Assignment 1: Recommender Systems

Group 28

Shuang Fan s3505847

Kaiteng Jiang s3479420

Shupei Li s3430863

1 Recommender systems

RMSE and MAE are chosen as metrics in Task 1.1-1.3. The predicted value in all task is a real number that may exceed the valid range of rating. To fix this problem, we truncate the predicted value according to the function,

$$f(\hat{x}) = \begin{cases} 1, & \hat{x} < 1, \\ \hat{x}, & 1 \leq \hat{x} \leq 5, \\ 5, & \hat{x} > 5. \end{cases}$$

It is worth noting that we set the random seed to 1 in all models that involve randomness, e.g. weights initialization, five-fold division, etc. The fixed random state ensures reproducibility.

When analyzing the complexity of the algorithm, we use the following notations.

- M: The number of movies.
- U: The number of users.
- R: The number of ratings.
- K: The number of features.

In Task 1.2 and 1.3, the number of features K is an integer specified by user. We regard K as a constant in the algorithm analysis.

1.1 Naive Approaches

1.1.1 Experimental Setup

During the sampling process of an "average rating" recommender, some users or some movies might disappear from the training sets. To fix this problem, we need to define a fall-back value. In our experiments, we build models of global average, user average, movie average and a linear combination of user and movie averages (with and without the intercept). A linear regression model can be expressed as,

$$pred = \alpha \cdot avg_{user} + \beta \cdot avg_{movie} + \gamma,$$

where γ is the intercept. After simplifing the variant γ, we get,

$$pred = \alpha \cdot avg_{user} + \beta \cdot avg_{movie}.$$

Considering the matrix $\mathbf{U}$ with UserID and MovieID, the global average should be the average value of all existed Ratings. For user average rating, we calculate every user's average rating. This is the avg_{user} term in the equation. For movie average rating, we calculate every movie's average rating. This is the avg_{movie} term in the equation. When implementing the linear regression algorithm, we calculate the coefficient and consider the situations that are with and without the intercept γ.

All experiments of Task 1.1 are run on a multi-core CPU Intel(R) Core(TM) i9-9880H CPU @ 2.30GHz.

Table 1: Results of Task 1.1

Algorithm	**Train RMSE**	**Train MAE**	**Test RMSE**	**Test MAE**	**Time**
GlobalAvg	1.1171	0.9339	1.1171	0.9339	0.93s
UserAvg	1.0277	0.8227	1.0355	0.8290	1.38s
MovieAvg	0.9742	0.7783	0.9794	0.7823	1.43s
LinearReg	**0.9145**	**0.7248**	**0.9002**	**0.7122**	13m 57s
LinearRegNI	0.9465	0.7586	0.9345	0.7487	13m 45s

1.1.2 Results

Table 1 reports RMSE, MAE, and the actual run time of global average, user average, movies average and a linear combination of user and movie averages(with and without the intercept). RMSE and MAE are the mean values of five folds.

1.1.3 Algorithm Analysis

Time Complexity

In the model construction stage, computing the global average rating has the $O(R)$ time complexity, since the number of addition operations is $R-1$ and the number of division operation is 1. When calculating the average movie rating, we need $O(R)$ addition operations and $O(M)$ division operations for all movies. Therefore, the time complexity of calculating the average movie rating is $O(M+R)$. Similarly, calculating the average user rating has the $O(U+R)$ time complexity. Linear regression can be written in the matrix form as follows.

$$Y = X\beta$$

where X is an $I \times J$ matrix. Then, the result of linear regression is,

$$\hat{\beta} = \left[X^T X\right]^{-1} X^T Y$$

The time complexity of estimating β is $O(I \times J^2 + J^3)$. I equals to R in Task 1.1. J is 3 if linear regression is with intercept term, and is 2 if without intercept term. Thus, the time complexity of linear regression algorithm is $O(U+R) + O(M+R) + O(c^2R + c^3) \to O(U+M+R)$, where c is a constant. In all five models, evaluating RMSE and MAE has the time complexity of $O(R)$. The following summarizes the time complexity of five models.

- Global average rating: $O(R)$.
- Average movie rating: $O(M+R)$.
- Average user rating: $O(U+R)$.
- Linear regression (with and without intercept): $O(U+M+R)$.

