Movie Rating Prediction based on Netflix Prize Data

Disney
PIRATES of the CARIBBEAN
ON STRANGER TIDES
IN THEATERS 5.20.11
So bad the director and everybody involved should be shot.
★★ Okay movie not really worth the time required to watch.
★★★ An average film that's entertaining, but missable.
★★★★ A great flick that everybody should see.
★★★★★ So amazing you will crap yourself.
CAPTAIN JACK SPARROW

CSC 522 ALDA Fall 2017
Advisor: Dr. Chi Min
Group ID 01
Junhua Ma
Zhangqi Zha
Gang Zhang

NC STATE UNIVERSITY

Overview

- ❑ Motivation
- ❑ Introduction
- ❑ Dataset
- ❑ Methods
- ❑ Experiments & Results

Motivation

Netflix problem is type type of recommender system problem

- Recommender systems
  - A subclass of information filtering system
  - To predict the "rating" or "preference" that a user would give to an item.

- Recommender systems have become increasingly popular in recent years.
  - Movies
  - Music
  - News

amazon

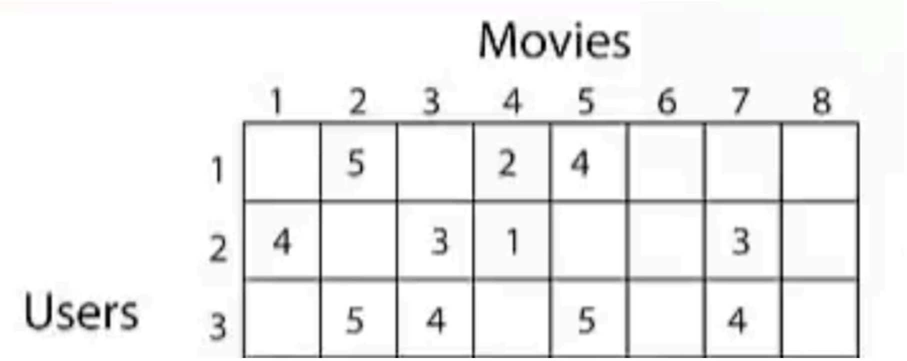
Spotify

facebook

Introduction – Netflix Prize

- Netflix Prize - a contest Netflix sponsored.
- In 2009, the award was grant to team “BellKor’s Pragmatic Chaos” with the improvement of 10% on RMSE.

Dataset Overview

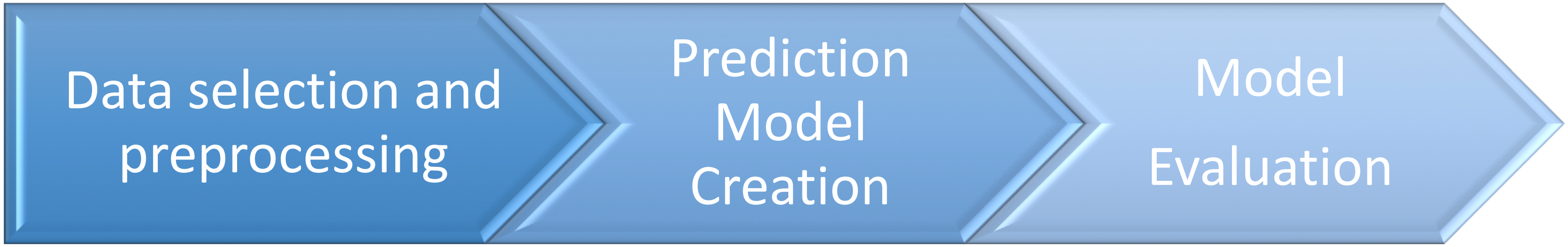


Movies

Users	1	2	3	4	5	6	7	8
1		5		2	4			
2	4		3	1			3	
3		5	4		5		4	

- 100 Million data points “user X rated movie Y a 4.0 on 2/12/05”
- 4 attributes: user ID, movie ID, movie title and date
- 5 classes: rating 1, 2, 3, 4, and 5
- 480,189 users rate the 17,770 movies
- Custom data: crawled online movie information: genre, director, actor

