The dataset composes two parts, which are numerical and characteristic.

Separate the dataset as numerical and characteristic. Analyze the numerical data. There are around 50% missing value. Therefore, the primary task is to full fill the numerical missing value.

For the numerical dataset, it is sparse and non-negative. Here, we decide to use nonnegative matrix factorization to full fill the missing value.



The figure shows the sample theory of matrix factorization. For a large U by D dimension matrix. Factorize it into two small U by K matrix and D by K matrix. Make sure the result of multiplier of two matrices that could approximate equal to original matrix.



To get the prediction of a new matrix R, we can calculate the dot product of the two vectors corresponding to _i and _j:

hat{r}_{ij} = p_i^T q_j = \sum_{k=1}^k{p_{ik}q_{kj}}

Therefore, the problem changes to how to obtain mathbf{P} and mathbf{Q}. Here we discuss to give random value to  mathbf{P} and mathbf{Q} as initial value. And calculate the difference of their product to origin matrix. Minimize the difference iteratively. We use gradient descent to find the minimum of difference.

The difference is the error of two matrices. It can be calculated by following equation:

_{ij}^2 = (r_{ij} - \hat{r}_{ij})^2 = (r_{ij} - \sum_{k=1}^K{p_{ik}q_{kj}})^2

Change the values of _{ik} and _{kj} in a direction which aim to minimize the error. So, the gradient of current value should be calculated.

frac{\partial}{\partial p_{ik}}e_{ij}^2 = -2(r_{ij} - \hat{r}_{ij})(q_{kj}) = -2 e_{ij} q_{k  
 \frac{\partial}{\partial q_{ik}

Obtain the gradient, the new value of _{ik} and _{kj} is:

'_{ik} = p_{ik} + \alpha \frac{\partial}{\partial p_{ik}}e_{ij}^2 = p_{ik} + 2\alpha e_{ij} q_{kj}   
'_{kj} = q_{kj} + \alpha \frac{\partial}{\partial q_{kj}}e_{ij}^2 = q_{kj} + 2\alpha e_{ij} p_{ik} 

The coefficient, alpha is a constant, which determines the rate of approaching the minimum. Usually choose a small value for alpha, here, use 0.0002. Because if alpha chooses a large value, it may take a risk of missing the minimum and end up oscillating around the minimum.

Using the above formula, we can then iteratively perform the operation until the error converges to its minimum. We can check the overall error as calculated using the following equation and determine when we should stop the process.

 = \sum_{(u_i,d_j,r_{ij}) \in T}{e_{ij}} = \sum_{(u_i,d_j,r_{ij}) \in

To avoid overfitting, a parameter beta is added and modify the squared error as follows:

_{ij}^2 = (r_{ij} - \sum_{k=1}^K{p_{ik}q_{kj}})^2 + \frac{\beta}{2} \sum_{k=1}^K{(||P||^2 + ||Q||^2)}

The new parameter beta is used to give a good approximation of http://www.quuxlabs.com/wp-content/latex/e1e/e1e1d3d40573127e9ee0480caf1283d6-ffffff-000000-0.pngwithout having to contain large numbers. In practice, beta is set to some values in the range of 0.02. The new formula shows as follow:

'_{ik} = p_{ik} + \alpha \frac{\partial}{\partial p_{ik}}e_{ij}^2 = p_{ik} + \alpha(2 e_{ij} q_{kj} - \beta p_{i  
'_{kj} = q_{kj} + \alpha \frac{\partial}{\partial q_{kj}}e_{ij}^2 = q_{kj} + \alpha(2 e_{ij} p_{ik} - \beta q_{kj

Here, we use a small test to analyze this method:

 

The left is origin matrix and the right matrix is after NMF, the approximate matrix. For the origin matrix, 0 mean missing value. Compare exist value to approximate value. The error is small enough, therefore, consider the approximate value is missing value.

The main problem of this method is the initial value of mathbf{P} and mathbf{Q}. We give random value to mathbf{P} and mathbf{Q}. Therefore, in iterative step, the changing is based on initial random value. Compare to KNN algorithm, both methods face the same problem. The result is good or bad depends on the random value.

In practice, we choose 2 million data as test data, the size of matrix is 2 million by 112 (drop all characteristic). I cost more than 30 hours to generate the approximate matrix. Cause it needs uncountable times to determine the stop variable and other coefficient. With the limit of time and equipment, we keep this method as theory analyzing and use another simple method to full fill the missing value.