Memory Complexity

Storing all ratings needs $O(R)$ memory. When calculating the global average rating, we only need to maintain one number that is the average value. It means that calculating the global average rating has $O(R) + O(1) \to O(R)$ memory complexity. As for computing the average movie rating, we firstly needs to maintain an $M \times 1$ array. Then, we calculate the mean value of the array. Therefore, the memory complexity of computing the average movie rating is $O(R+M+1) \to O(R+M)$. We can analyze the process of computing the average user rating similarly and obtain $O(R+U)$ complexity. When adopting linear regression models, we needs an additional $O(1)$ memory to store β. Besides, calculating RMSE and MAE requires $O(R)$ memory in all five models. The memory complexity of five models is listed as follows.

- Global average rating: $O(R)$.
- Average movie rating: $O(M+R)$.
- Average user rating: $O(U+R)$.
- Linear regression (with and without intercept): $O(U+M+R)$.

1.2 UV Matrix Decomposition

1.2.1 Experimental Set-up

The UV matrix decomposition is an element-wise update algorithm of the feature matrices $U_{m\times k}$ and $V_{k\times n}$, of which the multiplication UV approximates the utility matrix $M_{m\times n}$. The goal of update is to minimize the mean square error (MSE) function. For an element u_{rs} of the matrix U, the optimization problem can be written as,

$$u_{rs} = \min_x \sum_{j,m_{rj}\neq 0} \left[m_{rj} - \left(\sum_{k\neq s} u_{rk}v_{kj} + xv_{sj} \right) \right]^2 .$$

And for v_{rs},

$$v_{rs} = \min_y \sum_{i,m_{is}\neq 0} \left[m_{is} - \left(\sum_{k\neq r} u_{ik}v_{ks} + u_{ir}y \right) \right]^2 .$$

They both have closed-form solutions,

$$x = \frac{\sum\limits_{j,m_{rj}\neq 0} v_{sj} \left(m_{rj} - \sum\limits_{k\neq s} u_{rk}v_{kj} \right)}{\sum\limits_{j,m_{rj}\neq 0} v_{sj}^2},$$

$$y = \frac{\sum\limits_{i,m_{is}\neq 0} u_{ir} \left(m_{is} - \sum\limits_{k\neq s} u_{ik}v_{ks} \right)}{\sum\limits_{i,m_{is}\neq 0} u_{ir}^2}.$$

In the experiment, we visit every element of both U and V in a random order per epoch. And predicted values are clipped to the range of $[1, 5]$, as mentioned at the beginning of the section. Table 2 explains the hyperparameters in the algorithm.

Table 2: Hyperparameters in UV Matrix Factorization

Hyperparameter	**Meaning**
seeds	The random seed. Default: 1.
num_factors (K)	The number of features.
num_iter (N)	The maximum iteration.

1.2.2 Results

Table 3 reports the results of the task.

Table 3: Results of Task 1.2

K	N	**Train RMSE**	**Train MAE**	**Test RMSE**	**Test MAE**	**Time**
5	30	0.8270	0.6486	0.8979	0.7005	72m 25s

1.2.3 Algorithm Analysis

Time Complexity

Computing the sum $\sum_k$ needs time $O(K)$, and $\sum_j$ needs $O(M)$. Thus, a single update of an element of U costs $O(MK)$. Similarly, updating an element of V needs $O(UK)$. Since there are UK and KM elements in U and V respectively, an epoch which updates all the elements of U and V has a time complexity of $O(UK \cdot MK + KM \cdot UK) = O(MUK^2) \to O(MU)$. This is a very costly and slow algorithm, which explains the long running time in the experiment result. In fact, we have tried to set the dimension K to larger numbers, like 10, and it took many hours to run a single experiment, which became unacceptable.

Memory Complexity

We only need to initialize, store and update three matrices: U, V and M, thus the memory complexity is $O(UK + MK + R) \to O(U + M + R)$.