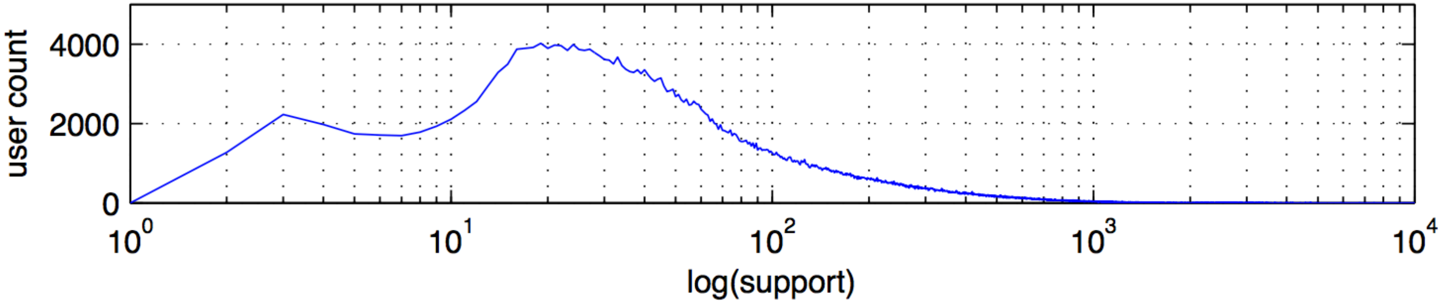
Methods - Preprocessing



Data selection and preprocessing

Prediction Model Creation

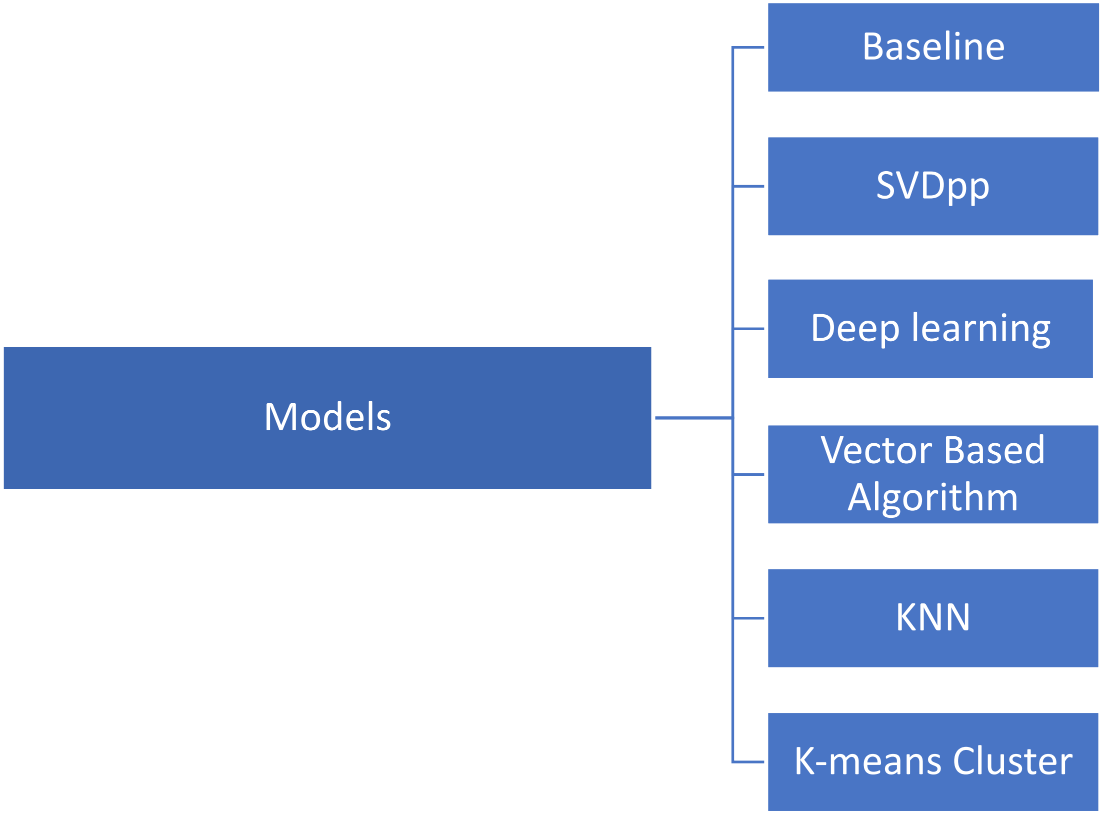
Model Evaluation



Data Prepressing

- Each User does not rate much on all movies (data sparsity), can not do random sampling on data.
- Stratify sampling: pick fewer movies and fewer user.
- Data transform to useable format.

