CS5691 : Pattern Recognition and Machine Learning Data Contest

Saurabh Jain (ME16B071) and Abhijeet Panpaliya (EE16B069)
IIT Madras

April 25, 2019

1 Movie Rating Prediction - Problem Description

The dataset given is a subset of ml-20m which describes 5-star rating and free-text tagging activity from Movie Lens, a movie recommendation service. It contains around 7 million ratings, split into about 5 million training and 2 million test ratings.

Each sample consists of a usedId , movieId and rating. The ratings are on a scale of 5 in steps of 0.5 (0.5,1.0, ...,5.0).

Number of Training Samples : 5264336

Number of Testing Samples : 2304988

2 Data Analysis

COMPLETE DATASET	
Number of Users	9970
Number of Movies	9998

.

TRAINING DATASET	
Number of Users	7632
Number of Movies	7823

TESTING DATASET	
Number of Users	5061
Number of Movies	9992

Now, the most important information regarding this, is the **number of users** and **movies in the testing dataset which are not present in the training dataset.**

TESTING BUT NOT TRAINING	
Number of Users	2338
Number of Movies	2175

Number of Movies for which **Genome Scores** are not given : **342,** for all other movies genome scores are available. Thus, the features of 9656 movies out of 9998 movies are known.

3 Approach

3.1 Movies

- All the Genome Scores of movies **(1128 dimension features)** are clustered into **10 clusters** using **K-Means Clustering Algorithm.** The number of cluster was set as a hyper-parameter and the best value came out to be 10.

- **Principal Component Analysis (PCA)** is then used to project these features onto a 512 dimensional feature space.
- A training matrix containing reduced genome scores (512 dimensional) for the set of movies present in the training dataset is created and the average rating of these movies is stored in another vector (score vector).
- A **Linear Regression Model** is then fit over the training matrix and the score vector.
- Now, for the complete training dataset, this Linear Regression Model is used to predict the movie rating. These rating take in consideration only the features of movies and user specific rating are not yet added.
- Now, for the movies for which genome scores were not available, mean of the ratings for those movies in training set is used. For the movies which are not present in the training dataset, overall mean over the training dataset is used.

3.2 Users

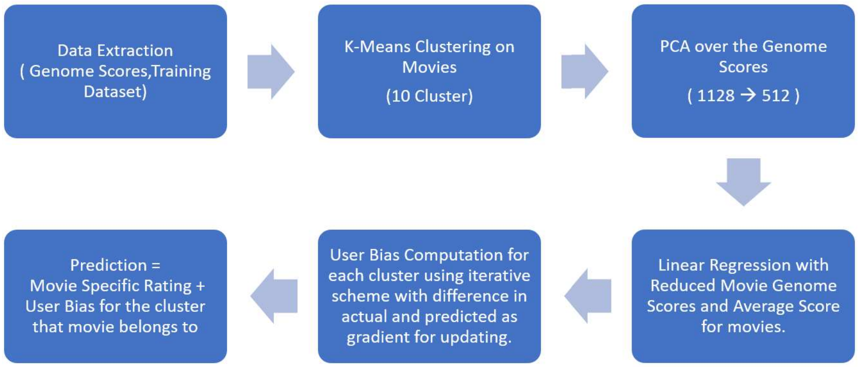
- For every user, a bias is trained corresponding to each of the 10 cluster. The training is done by running multiple iterations and sequentially updating user bias based on the difference in actual and predicted rating.

$$\text{Predicted Rating} = \text{Movie Specific Prediction} + \text{User Bias}$$

- Suitable learning step is chosen to make sure training error reduces at every iteration.
- If the user bias for a particular set of cluster is zero, the mean of that user over the clusters where bias is not zero is used.
- For the users, which are not there in training dataset, the bias for a cluster is defined as the mean of the bias of all the users in that particular cluster.

3.3 Prediction

- The prediction for every sample in the testing dataset is the sum of the Movie Specific Prediction and User Bias for the cluster that movie belongs to.
- If the predicted value is greater than 5.0, it is changed to 5.0. Similarly the prediction is changed to 0.5 . if it is less than 0.5.
- This prediction is rounded off to one decimal place and used as submission.


Data Extraction
(Genome Scores,Training Dataset)
K-Means Clustering on Movies
(10 Cluster)
PCA over the Genome Scores
(1128 → 512)
Linear Regression with Reduced Movie Genome Scores and Average Score for movies.
User Bias Computation for each cluster using iterative scheme with difference in actual and predicted as gradient for updating.
Prediction = Movie Specific Rating + User Bias for the cluster that movie belongs to

3.4 Performance of Model

Performance (MSE)	
Training Data	0.6341
Test Data	0.76668

4 Other Approaches

4.1 Baseline Model

Each prediction is sum of a User specific bias, Movie specific bias and a Constant bias. The model is trained with an iterative scheme with gradient as the error in prediction. .

$$\text{Prediction} = \text{B[User]} + \text{C[Movie]} + \text{U}$$

The testing accuracy came out be **0.8381**

4.2 Logistic Regression over Movies

The training dataset is trained using **Logistic Regression Model**. Multi- Class Logistic Regression with 10 classes was used with userId, movieId and genome_scores as the feature matrix.

The testing accuracy came out be **1.08352.**

4.3 User Specific Model

Linear Regression model for every user was trained based on the samples in the training data corresponding to that user. Thus, 9970 linear models were used.

The testing accuracy came out to be **1.28199.**