Project 1 - Classification

CSCI 5523 - Introduction to Data Mining
University of Minnesota

Due - October 28, 2020

Instructions and Experiments

Note: Please read the entire project description before you begin. The goal of this project is to analyze the performance of classification algorithms on several synthetic and real-world data sets. This will be done in the following steps:

- First, you will explore the data sets.
- Next, you will perform a series of experiments on which you will be asked to answer a series of questions. For these experiments, you will be running a python Jupyter notebook.
- Compile your answers in the form of a report.

Python Jupyter Notebooks

We recommend installing Jupyter using Anaconda as it will also install other regularly used packages for scientific computing and data science. Some pointers to setup Jupyter notebooks on your system:

- Video link - *https://www.youtube.com/watch?v=MvN7Wdh0Juk*
- Medium Link - *https://medium.com/@neuralnets/beginners-quick-guide-for-handling-issues-launching-jupyter-notebook-for-python-using-anaconda-8be3d57a209b*
- Tutorials link - *https://www.dataquest.io/blog/jupyter-notebook-tutorial/*, *https://www.youtube.com/watch?v=3C9E2yPBw7s*

Before you Begin

- Visually explore the data sets in the experiments below, and consider the following:

  types of attributes

  class distribution

  which attributes appear to be good predictors, if any

  possible correlation between attributes

  any special structure that you might observe

  Note: The discussion of this exploration is not required in the report, but this step will help you get ready to answer the questions that follow

- Use precision and recall to measure performance.

- Your goal is to learn everything that you can about the dataset. Answer the questions below as a starting point, but you should dig further. What more can you discover? The goal of this assignment is to give a helping hand for you to discover the most interesting and surprising things.

Report and Submission

- Collect output from your experiments. Submit all Jupyter notebook (cell displaying output) electronically as a single zipped file using the Project 1 Canvas submit tool. A submission not adhering to this policy will not be graded and you will get zero.

- Write a report addressing the experiment questions. **The report has to be submitted in PDF format electronically using Project 1 Canvas submit tool**. Your project will be evaluated based only on what you write on the report.

- If you are a UNITE student, you should upload your Jupyter notebook (cell displaying output) on canvas like other students and submit your project report via UNITE as homework.

- Your Jupyter notebook should be submitted electronically - we will look at your output if something is ambiguous in your report. Copy and paste the output from the Jupyter notebook into your report only to the limited extent needed to support your answers.

1 Problem 1 [20 points]

The files for this problem is under Experiment 1 folder. Datasets to be used for experimentation: **telecom_churn.csv**. Jupyter notebook: **Exploratory data analysis.ipynb**. In this experiment, we will do exploratory data analysis to get a better sense of data. The dataset contains record of telecom customer along with the label "churn". Churn = "true" signifies that the customer has left the company and churn = "false" signifies that the customer is still loyal to the company. Answer the following questions

1. How many records are there in the dataset?
2. How many features are there? Name each feature and assign it as binary, discrete, or continuous.
3. As a data scientist, your job is to build a model that identifies customers intending to leave your company. To do that, we prepare our data for the machine learning model. We can have the most advanced algorithm, but if our training data is terrible, our result will be poor. According to your intuition, which features are irrelevant. Briefly explain your reasoning.
4. Are there any missing values in the data?
5. For the continuous features, what is the average, median, maximum, minimum, and standard deviation values?
6. What is the average number of customer service calls made by a customer to the company?
7. In our dataset, data comes from how many states?
8. What's the distribution of the "Churn" feature. Is the feature skewed?
9. What's the highest and lowest "total day charge" encountered by the customer? If we sort the dataset in ascending and descending order by "total day charge," what observation can you make regarding the connection between "total day charge" and "churn" rate?
10. What's the average number of customer service calls made by the user who has churned out of the company? Compare and contrast it with the average number of customer service calls made by the user who is still with the company.
11. Compare and contrast the average values of numerical features for churned and non-churned users? As a data scientist, what strategy will you recommend to the company to retain more customers?
12. Assume you have devised a model which states that if "international plan" = 'no', then the customer will not churn (i.e., "churn" = False). Report accuracy, precision and recall concerning "churned" class.