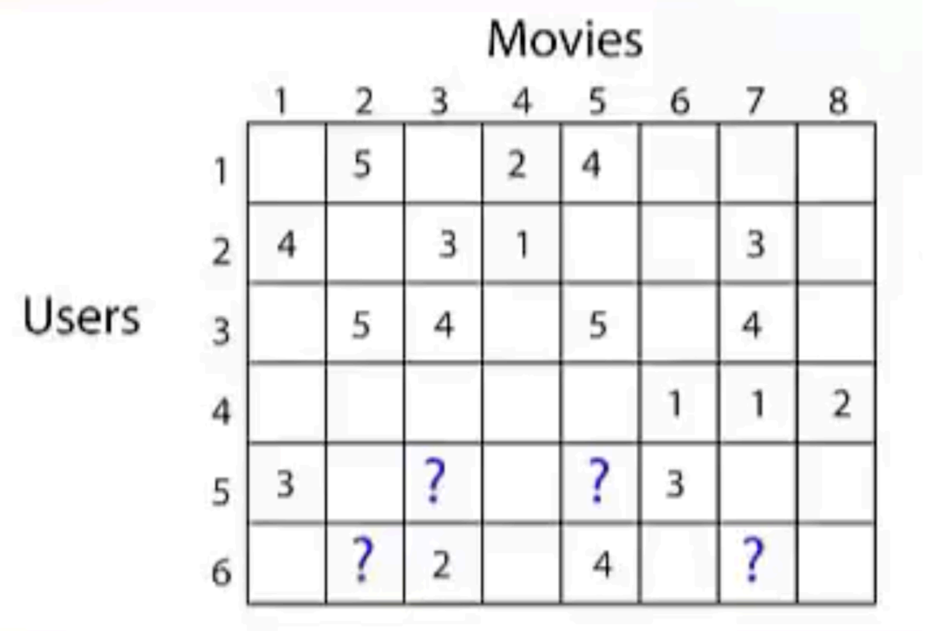
Methods - Models



Models - Baseline

Baseline predictor:
mean rate with bias on specific movie and specific user

$$r_{ui} = \mu + b_i + b_\mu$$



		Movies							
		1	2	3	4	5	6	7	8
Users	1		5		2	4			
	2	4		3	1			3	
	3		5	4		5		4	
	4						1	1	2
	5	3		?		?	3		
	6		?	2		4		?	

Models - SVDpp

SVDpp predictor:

- SVD approximates matrix A by:

$$A = U\Sigma V^*$$

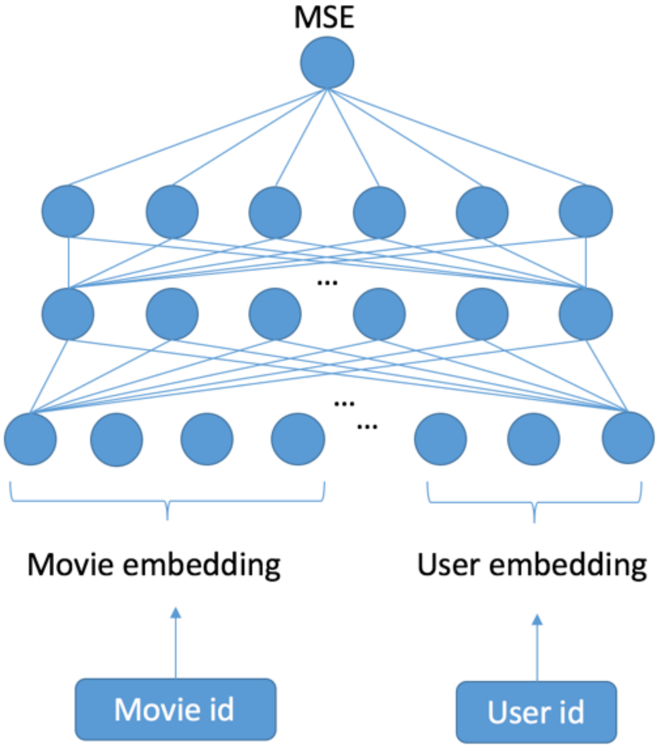
- In recommender system, user-movie interactions are modeled as inner products in the latent factor space.

