Report

**Linear regression**

Firstly, we select some original features to do simple linear regression. We choose age, gender, occupation and genres. Cross validation shows that Mean Absolute Prediction Error is 0.91.

Then we try to add some new features. We notice that in the dataset, there are features describing users and there are features describing movies. However, none of these features capture a match between a user and a movie. Intuitively, users have preferences for specific genres of movies. According to work by Yashodhan, a UCSD student who met same issue when dealing with movieLens dataset, we decide to define a new feature to capture this as follows. For a user u, we define a vector ug where ugi is the average rating given by the user for movies which include the genre i. For a movie m, we define a vector mg that denoting the genres listed for that movie. Finally we compute the dot product of vectors ug and mg then divide it by the number of genres ng. We compute the value (ug\*mg)/ng and use it as a feature in linear regression.

We add this new feature to linear regression and use cross validation to test this model. Mean Absolute Prediction Error becomes 0.79, which is a great improvement. Also we take a look into weighting of each feature and find that weighting of this new feature is relatively high, which suggests that this new feature that capture the match between user and movie.

Also, we notice that some users tend to give more liberal ratings while others tend to be more cautious. Similarly, some items tend to be rated more highly than others. So we introduce two additional features. One is average rating given by each user and the other is average for each movie. Cross validation shows this new linear regression model performs really well. Mean Absolute Prediction Error is 0.71.

Consider that male and female users may show different tendency when get old, we introduce a new feature that age\*gender. But it does not make big progress.

**Cross validation**

We use cross validation to test our model. We divide our dataset into train set and test set randomly. 95% of original dataset is train set and other 5% is test set. We use Mean Absolute Prediction Error (MAPE) to calculate error between true value and prediction.