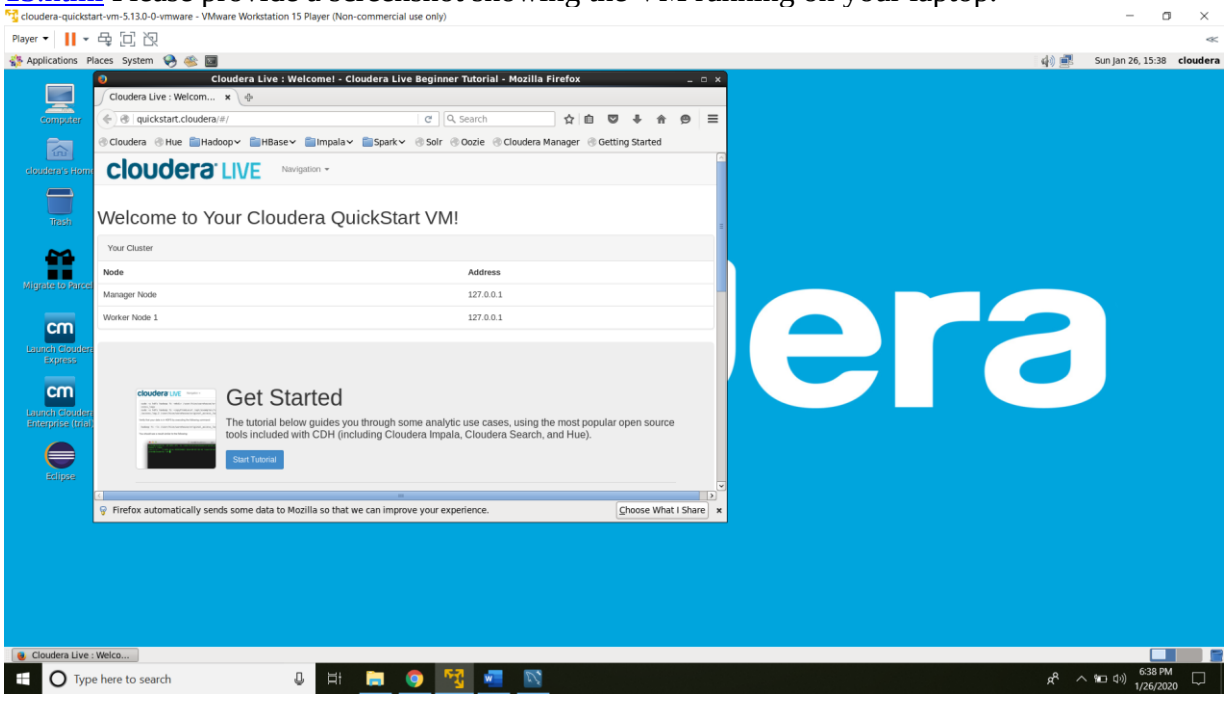
14. **What is a best practice when deploying a change to a recommender system?**
    A. Deploy the change to all users in one step so results from multiple experimental groups are not intermingled in log files
    B. Deploy the change to 50% of users so the experimental and control groups are of equal size
    C. Deploy the change to a small proportion of users then ramp up in phases to limit possible negative impacts Yes
    D. Run the original and changed versions on alternating days to isolate negative impacts by day

15. **When is it most appropriate to use a lambda function in Python?**
    A. When the function will be used only once Yes
    B. When the function will be reused multiple times
    C. When the function has no input parameters
    D. When the function has no return value

Part II. Hands on (40 points)

1. Run the Cloudera QuickStart VM 5.13 (10 points)

For this question, you should download the Cloudera QuickStart VM (5.13 version) and run it on your local machine. The VM can be downloaded from here: https://www.cloudera.com/downloads/quickstart_vms/5-13.html Please provide a screenshot showing the VM running on your laptop.



2. Predictive Analytics Application (30 points)

Select a publicly available dataset (e.g., from the UC Irvine Machine Learning Repository, Kaggle, etc.) of your interest and then build and evaluate a predictive model of your choice using any tool \ programming language you prefer.

In general, use best practices when building and evaluating your models: optimize parameters on validation data, perform final evaluation on test data, etc.

As part of this question, you should present a brief overview of your predictive modeling process and discuss your results. Your overview should i) clearly describe the specific predictive task you selected and the class it belongs to (e.g., classification, clustering, etc.) as well as cover ii) why you selected the particular machine learning algorithm for the predictive task you selected. Make sure you also present iii) information about the model "goodness" (possible things to think about: confusion matrix, predictive accuracy, classification error, precision, recall, f-measure, ROC curves).

The presented overview should not exceed 2 pages.