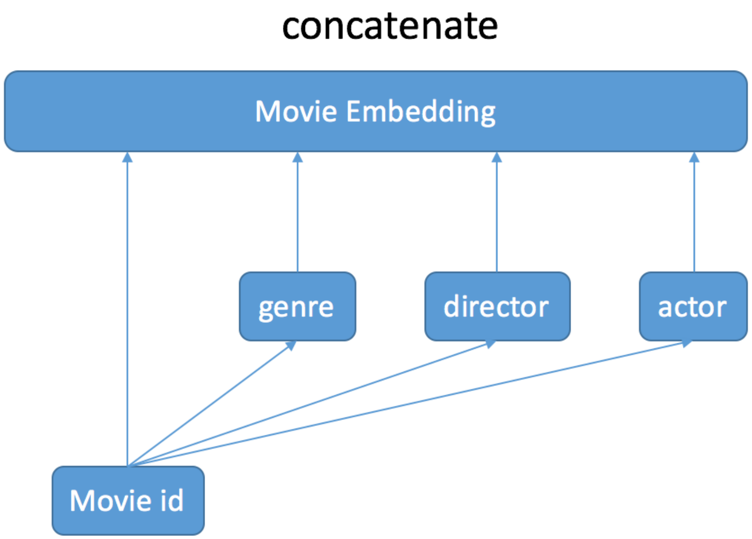
- Movie is associated with a vector q_i
- User is associated with a vector p_μ
- Inner product of $q_i^* p_u$ approximates the rating
- Adding the basel rediction would be :

$$r_{ui} = \mu + b_i + b_\mu + q_i^* p_u$$

Models - Deep Learning

Deep Learning predictor:

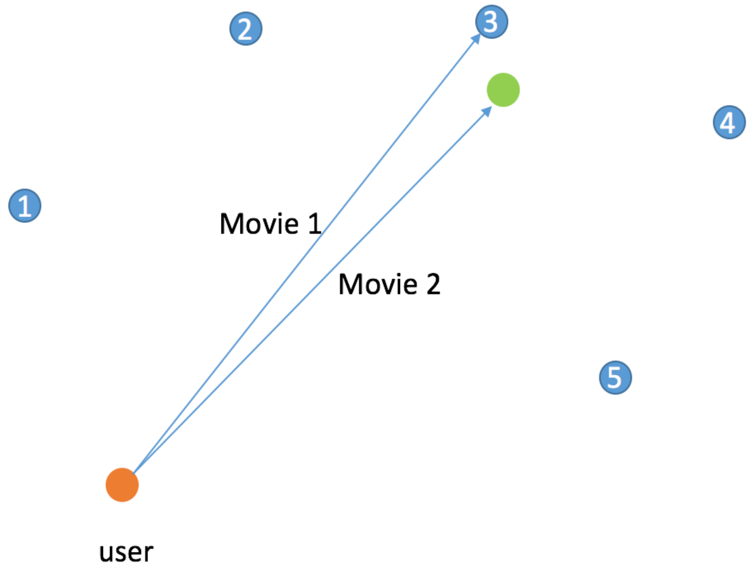




Models - Vector Based Algorithm

Vector Based Algorithm Predictor:

- Inspired by transE
- Data we have is triplets as (movie, user, rate)
- We train embedding for each movies, users and rates(1 to 5), and want to minimize the distance between movie + user and rate

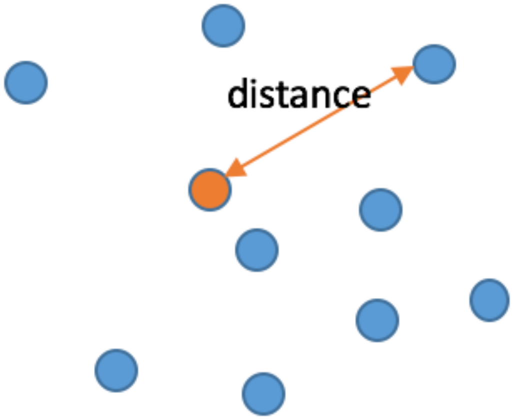


$$\sum_{(m,u,r)\in S\,(m',u',r')\in S'} [\alpha + d(m+u,r) - d(m+u,r)]_+$$

Models – KNN and Other Methods

K-Nearest-Neighbors predictor:

- User based kNN and movie based kNN



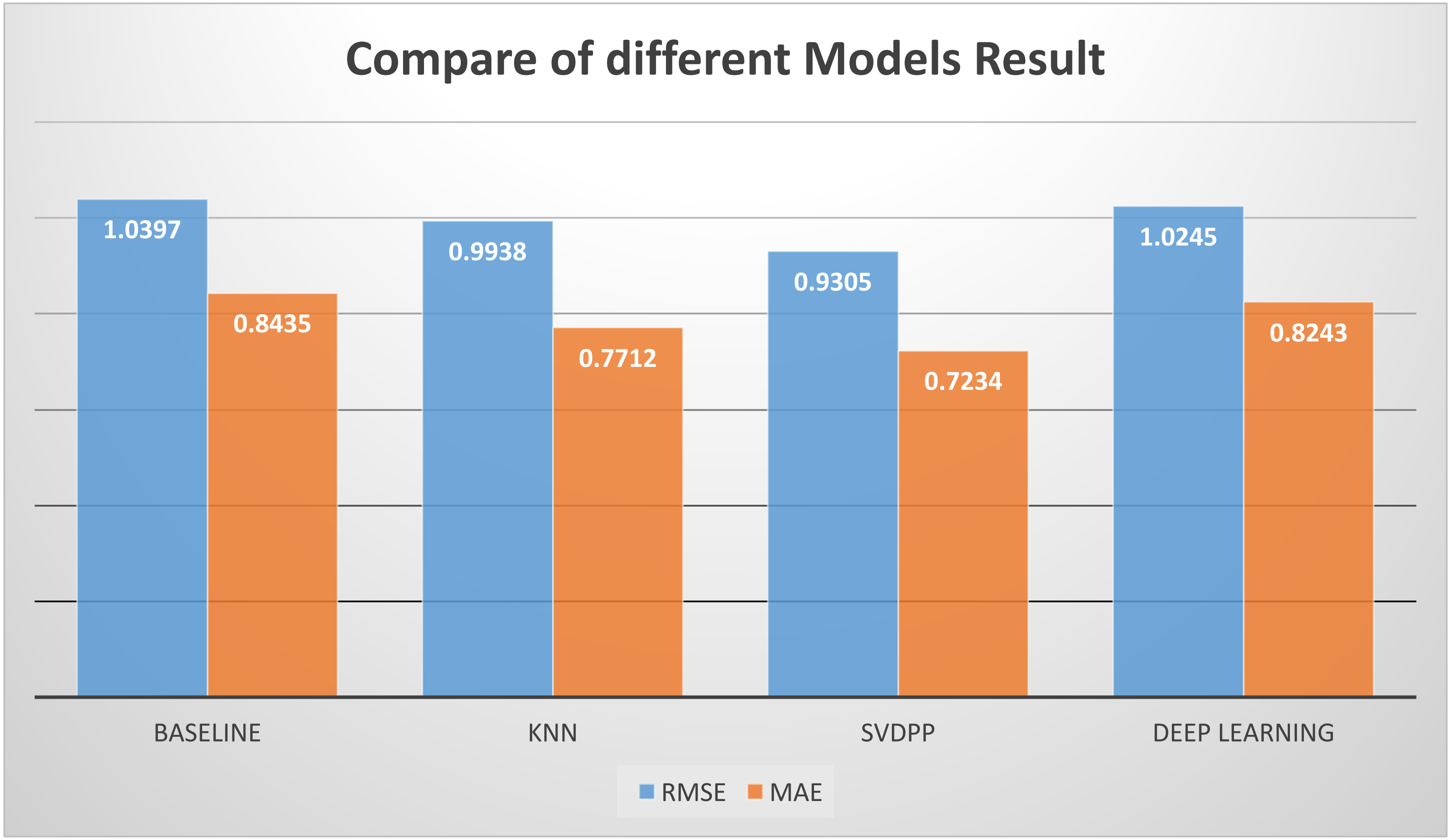
How to measure distance?

