Data Mining

Final Term Project

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

CS 634 Final Term Project

- This is a single person project.
- On the first page of the project report, indicate your first name, last name, NJIT UCID, preferred email address, and the option number you choose (otherwise you will lose 5 points).
- There are six options. You choose one of the six options.

Submission Rules

- Submit ONE SINGLE file. Embed your last name, first name in your project file name. For example, if your name is John Smith, then your file name should read: smith_john_finaltermproj.doc. Only a doc or pdf file is accepted. No tar/zip/rar is allowed. Your final term project will automatically lose 10 points if this submission rule is violated.

- This is a single person project.

- Submit your final term project file to Canvas under Final Term Project Submission Site before the deadline.

Late Project Policy

A project is late if it is not submitted to Canvas before the deadline. If you turn in your project *n* days late, your total point will be deducted by (50 × *n*) points. For example, suppose you turn in your project 1 day late (if you turn in your project after the deadline on the due date, it is also considered as 1 day late). Then, you lose (50 × 1) = 50 points automatically, and your total point is 50 points. Further, suppose you lose 10 points in documentation. Thus, you receive (50 –10) = 40 points in total.

For all late submissions of the project, they must be emailed to me at wangj@njit.edu. The email subject and the file name in the email must both be called

Lastname_Firstname_finaltermproj.doc

(where you should fill in your last name and first name).

Note: Each student should submit one final term project only. If the student has submitted his/her final term project (even incomplete) in Canvas, the student is NOT allowed to send another final term project to wangj@njit.edu. Your project will automatically lose 80 points if this rule is violated.

Final Term Project: Option 1

Supervised Data Mining (Classification)

- This option is to implement 2 classification algorithms of your choice on 1 dataset of your choice (each of the 2 algorithms must run on the dataset).
- Your final term project documentation must indicate clearly the algorithms and dataset you used in the project.

Final Term Project: Option 1
General Sources of Algorithms/Software

http://davidmlane.com/hyperstat/Statistical_analyses.html

http://statpages.org/javasta2.html

http://pcp.sourceforge.net/

http://www.cs.waikato.ac.nz/ml/weka/

http://www.r-project.org/

http://mlflex.sourceforge.net/

Han Book 3rd ed Chapters 8, 9

Kumar Book 1st ed Chapters 4, 5

Final Term Project: Option 1
Specific Algorithms and Tools used in the Project

- There are 5 categories of algorithms listed on the following pages (categories 6-10 are software tools and platforms). The 2 classification algorithms you choose must come from two different categories. In the same category, only one algorithm can be chosen from that category.
- In your final term project documentation, for each algorithm you choose, specify clearly the category number/name and algorithm name in that category.

Category 1 (Support Vector Machines)

- LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) linear kernel
- LIBSVM polynomial kernel
- LIBSVM radial basis function (RBF) kernel
- LIBSVM sigmoid kernel
- Gist (http://www.chibi.ubc.ca/gist/) Pick a kernel of your choice and specify the kernel used in your project

Category 2 (Random Forests)

- CRAN (http://cran.r-project.org/web/packages/randomForest/index.html)
- Willows (http://c2s2.yale.edu/software/Willows/)
- Weka (https://www.cs.waikato.ac.nz/ml/weka/)

Category 3 (Decision Trees)

Refer to Weka

http://www.cs.waikato.ac.nz/ml/weka/

- ADTree
- J48 (C4.5)
- LMT
- M5P
- NBTree

Category 4 (Bayesian Networks)

- Weka BayesNet
  https://www.cs.waikato.ac.nz/ml/weka/

- JBNC (http://jbnc.sourceforge.net/)

Category 5 (Naïve Bayes)

Refer to Weka:

https://www.cs.waikato.ac.nz/ml/weka/

- AODE
- ComplementNaiveBayes
- NaiveBayes
- NaiveBayesMultinomial
- NaiveBayesSimple
- NaiveBayesUpdateable

Category 6 (R Package)

- Refer to

  http://www.r-project.org/

  Pick any classification tool in R.

Category 7 (Mathematical Package)

- MATLAB available at NJIT

Category 8 (RapidMiner)

www.rapidminer.com

http://sourceforge.net/projects/rapidminer/

Category 9 (Weka)

https://www.cs.waikato.ac.nz/ml/weka/

Category 10 (Python)

https://scikit-learn.org/

Final Term Project: Option 1

Sources of Data

http://archive.ics.uci.edu/ml/

http://www.cs.ucr.edu/~eamonn/time_series_data/

http://aws.amazon.com/datasets

http://www.trustlet.org/wiki/Repositories_of_datasets

Final Term Project: Option 1

Submission (One File)

A word or pdf file (final term project report) containing:

- Source code of your classification algorithms.
- The websites where the software and complete dataset can be downloaded.
- All related documentation and documents including the manual you developed and **screenshots** showing the running situation and input/output of your programs. This report should be written in a tutorial style to explain through **screenshots** and examples how to run your tools on the dataset you choose.