1.3 Matrix Factorization

1.3.1 Experimental Set-up

Matrix factorization algorithm in `gravity-Tikk.pdf` consists of three stages — initialization, gradient descent, and evaluation. In the experiments, we initialize feature matrices $\mathbf{U}_{I\times K}$ and $\mathbf{M}_{K\times J}$ from a Gaussian distribution $N \sim (0, 0.1)$, where I, J, and K are maximal UserID, maximal MovieID, and the number of features respectively. `gravity-Tikk.pdf` combines gradient descent and regularization strategies. Main update formulas are,

$$\begin{aligned} u_{ik}^{(t+1)} &= u_{ik}^{(t)} + \eta \cdot \left(2e_{ij} \cdot m_{kj}^{(t)} - \lambda \cdot u_{ik}^{(t)}\right) \\ m_{kj}^{(t+1)} &= m_{kj}^{(t)} + \eta \cdot \left(2e_{ij} \cdot u_{ik}^{(t)} - \lambda \cdot m_{kj}^{(t)}\right) \end{aligned}$$

where η, λ are hyperparameters, and t represents the t th iteration. To enhance the efficiency of program, we update the weights based on the rows or columns of matrices, that is,

$$\begin{aligned} U^{(t+1)}[i,:] &= U^{(t)}[i,:] + \eta\left(2e_{ij}M^{(t)}[:,j] - \lambda U^{(t)}[i,:]\right) \\ M^{(t+1)}[:,j] &= M^{(t)}[:,j] + \eta\left(2e_{ij}U^{(t)}[i,:] - \lambda M^{(t)}[:,j]\right) \end{aligned}$$

Termination condition is achieving the maximum iteration specified by user.

We try five different sets of hyperparameters to improve the model performance. Table 4 summarizes all hyperparameters in matrix factorization algorithms. All experiments of Task 1.3 are run on a multi-core CPU Intel(R) Core(TM) i7-10875H CPU @ 2.30GHz. We adopt multiprocessing programming to speed up the program.

Table 4: Hyperparameters in Matrix Factorization

Hyperparameter	**Meaning**
seeds	The random seed. Default: 1.
num_factors (K)	The number of features.
num_iter (N)	The maximum iteration.
regularization (λ)	Regularization rate.
learn_rate (η)	Learning rate.

1.3.2 Results

Table 5 reports RMSE, MAE, and the actual run time of matrix factorization algorithm on MovieLens 1M data, with different hyperparameter settings. RMSE and MAE are the mean values of five folds.

According to Table 5, the suggested setting is not the optimal choice.

1.3.3 Algorithm Analysis

Time Complexity

The time complexity of initializing U and M depends on the implementation of random number generation algorithm. For simplicity, we assume the time complexity of the initialization stage is $O(1)$. According to our Python implementation, the `for` loop to update weights has the time complexity $O(R)$. In the `for` loop, calculating error, computing gradients, and updating weights all have the time complexity $O(K) \to O(1)$. To

Table 5: Results of Task 1.3

K	N	λ	η	**Train RMSE**	**Train MAE**	**Test RMSE**	**Test MAE**	**Time**
*10	75	0.05	0.005	0.7689	0.6036	0.8686	0.6785	18m 37s
20	75	0.05	0.005	**0.7003**	**0.5475**	0.8848	0.6878	18m 38s
10	100	0.05	0.005	0.7673	0.6020	0.8670	0.6793	24m 38s
10	75	0.01	0.005	0.7627	0.5949	0.8807	0.6829	18m 29s
10	75	0.05	0.001	0.7980	0.6292	**0.8611**	**0.6763**	18m 34s

* This set of hyperparameters is suggested by Task 1.3.

evaluate the performance, we need to traverse the rating table, which leads to $O(R)$ time complexity. Calculating RMSE and MAE also has time complexity $O(R)$. Therefore, the time complexity of our implementation is $O(R)$.

Memory Complexity
We need to initialize U, M, and rating table at the beginning of the algorithm. This step requires $O(UK + MK + R) \to O(U + M + R)$ memory. In the `for` loop, storing error value and gradients needs $O(1)$ memory. During the evaluation, storing predicted values requires $O(R)$ memory and storing metrics needs $O(1)$ memory. Therefore, the memory complexity of our implementation is $O(U + M + R)$.

1.4 Comparison of Algorithms

Table 6 compares the performance of previously mentioned algorithms, which only includes the model with the best performance on test data if multiple hyperparameters have been explored.