Calculate mean and variance for each user's rate
then use KL divergence as distance or just use correlation coefficient

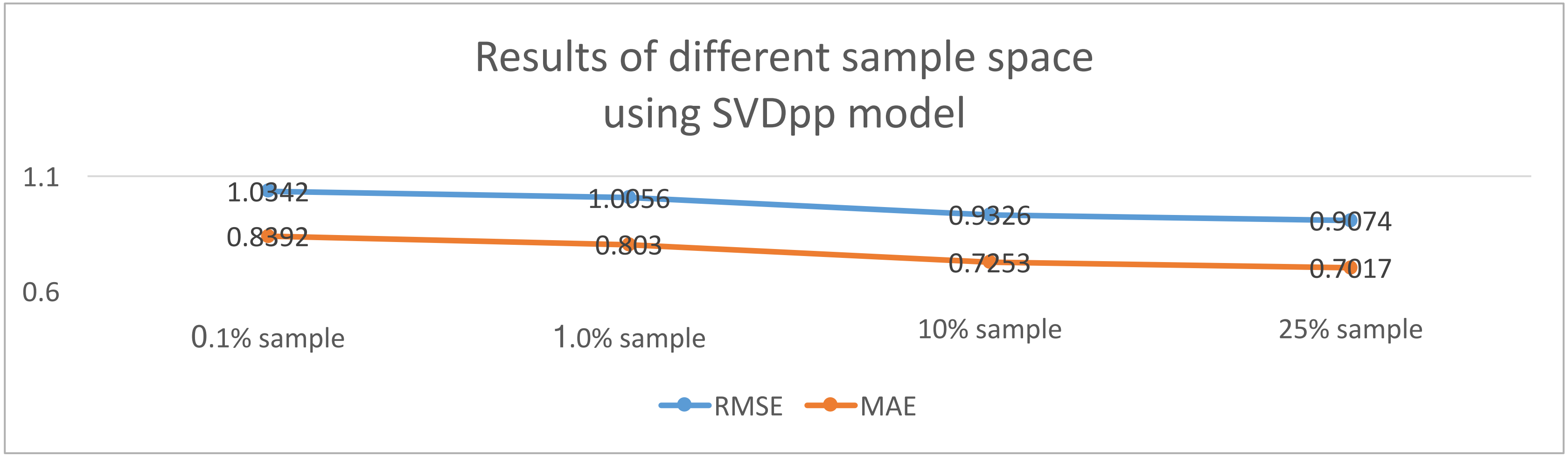
ANN based Clustering and Visualization

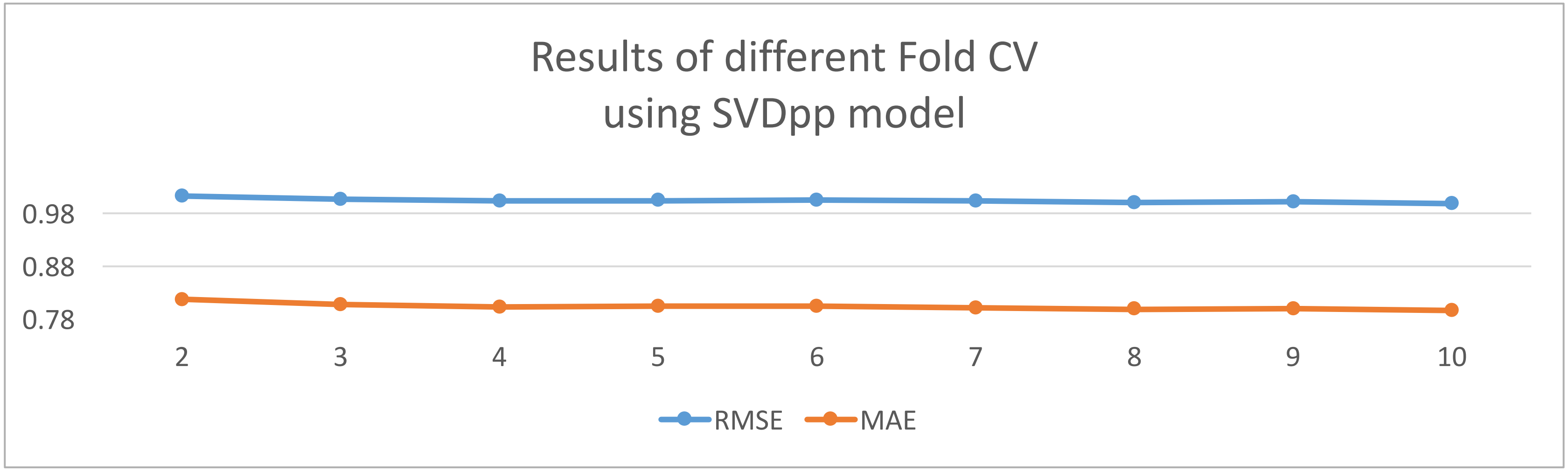
- Use movie and user embedding trained by ANN for k-means clustering
- Use t-SNE for visualization result in 2D space.