13. Calculate P(churn = True | international plan = 'yes'), P(churn = False | international plan = 'yes'), P(churn= True | international plan = 'yes'), P(churn = False | international plan = 'no'). Given that the customer has churned, what are the probabilities that the customer has opted/not-opted for the international plan? Similarly, given that the customer has not churned, what are the probabilities that the customer has opted/not-opted for the international plan?

14. Given that the customer has opted for an international plan, what are the chances of the customer staying/not-staying with the company. Compare and contrast with the customer who has not opted for an international plan.

15. Calculate the probability of customers leaving the company, given that he has not made any customer service call. Compare and contrast it with the customer making 1,2,3,4,5,6,7,8,9 customer service calls. Plot the probability of customers leaving the company as customer service calls increase.

16. Assume you have devised a model which states that if "international plan" = "yes" and the number of calls to the service center is greater than 3, then the customer will churn (i.e., "churn" = True). Report accuracy, precision and recall concerning "churned" class.

2 Problem 2 [20 points]

The files for this problem is under Experiment 2 folder. Datasets to be used for experimentation: **telecom_churn.csv**. Jupyter notebook: **Decision Trees and kNN.ipynb**. In this experiment we will apply and visualize decision trees, kNN, finetune parameters and learn about k-fold cross validation etc. To visualize decision tree we need additional packages to be installed i.e. **Graphviz** and **pydotplus**. Answer the following questions:

1. For the synthetic dataset, we separate two classes by training a decision tree. What does the boundary look like when we overfit (max_depth ≥ 4) and underfit (max_depth = 1) the decision tree on data. For both cases, paste the decision tree and the decision boundary from Jupyter notebook output.

2. Decision tree classifier $sklearn.tree.DecisionTreeClassifier$ has parameter "max_depth" which defines the maximum depth of the tree and "criterion" which measure the quality of the split. What happens if we don't specify any value for both parameters?

3. For Bank Dataset, what are the 5 different age values that the decision tree used to construct the split the tree? What is the significance of these 5 values?

4. Briefly describe the K-nearest neighbor algorithm? What is the importance of "K" in K-nearest neighbor? Shortly describe the default metric used in sklearn.neighbors.KNeighborsClassifier ?

5. Why do we do cross-validation? Briefly describe the importance of it.

6. For the customer churn prediction task, we show that the accuracy of the decision tree is 94% when max_depth is set to 5. What happens to accuracy when we leave the value of max_depth to its default value? Explain the rise/fall of accuracy.

7. Given a dataset d, with n sample and m continuous features, what does Standard Scaler *sklearn.preprocessing.StandardScaler* do? Given dataset d = [[0, 0], [0, 0], [1, 1], [1, 1]], write down its scaler transformation.

8. What does GridSearchCV *sklearn.model_selection.GridSearchCV* do? Why is it important for decision trees and k-nearest neighbors?

9. How many decision trees do we have to construct if we have to search the two-parameter space, max_depth[1-10] and max_features[4-18]? If we consider 10-fold cross-validation with the above scenario, how many decision trees do we construct in total?

10. For the customer churn prediction task, what is the best choice of k[1-10] in the k-nearest neighbor algorithm in the 10-fold cross-validation scenario?

11. For MNIST dataset, what was the accuracy of the decision tree [max_depth = 5] and K-nearest neighbor [K = 10]? What are the best parameters and accuracy for holdout dataset for decision trees when we used GridSearchCV with 5 fold cross-validation?

3 Problem 3 [20 points]

The files for this problem is under Experiment 3 folder. Datasets to be used for experimentation: **spam.csv**. Jupyter notebook: **Naive Bayes Spam.ipynb**. The dataset contains 5,574 messages tagged according to ham (legitimate) or spam. In this experiment we will learn about text features, how to convert them in matrix form and Naive Bayes algorithm. Answer the following question :

1. How many records are there? What's the distribution of the "label" class. Is it skewed?

2. How many unique SMS is there in the dataset? What is the SMS that occurred most frequently and what is its frequency?

3. What is the maximum and minimum length of SMS present in the dataset? Plot the histogram of the length of SMS with bin size 5,10,20,30,40,50,100,200. What do you hypothesize with the plots?

