CS378: Introduction to Data Mining, Final Project Description

November 3, 2006

For your final project, you will be required to perform a data mining task relating to the Netflix prize dataset. The dataset is available from the course website (the password will be emailed to the cs378 email list). *Do not distribute this dataset or make it otherwise publicly available.* For the project, you will be given a subset of the data. Details about the format of the data will be supplied in a readme document included in the dataset tar file.

As a reminder, the final project is worth 25% of your final grade. Before the final project deadline, you will have two intermediate deadlines: a project proposal deadline and a project progress deadline. For your project proposal, you should list your top three choices. Further, if you feel very passionate about any one of your choices, you can optionally write a brief paragraph about why you prefer your choice. Your final project will be due approximately Dec. 10.

Project Options

Listed below are several possible projects. For each of projects 1-3 (listed below), you will be need to focus on one of the following two types of feature sets.

- *A. Sparse Features:* In this representation, each reviewer is represented as a sparse vector of the set of movies he/she reviewed. Similarly, each movie should be represented by the set of reviewers who reviewed the movie.
- *B. Induced Features:* These are defined as features that are computed indirectly via properties of the dataset. Two basic such features are (1) average rating of reviewers for a given movie, or (2) average rating of movies for a given reviewer. Other examples include temporal features. For example, a number of previous reviews a reviewer has submitted, or a feature indicating whether the specified date is a holiday.

  In addition to coming up with your own features, you must also use the following induced features:

  - Average movie rating
  - Average reviewer's ratings
  - Movie year (this information is included in the supplied data)
  - Weekend vs. Weekday
  - Holidays (the choice of holidays used is up to you)

Project 1: Classification

The goal of this project is to predict Netflix scores by building a classifier.

Your classifier type should be one of the following:

- K-nearest neighbor. Predict movie ratings based on nearest neighbors' ratings. Your algorithm should use either sparse feature types or induced feature types. You should experiment with various values of k, various types of distance measures and different normalization schemes.
- Parametric classification. Using the WEKA machine learning package (Google 'weka'), you should test classifiers such as Decision Trees, Support Vector Machines, etc. Your classifiers should use either sparse feature types or induced feature types.

You should develop a methodology to test your classifier. Is cross-validation an appropriate tool to use here? You should report accuracy and confusion matrices. You should also propose and evaluate one or two other metrics for evaluating the quality of your predictions. Perhaps some reviewers are easier to predict than others?

Project 2: Clustering

This project option involves unsupervised clustering of the Netflix dataset. You can either cluster using vector clustering algorithms (via the G-Means clustering package), or cluster using a graph-based representation of your data (using the Graclus package).

For vector-based clustering, you should download a software package called G-Means (Google 'gmeans'). This package includes k-means as well as other variants. Your project should focus on clustering using either the sparse or the dense feature representations described above.

For graph-based clustering, you should download Graclus (Google 'graclus'). Here, you must form a similarity matrix that compares all pairs of movies, or a matrix that compares all pairs of reviewers. You should experiment with standard similarity measures (e.g. correlation). You should also propose and experiment with your own similarity measure(s). You may need to use a smaller subset of the data to compute this for all pairs of movies or pairs of reviewers. Additionally, you should come up with your own similarity measures.

Finally, you must analyze your results, both quantitatively and also qualitatively.

Project 3: Recommendations

This project requires you to build a system that recommends movies to users. You should use a K-nearest neighbor approach here. Your project should focus on either the sparse or induced feature representation. Given a specific reviewer, your algorithm should find his/her nearest neighbors and determine which movies among the neighbors are (1) very popular, (2) rated highly, and (3) not rated by the candidate reviewer. As with classification, you should experiment with different distance measures, normalization schemes, various values of k, etc.

You must devise a reasonable measure to quantify your results. Your algorithm should be trained on data before a given date (i.e. this is the training set) and should be evaluated based on reviews occurring only after the given date (i.e. the test set). A good prediction of a movie X for reviewer Y would imply that Y rated X highly in the test set.

Project 4: Data Cube

Load the Netflix data into a data cube. There are several open source Data Cube projects - Mondrian seems like it is the most mature. Once your data cube is implemented, you give a

thorough analysis of the data, provide interesting plots and draw conclusions. Your project should also test the speed of the data cube implementation. How long does a typical data cube query take. For simple queries, how much faster is the cube over a basic grep of the raw data file?

Project Options Summary

There are 10 different possible projects:

- 1a: Classification: KNN, Sparse Features
- 1b: Classification: KNN, Induced Features
- 1c: Classification: Parametric Classifiers, Sparse Features
- 1d: Classification: Parametric Classification, Induced Features
- 2a: Clustering: Vector-based Clustering, Sparse Features
- 2b: Clustering: Vector-based Clustering, Induced Features
- 2c: Clustering: Graph-based Clustering
- 3a: Recommendations: KNN, Sparse Features
- 3b: Recommendations: KNN, Induced Features
- 4: Data Cubes

Project Progress Requirements

For your progress deadline, you must compute the following properties on the data set:

- Average review scores across all reviewers
- Distribution of these scores
- Top 10 most highly rated movies
- Number of reviews as a function of time (discretized at a month level)
- The reviewer whose score distribution has the highest entropy

You should also investigate five other interesting properties of the dataset that are relevant to your project. Be sure to describe how each of these properties are relevant to your project. This project progress report is due in class on Thursday, Nov. 16.

Final Project Deadline

Your final project writeup should be approximately 7-10 pages. This writeup should include the following

- Overview of the project and your analysis of the dataset.
- Description of the algorithms you used and the associated feature representation

- Overview of your experimental methodology. Specifically, how are you quantifying your results?
- Results: Provide a detailed analysis of your results. This should include several tables and graphs. Be sure to address any surprising or unexpected findings.

You should turn in all code (submitted electronically, more details later) that you used. Any third party software used should be acknowledged in your writeup.