Experiments & Results



Experiments & Results



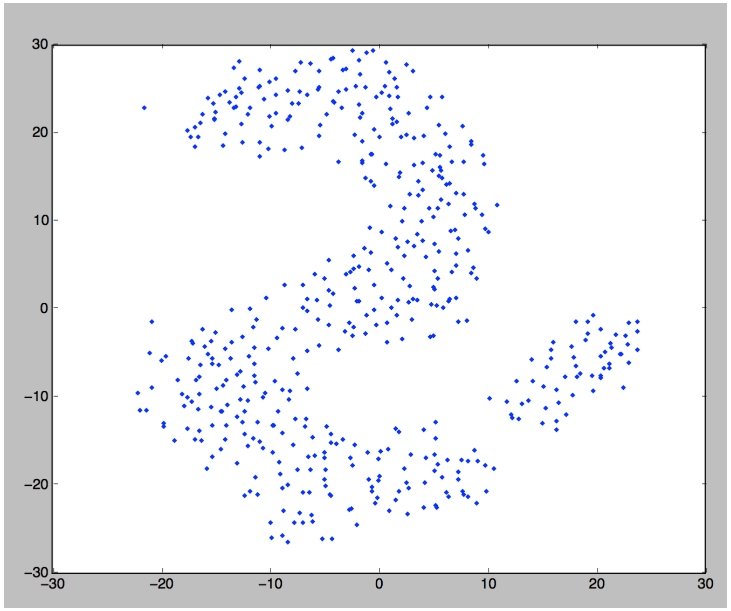


NC STATE UNIVERSITY

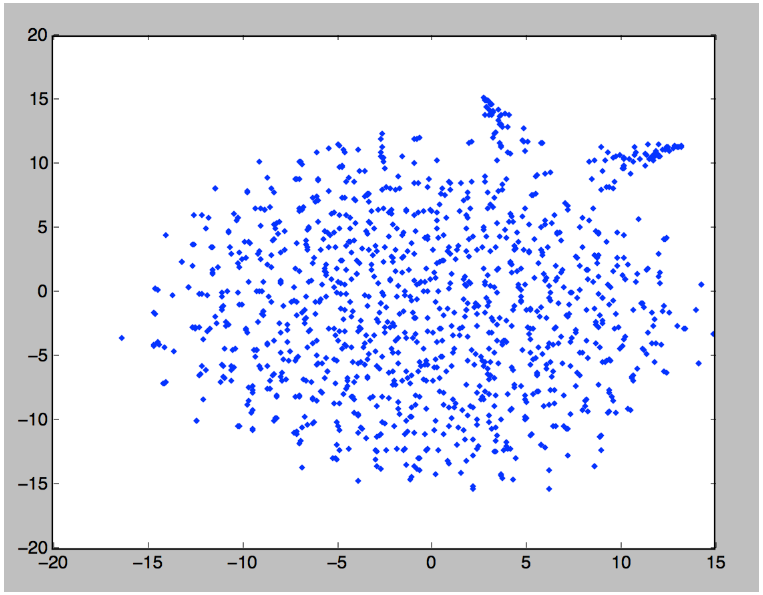
Experiments & Results

Visualized trained movies (500) and users (200,000) embedding using t-SNE

movies



users



NC STATE UNIVERSITY