Table 6: Algorithm Comparison

Algorithm	**Train RMSE**	**Train MAE**	**Test RMSE**	**Test MAE**	**Time**
GlobalAvg	1.1171	0.9339	1.1171	0.9339	0.93s
UserAvg	1.0277	0.8227	1.0355	0.8290	1.38s
MovieAvg	0.9742	0.7783	0.9794	0.7823	1.43s
LinearReg	0.9145	0.7248	0.9002	0.7122	13m 57s
LinearRegNI	0.9465	0.7586	0.9345	0.7487	13m 45s
UV Decomposition	0.8270	0.6486	0.8979	0.7005	72m 25s
Matrix Factorization	**0.7980**	**0.6292**	**0.8611**	**0.6763**	18m 34s

According to Table 6, Matrix Factorization algorithm outperforms all the other algorithms both on training set and test set, while its real run time is acceptable.

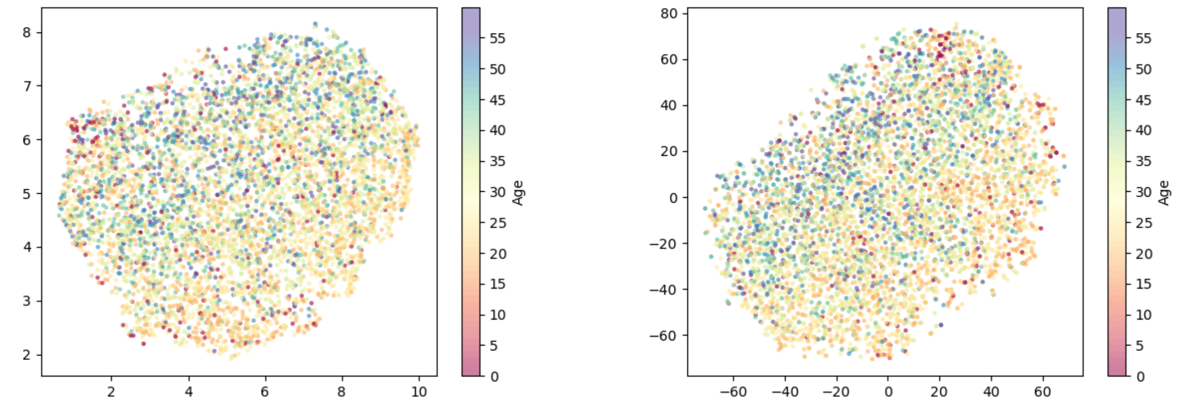
2 Data visualization

2.1 An Introduction to PCA, t-SNE and UMAP

In order to visualize the features of users and movies, we first need to utilize dimensionality reduction techniques to project these features to a 2-dimension space while keeping most of the information. Principal Component Analysis, or PCA, is the most classic one among the three techniques, which linearly transforms the data to a new coordinate system with fewer dimensions. It first normalizes the data, then finds the eigenvectors with the order of corresponding eigenvalues from large to small, so as to construct the first and remaining principal components. The other two techniques work similarly. The t-distributed stochastic neighbor embedding (t-SNE) uses a Gaussian distribution for the relationship between data points in the original space with high dimension and creates an embedding in the low-dimension space with a Student's t-distribution, optimized by gradient descent. UMAP constructs a high dimensional graph representation of the data with edge weights representing the likelihood that two points are connected, then it optimizes the low-dimension layout in a way very similar to t-SNE.

2.2 Visualization of Users' Features

We experiment on three features "Gender", "Age" and "Occupation" with all three techniques. Only the label "Age" is a little more significant in clustering. From Figure 1, it can be seen that younger and older users who are represented by warm and cool colors are distributed on different sides of the images respectively.



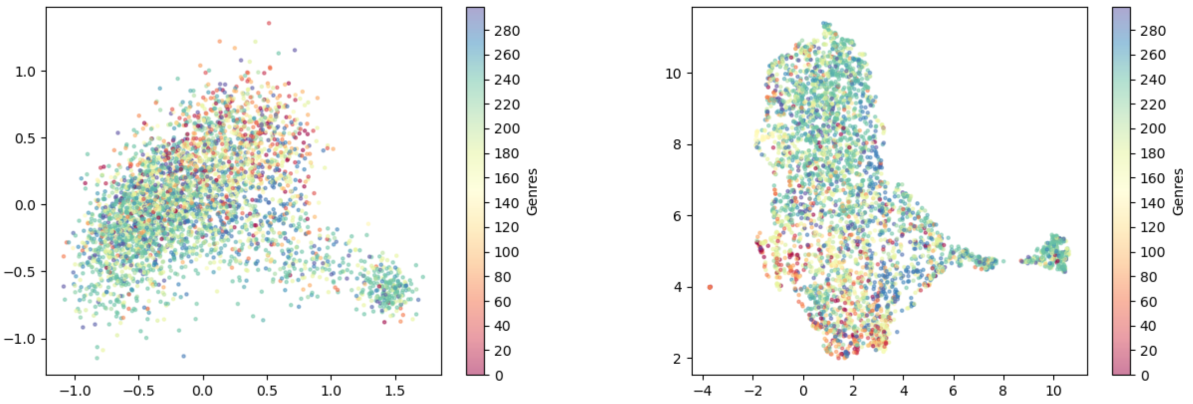
(a) UMAP with label "Age"

