TEAM 14

MOVIE RECOMMENDATION SYSTEM

PROBLEM STATEMENT

To build a movie recommendation engine for users, based on existing movie ratings using only collaborative filtering approach.

DATASET

1 Million
OBSERVATIONS
[Primary dataset]

100K OBSERVATIONS
[Secondary dataset]

6000 user ratings on 4000 movies

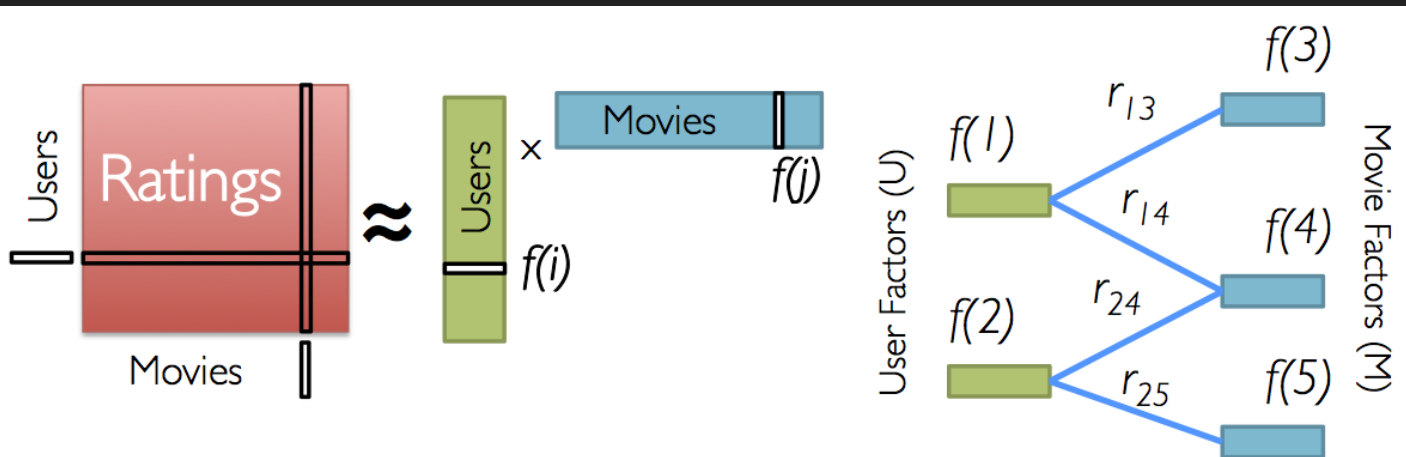| userId	movieId	rating	timestamp

MODELS

ALS

KNN

COSINE SIMILARITY

ALS, Alternating Least Square

Iterate:

$$f[i] = \arg \min_{w \in \mathbb{R}^d} \sum_{j \in \mathrm{Nbrs}(i)} \left(r_{ij} - w^T f[j]\right)^2 + \lambda ||w||_2^2$$

ALS, Alternating Least Square

30% TESTING SET

70% TRAINING SET

1M
OBSERVATION

```
Root-mean-square error = 1.144928640191194
```

Fig. RMSE value

```
+------+-------+------+----------+----------+
|userId|movieId|rating| timestamp|prediction|
+------+-------+------+----------+----------+
|   409|    471|   3.0| 967912821| 2.7708223|
|   372|    471|   3.0| 874415126| 1.8651723|
|   603|    471|   4.0| 954482443| 3.3895032|
```

Fig. Test set predictions

ALS, Alternating Least Square

CHALLENGES

1. **Cold start strategy**

- In production, for new users or items that have no rating history and on which the model has not been trained

- During cross-validation, it is actually very common to encounter users and/or items in the evaluation set that are not in the training set
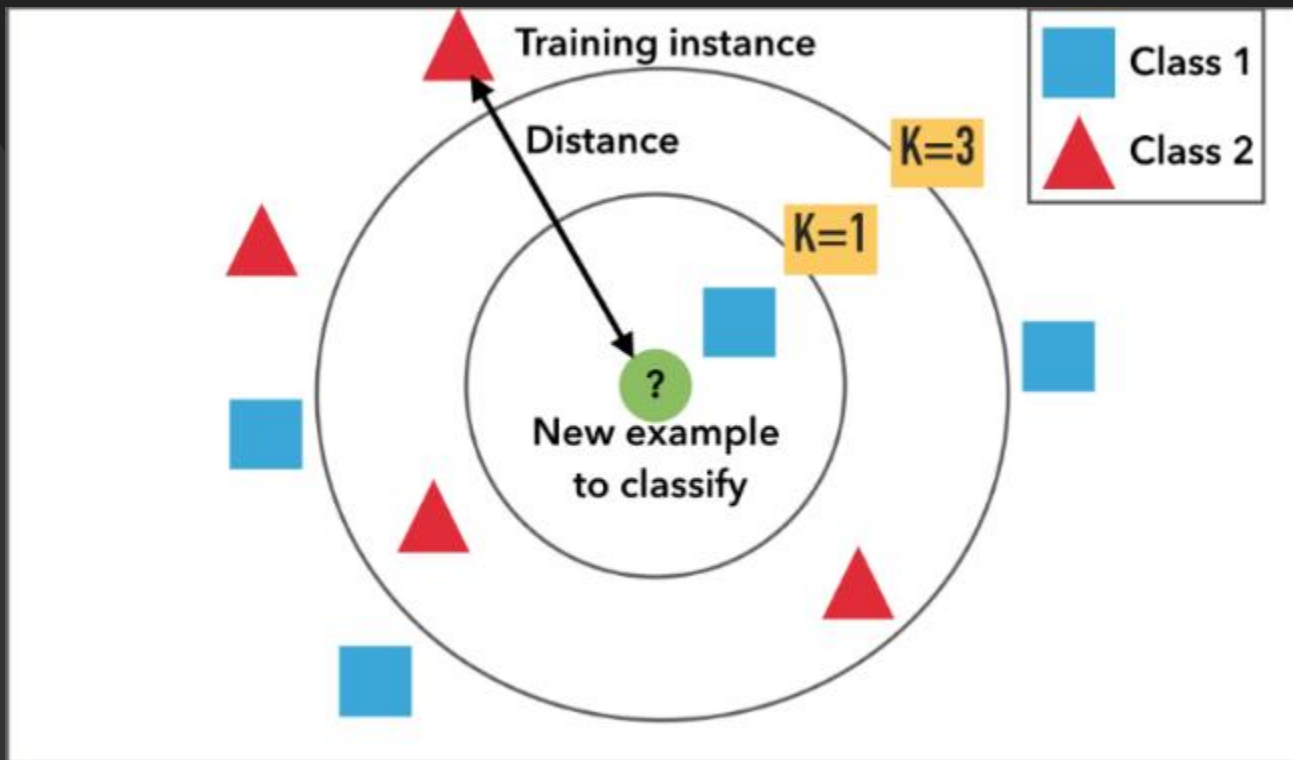
PREDICTIONS

