COMP7404

Project Report

Topic: Collaborative Filtering - Latent Factor Model

Group C

1. Introduction  
    How to make a recommendation to customer is a challenging problem in variety areas including retail, financial industry, e-marketing and search engineer. For example, in an online shop system, we are trying to recommend items to a particular customer. Our recommendation should match the interest or preferences of the customer. If we can make an accurate recommendation, it can boost up the sales income and turnover.  
     
    Recommender System is a system that tries to produce a list of recommendations based on the available information. In general, there are two approaches: content-based filtering and collaborative filtering.  
     
    In this report, we will briefly talk about different approaches and methodologies at the beginning. Then we will focus on one methodology, Latent Factors Model, of collaborative filtering.
2. Content-based Filtering  
    Content-based filtering is based on the attributes or features of customer and item. For example, in an online shop system, when we make a recommendation to a particular customer, we look into the historical data (such as purchase record or borrowing record) of that particular customer. We find out the attributes of those item in historical data, then we can suggest items with a similar attributes.  
     
    In this approach, we need to have the attributes of each item or customer. However, collecting that information is time-consuming. Imagine that, we are working with 10,000 items. We need to label all item one by one.

1. Collaborative Filtering  
    In contrast to content-based filtering, collaborative filtering does not require the attributes or features of item or customer. The idea of collaborative filtering is using other customer’s behavior in the past to make prediction of interest or preference of a particular customer. For example, if two customers (A and B) have bought same set of item in the past, we can assume that these two customers are “similar”. Based on this assumption, we can find out items which bought by customer A but not customer B and then recommend it to customer B.  
     
    There are many difference methodology to collaborative filtering problem. We can classify these method into three category: memory-based, model-based and hybrid.  
     
    Memory-based methodology is try to measure the similarity between customer or item. For example, if item A and item B are similar, then we can recommend item B to customer who bought item A in the past. This example is a item-based methodology. Similarly, there is user-based methodology.  
     
    Model-based methodology is try to build a model between customer and item, then using this model to do prediction. There are few difference model-based methodology including latent factors model, Clustering model, Bayesian network, principal component analysis (PCA).  
     
    Hybrid methodology is a combination of memory-based and model-based method.
2. Problem formulation

In a movie recommendation system, we have two entities, customers and movies. Customers have rated the movies they had watched and we would like to introduce other movies that they may be interested based on rating from themselves and a rating from other customers. Based on those rating information, we can construct a “customer to item” matrix to represent the preference between user between user and movie. However, there are many missing value in this matrix. Before we can predict which movie a given customer may interested, what we need to do is to fill in all the missing values of the matrix. In other words, we are trying to predict the rating of movies customers give even they have not watched it.

1. Latent Factors Model  
    Latent factors model, sometimes called matrix factorization, is a mathematical technique to decompose one matrix into two. The observable variables matrix Y (m x n) is decomposed and can be approximated by two latent variables matrices theta (m x k) and X (n x k), where k << m, n.   
     
    Recommendation based on hidden factors found using the latent factor model is more robust than using content-based filtering approach which compares observable item similarities. When using latent factor model, we will define the cost function as the difference between the observable matrix Y and the predicted matrix, which is the dot product of the two latent variables matrices. We can then compute the gradients of the cost function with respect to the two latent variables matrices and minimize the cost function using gradient descent.
2. Demonstration  
    To demonstrate the idea of collaborative filtering using latent factors model, we will use a dataset with about 900 users with 1600 movies. There are basically three variables: the name of the movies, the user’s preferences to the movies (which range from 0 to 5) and an indicator matrix R which return 1 if there is a score on a movie for a particular user.  
      
    In the data pre-processing step, we will shortlist the 100 mostly rated movies in order to filter out the rare rated movie and we will normalize the data so the mean score for any given movie is zero.  
     
    Under the latent factors model, we will decompose the original matrix Y (m x n) into two latent factors matrices, theta (m x k) and X (n x k), where k is the number of latent factors. We assume the number of latent factors be 3 in our analysis. To make a prediction on the values of Y (Y ~ Y\_hat = Theta.dot(X.T)), we will first initialize the values in Theta and X with some standard normal numbers. The cost function is defined as the sum square error between the prediction and Y, adjusted by the indicator matrix and regularization term. We can then train our model by minimizing the cost function using an advance gradient descent algorithm called LBFGS.  
     
    Interesting result was found by investigate matrix X, which represents the tendency of movies to hidden features. Movies whose has a high score on the first, second and third hidden feature criminal related films, feminine related films and commercial/sci-fi/action films respectively. The type of movie in these 3 category are fairly distinct which indicate our model did a good job in finding useful hidden features. We can then recommend movie with highest predicted score to any given user.
3. Summary  
    We have introduced two main categories of the recommender system, which is the content-based approach and collaborative filtering approach. Under the collaborative filtering approach, it can further classify into memory-based and model-based method. Our focus in the project is to demonstrate how to use a model-based method, latent factors model, to do movie recommendation to the users. In our demonstration, our model successfully learns some latent factors. We can use it to classify and make recommendation automatically.