(b) t-SNE with label "Age"

Figure 1: Visualization of users' features

2.3 Visualization of Movies' Features

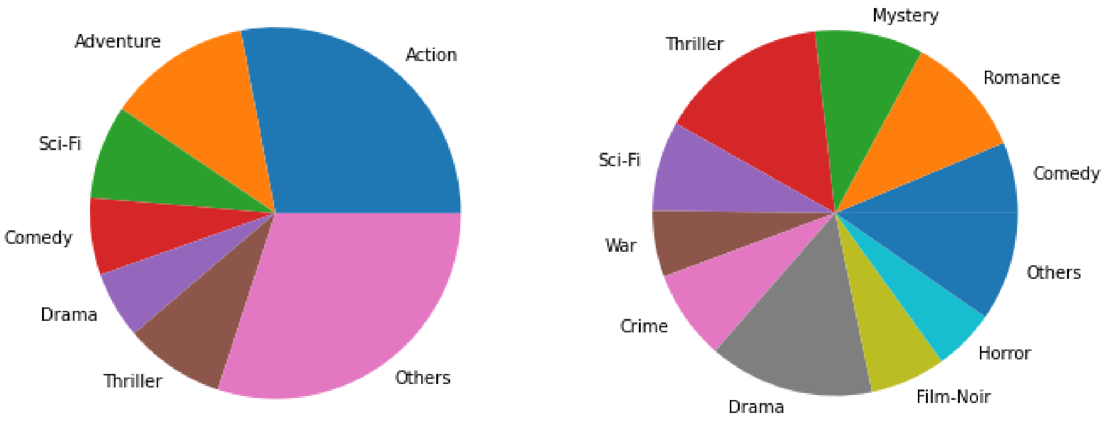
Similarly, we find out that the label "Genres" has better effect on clustering. We also count the frequencies of every single theme in "Genres". It turns out that Genres with smaller numbers in warm colors and those with larger numbers in cool colors do have different distributions of themes.



(a) PCA with label "Genres"

(b) UMAP with label "Genres"