Movies liked by user 6 :
- Sense and Sensibility
- Father of the bride – part 2
- Eye for eye

Movies recommended to user 6 :
- Police academy
- Lost Tango in Paris
- Zero dark thirty

KNN, k-nearest neighbours

COSINE SIMILARITY

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Fig. Formula to calculate distance

```
Loading movie names...

Top 10 similar movies for Silence of the Lambs, The (1991)
Shawshank Redemption, The (1994)        score: 0.9820460664993174        strength: 215
Dial M for Murder (1954)        score: 0.981581215947862         strength: 58
To Kill a Mockingbird (1962)    score: 0.9797064378902174        strength: 171
Glory (1989)    score: 0.9795842477300154        strength: 136
Casablanca (1942)       score: 0.9790976051286594        strength: 188
Sling Blade (1996)      score: 0.9780578979710847        strength: 112
12 Angry Men (1957)     score: 0.9774123548400772        strength: 97
Christmas Carol, A (1938)       score: 0.9769453088294542        strength: 59
Third Man, The (1949)   score: 0.9768277198741823        strength: 51
Rear Window (1954)      score: 0.97586974154218 strength: 160
```

Fig. Predictions for the movie

WHY THESE ALGORITHMS ?

ALS

- It is the most common and effective in matrix factorization.

KNN

- KNN did the most accurate prediction of the movies.

COSINE SIMILARITY

- This is the parameter by which we calculated the similarity score and number of users that have rated.

INSIGHTS

- Effect of ALS on size of data.

SIZE OF THE DATA	RMSE VALUE
1 Million records	0.904
100K records	1.14

INSIGHTS

- How good KNN performs !

KNN ALGORITHM PREDICTION

```
print('Recommendations for {}:'.format(fav_movie))
for i, (idx, dist) in enumerate(raw_recommends):
    print('{0}: {1}, with distance ''of {2}'.format(i+1, reverse_hashmap[idx], dist))

Recommendations for Silence of the lambs:
1: Goodfellas (1990), with distance of 0.4157697558403015
2: Sixth Sense, The (1999), with distance of 0.4073669910430908
3: American Beauty (1999), with distance of 0.39630424976348877
4: Fight Club (1999), with distance of 0.3960478901863098
5: Star Wars: Episode V - The Empire Strikes Back (1980), with distance of 0.382604956
6: Saving Private Ryan (1998), with distance of 0.38189244270324707
7: Braveheart (1995), with distance of 0.37866610288619995
8: Star Wars: Episode IV - A New Hope (1977), with distance of 0.36865437030792236
9: Terminator 2: Judgment Day (1991), with distance of 0.3682152032852173
10: Fugitive, The (1993), with distance of 0.36772972345352173
11: Matrix, The (1999), with distance of 0.3669782280921936
12: Godfather, The (1972), with distance of 0.36455029249191284
13: Usual Suspects, The (1995), with distance of 0.3641975522041321
14: Jurassic Park (1993), with distance of 0.36222052574157715
15: Schindler's List (1993), with distance of 0.35961949825286865
16: Fargo (1996), with distance of 0.34928375482559204
17: Seven (a.k.a. Se7en) (1995), with distance of 0.3445446491241455
18: Forrest Gump (1994), with distance of 0.3235422372817993
19: Shawshank Redemption, The (1994), with distance of 0.29697757959365845
20: Pulp Fiction (1994), with distance of 0.24819225072860718
```

IMDB PREDICTION

THANK YOU

SURYA USHASRI
SAURABH SANKHE
VIRAJ SAPRE
MADHUR NAGARAJ
SHARAVANKUMAR HIREGOUDAR