Final Term Project: Option 1
Submission (cont.)

- In the final term project report, each student must present experimental results that show the comparison of classification accuracies between the two classification algorithms used in the project.
- In evaluating classification accuracy, each student should use the 10-fold cross validation method (if your algorithms predict labels) or present ROC and AUC (if your algorithms predict continuous probability values).

Final Term Project: Option 2

Unsupervised Data Mining (Clustering)

Part 1

Generate a set S of 500 points (vectors) in 3-dimensional Euclidean space. Use the Euclidean distance to measure the distance between any two points. Write a program to find all the outliers in your set S and print out these outliers. If there is no outlier, your program should indicate so. Use any programming language of your choice (specify the programming language you use in the project).

Next, remove the outliers from S, and call the resulting set S'.

Final Term Project: Option 2

Part 2

(1) Write a program that implements the hierarchical agglomerative clustering algorithm taught in the class to cluster the points in S' into *k* clusters where *k* is a user-specified parameter value.

(2) Repeat part 1 and (1) above on two additional different datasets.

Notes on the hierarchical agglomerative clustering algorithm

In determining the distance of two clusters, you should consider the following definitions respectively:

- the distance between the nearest two points in the two clusters,
- the distance between the farthest two points in the two clusters,
- the average distance between points in the two clusters,
- the distance between the centers of the two clusters.

Use the definition that yields the best performance where the performance is measured by the Silhouette coefficient.

Final Term Project: Option 2

Submission (One File)

A word or pdf file (final term project report) containing:

- Source code of your clustering algorithm.
- The website where the complete datasets can be downloaded.
- All related documentation and documents including the manual you developed and **screenshots** showing the running situation and input/output of your programs. This report should be written in a tutorial style to explain through **screenshots** and examples how to run your tool on the datasets you choose.

Final Term Project: Option 3

This option is to implement a graph mining or graph clustering system. You can use any heuristics published in data mining articles and implement the heuristics using any programming language of your choice (specify the programming language you use in the project). Use any data set of your preference (specify the source of data, e.g. http://snap.stanford.edu/data/index.html). The output should be displayed using a visualization tool such as Graphviz or Cytoscape, etc. (specify the name of the visualization tool you use).

Final Term Project: Option 3

Submission (One File)

A word or pdf file (final term project report) containing:

- Source code of your graph clustering algorithm.
- The website where the complete dataset can be downloaded.
- All related documentation and documents including the manual you developed and **screenshots** showing the running situation and input/output of your programs. This report should be written in a tutorial style to explain through **screenshots** and examples how to run your tool on the dataset you choose.

Final Term Project: Option 4

This option is to implement a text mining system.

You can use Reuters-21578 dataset at

http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection

To parse the documents, you can use

Stanford CoreNLP tools at

http://nlp.stanford.edu/software/corenlp.shtml

Represent each document by a set of keywords.

Find and print out association rules among the keywords or terms in the documents.

Final Term Project: Option 4

Submission (One File)

A word or pdf file (final term project report) containing:

- Source code of your text mining algorithm.
- The website where the complete dataset can be downloaded.
- All related documentation and documents including the manual you developed and **screenshots** showing the running situation and input/output of your programs. This report should be written in a tutorial style to explain through **screenshots** and examples how to run your tool on the dataset you choose.

Final Term Project: Option 5

Data Mining Using Hadoop/Spark in the Cloud

- This option is to implement a data mining algorithm (e.g., association mining, classification, clustering, etc.) of your choice on a dataset of your choice using Hadoop/MapReduce/Spark.

- Your final term project documentation must indicate clearly the algorithm, the dataset and the cloud environment you used in the project.

Final Term Project: Option 5

Infrastructure and Software

- Cloud infrastructure: Amazon EMR

- Programming software: Hadoop/MapReduce/Spark on an AWS cluster in a master slave fashion with multiple nodes using a programming language of your preference (e.g., Java)

Final Term Project: Option 5

Submission (One File)

A word or pdf file (term project report) containing:

- Source code of your data mining algorithm implementation using Hadoop/MapReduce in the cloud.
- The website where the complete dataset can be downloaded.
- All related documentation and documents including the manual you developed and **screenshots** showing the running situation and input/output of your programs. This report should be written in a tutorial style to explain through **screenshots** and examples how to run your programs on the dataset you choose.

Final Term Project: Option 6
Deep Learning

- When you choose this option, I will provide deep learning source code and datasets developed by our machine learning and data mining group and posted on GitHub. You will submit your project report directly to me, not to Canvas, before the due date.

- Email me at wangj@njit.edu to get details of this option.

Project Grading

- There is a limit on the file size in Canvas. So, keep your project file small to avoid any problem that may occur when submitting the file in Canvas.
- The project file should contain the source code and documentation including **screenshots**. The screenshots are used to demonstrate the running situation of your programs, particularly how the programs execute and produce output based on different input data and user-specified parameter values.