Figure 2: Visualization of movies' features



(a) First 100 genres

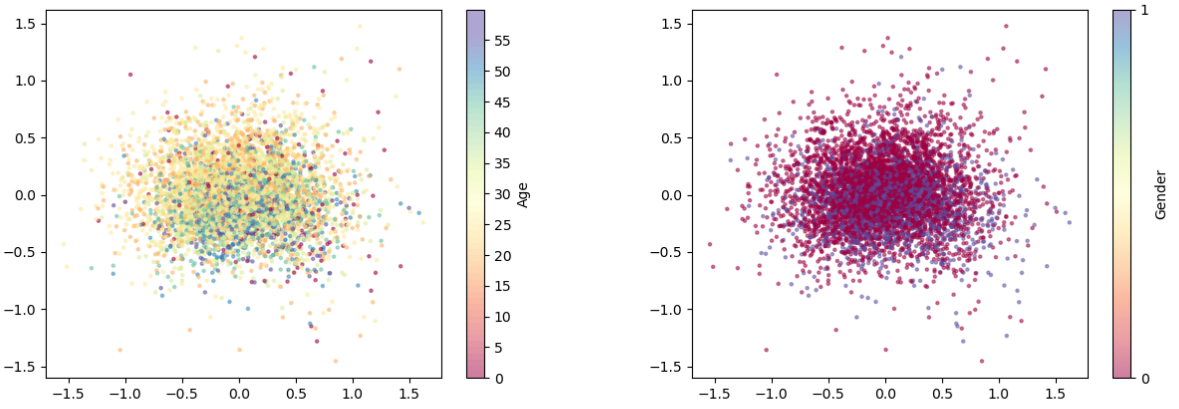
(b) Last 100 genres

Figure 3: Genres distributions

3 Contributions

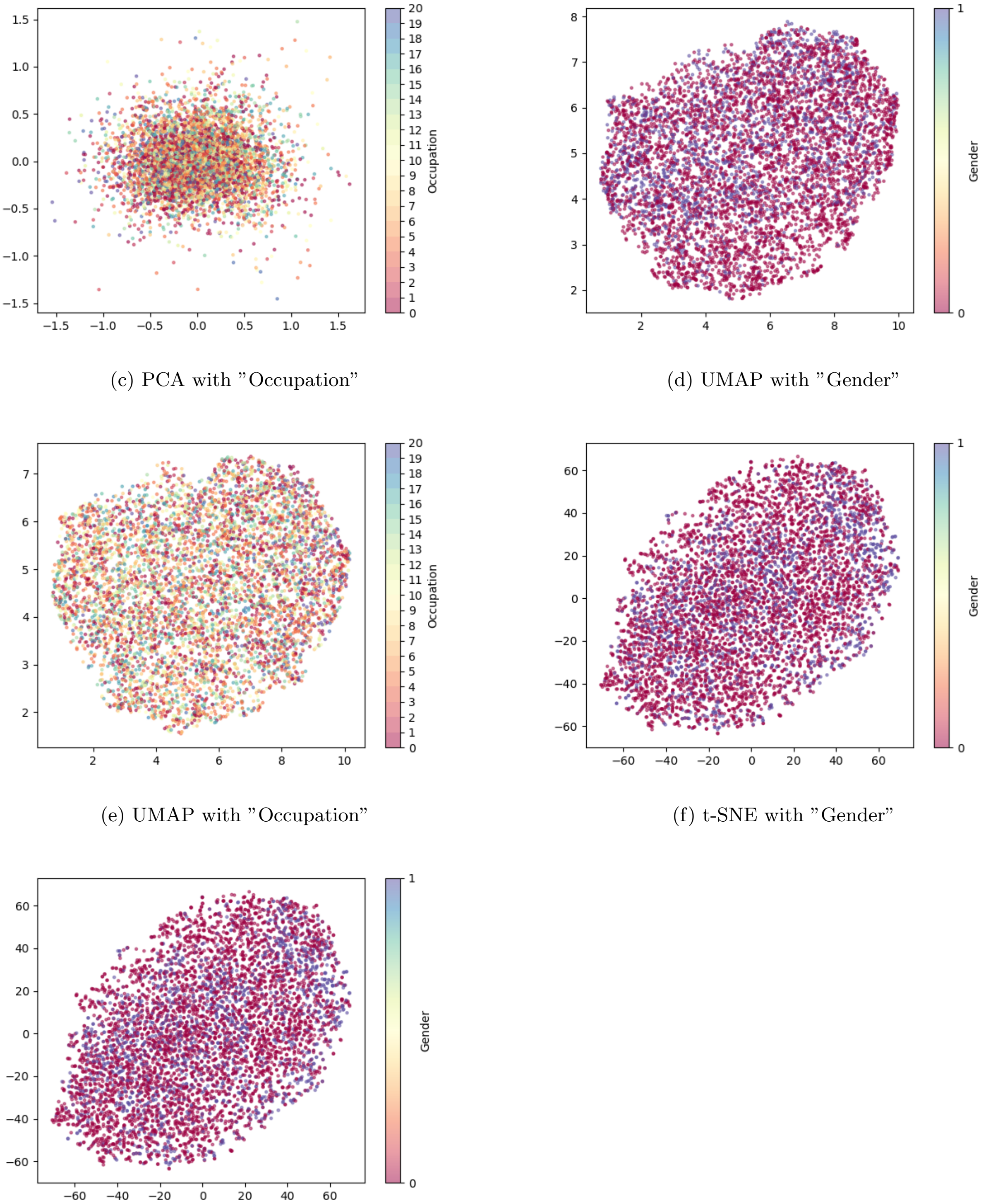
Name	Tasks
Shuang Fan	Task 1.1 code, Task 1.1 report
Kaiteng Jiang	Task 1.2 code, Task 1.2 report, Task 2 code, Task 2 report
Shupei Li	Task 1.3 code, Task 1.3 report