4. Plot the histogram of the length of SMS for each label separately with bin size 5,10,20,50 i.e. histogram of the length of all ham SMS and histogram of the length of all spam SMS. What can you perceive after examining the plots?

5. In the Bag of words approach, we convert all strings into lower cases. Why did we do that, and why is it important? Can we convert all strings into the upper case and still fulfill our original goal?

6. What does CountVectorizer achieve? What will happen if we set stop_words = "english". Give five examples of stop-words in English.

7. Given a dataset, how do we generate a document-term matrix? Do we first generate document-term matrix and then separate matrix into train/test or first separate the data into train/test and then generate document-term matrix based on train dataset and afterwards generate matrix for test set? Explain your reasoning.

8. Using bag of words approach, convert documents = ['Hi, how are you?', 'Win money, win from home. Call now.', 'Hi., Call you now or tomorrow?'] to its document-term matrix.

9. How many features are created while making document-term matrix for SMS dataset? Can you think of a method to reduce the number of features? List the pros and cons of the method.

10. For our input dataset, which Naive Bayes model should we use, Gaussian Naive Bayes or Multinomial Naive Bayes? Explain your reasoning ? Report accuracy, precision, recall and F1 score for the spam class after applying Naive Bayes algorithm.

4 Problem 4 [20 points]

The files for this problem is under Experiment 4 folder. In this assignment, we provide three real-world datasets for classification, i.e., **Iris dataset** (*https://archive.ics.uci.edu/ml/datasets/Iris*), **Thyroid dataset**
(*https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease*), and **Diabetes dataset**
(*https://www.kaggle.com/uciml/pima-indians-diabetes-database*). Also, we give three jupyter notebooks, one for each dataset, in which we have applied k-nearest neighbor, decision-tree, and Naive Bayes Algorithm without any parameter tuning.
Write a short report for each dataset; you can be as innovative as you want, giving your analysis about the dataset, your observations, and comments. It should be no more than half a page to a page in length. It can include a description of the dataset, the number of observations, missing value or not, testing strategy you deployed, classification accuracy of algorithms, etc.

5 Problem 5 [20 points]

In this experiment, you will participate in a Kaggle competition **Titanic: Machine Learning from Disaster**. Kaggle is the world's largest community of data scientists that host data science competitions that can be intimidating for beginners to join. The goal of the assignment is to make you familiar with the Kaggle competitions and give you a path forward to hone your data science skills after you have completed the data mining class. By participating in the competition, you will have the opportunity to learn from other people's machine learning code and contribute to the community. On a side note, some of the listed competitions have over $\$1,000,000$ prize pools.

Proceed to the following link *https://www.kaggle.com/c/titanic/overview* and follow the description meticulously. Then check out Alexis Cook's Titanic tutorial *https://www.kaggle.com/alexisbcook/titanic-tutorial* and get ready to make your submission. In a few seconds, your submission will be scored, and you'll receive a spot on the leaderboard. For each of the questions below, write down the accuracy obtained, position on the leaderboard, and put the screenshot of the leaderboard where your username, rank, and accuracy are clearly visible.

1. Submit example gender_submission.csv file that predicts that all female passengers survived, and all-male passengers died.

2. Submit the gender_submission.csv file that predicts that all passengers survived.

3. Submit the gender_submission.csv file that predicts that all passengers died.

4. Submit the model output of the random forest model as detailed in the tutorial. Try to play with the number of decision trees (we constructed 100) and see if accuracy improves.

5. Copy and Edit the kernel in *https://www.kaggle.com/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy*. Submit the model output from the kernel, write a short (half page to a page) report on what the kernel does and include the position on the leaderboard, screenshot as detailed above.