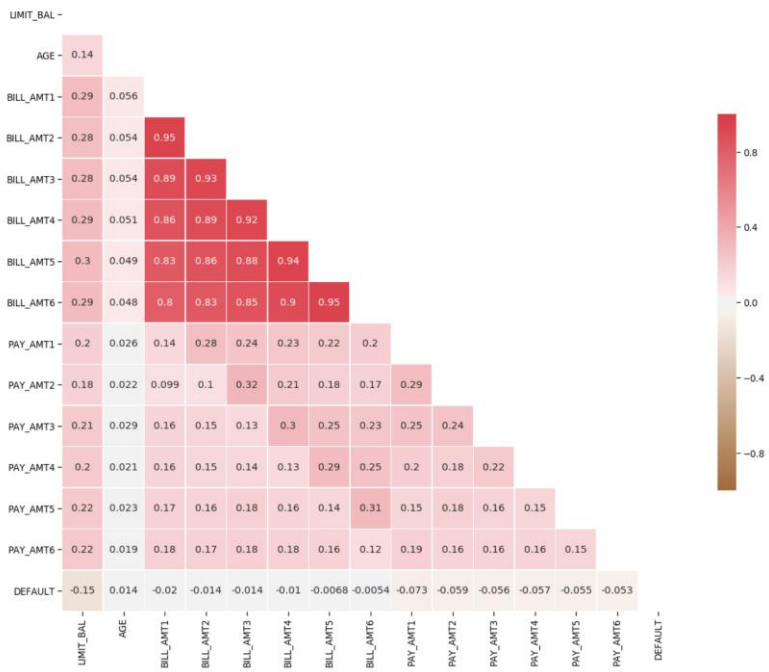
Data set: UCI ML repository: default of credit card clients Data Set
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Problem: Classification problem - Predict if the customer will default on the next payment
Data Description:

Variable	Description
ID	Customer ID
LIMIT_BAL	Credit card limit
Sex	1=male, 2=female
Education	1=graduate school, 2=university, 3=high school, 4=others
Marriage	1 = married; 2 = single; 3 = others
Age	Customer age in years
PAY_0 – PAY_6	Past payment history status from September to April 2005 -2 = no amount due, -1 = pay duly, 0 = revolving credit, 1-8 = payment delay for 1-8 months
BILL_AMT1-6	Credit card bill amount from September to April 2005
PAY_AMT1-6	Previous month amount paid from September to April 2005
default payment next month	Target Variable (1=yes, 0=no)

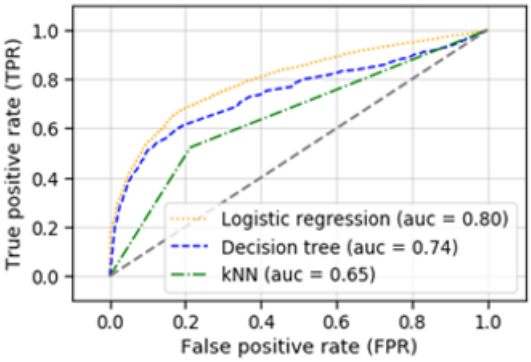
Since it is a classification problem, I will consider classification algorithms like Logistic regression, KNN and decision trees.

First I ran basic EDA on data to understand it in detail. There is 1 ID column, 23 predictors and 1 target variable – 30000 observations. Below I see that some variables are highly correlated with each other – bill amount variables. This could lead to multicollinearity in model. I also noticed that credit limit and age variables have right skewed distributions, so I can try applying transformations to fix it. Further, most of the categorical variables have categories with very less frequency so I can club these into "others" category



- After the EDA, I transformed the skewed continuous variables to reduce skewness in their distributions, grouped redundant categories for categorical variables and combined the bill and pay amount variables to get pay ratio variables.
Then I ran Decision tree, KNN and Logistic regression models. I split the data into train and test in ratio 70:30. I then trained the 3 models on training data and figured out the best tuning parameters by using inner and outer k fold cross validation methods, where data in trained on inner k folds (k=5) and performance is validated on outer 5 folds of data. To do this I used the gridsearchcv function and optimized the following parameters:
Logistic regression – C(regularization strength) , penalty – lasso or ridge
Decision tree – max_Depth, splitting criterion, min_samples_leaf, min_samples_split
KNN- no of neighbors, weights(uniform or distance)

Once I had the optimum tuning parameters for all 3 algorithms, I ran these models on the test data set. Below are the ROC curve and performance measures.



Algorithm	Precision (positive class)	Recall (positive class)	F1 score (positive class)	Accuracy
Logistic regression	0.59	0.61	0.6	0.79
KNN	0.46	0.52	0.49	0.72
Decision tree	0.63	0.52	0.57	0.8

Based on auc from ROC curve, I would choose to us logistic regression method as final model since it has highest AUC. Below is the confusion matrix for the logistic regression model