Appendix: Complete Results of Data Visualization



(a) PCA with ”Age”

(b) PCA with ”Gender”



(c) PCA with ”Occupation”

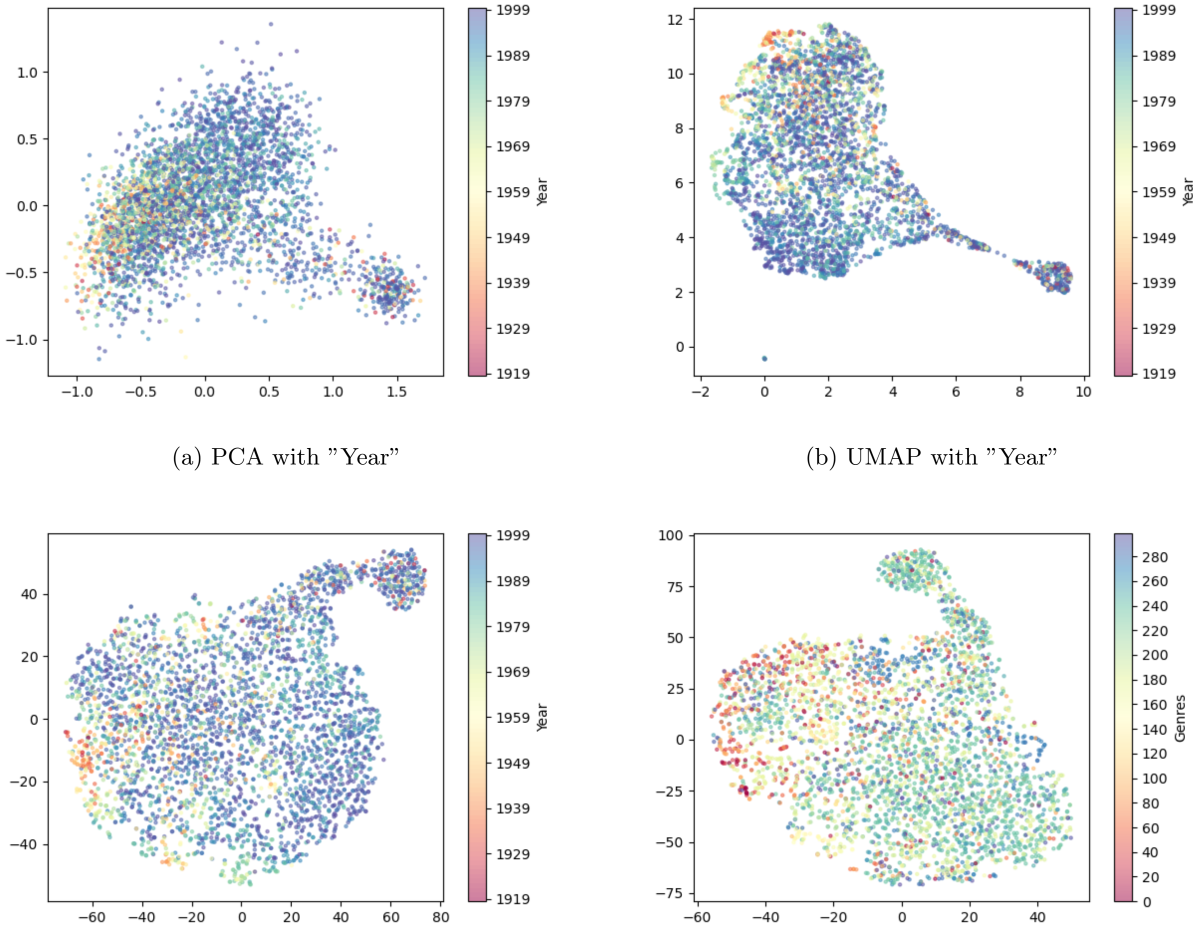
(d) UMAP with ”Gender”

(e) UMAP with ”Occupation”

(f) t-SNE with ”Gender”

(g) t-SNE with ”Gender”

Figure 5: Visualization of users’ features



(a) PCA with ”Year”

(b) UMAP with ”Year”

(c) t-SNE with ”Year”

(d) t-SNE with ”Genres”

Figure 6: Visualization of movies’ features