A Lower-Bound on the Number of Rankings Required in Recommender Systems Using Collaborative Filtering

Peter Marbach
University of Toronto
ENS/INRIA Paris, France
Overview

- Motivation
- Problem Formulation
- Results
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *,**,***,...,*****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings

Recommender Systems Using Collaborative Filtering
Motivation

Problem Formulation

Results

Netflix Prize

Netflix:
- Online Video Rental
- Users Rank Movies: *, **, ..., *****
- System Provides Recommendations (Ranking Predictions)

Netflix Prize
- Improve Netflix Prediction System by 10%
- Prize: 1 Million Dollar
- Data Set of User Rankings
  - Training Set
  - Data Set

Training Set
- 480,000 Users
- 30,000 Movies
- 100,000,000 Rankings
Netflix Prize

- **Netflix:**
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- **Netflix Prize**
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- **Training Set**
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- **Netflix:**
  - Online Video Rental
  - Users Rank Movies: *,**,...,*****
  - System Provides Recommendations (Ranking Predictions)

- **Netflix Prize**
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- **Training Set**
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *,**,...,*****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- **Netflix:**
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- **Netflix Prize**
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- **Training Set**
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *,**,....,*****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings

Recommender Systems Using Collaborative Filtering
Netflix Prize

Netflix:
- Online Video Rental
- Users Rank Movies: *, **, ..., *****
- System Provides Recommendations (Ranking Predictions)

Netflix Prize
- Improve Netflix Prediction System by 10%
- Prize: 1 Million Dollar
- Data Set of User Rankings
  - Training Set
  - Data Set

Training Set
- 480,000 Users
- 30,000 Movies
- 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *,**,...,*****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- **Netflix:**
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- **Netflix Prize**
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- **Training Set**
  - 480,000 Users
    - 30,000 Movies
    - 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Netflix Prize

- Netflix:
  - Online Video Rental
  - Users Rank Movies: *, **, ..., *****
  - System Provides Recommendations (Ranking Predictions)

- Netflix Prize
  - Improve Netflix Prediction System by 10%
  - Prize: 1 Million Dollar
  - Data Set of User Rankings
    - Training Set
    - Data Set

- Training Set
  - 480,000 Users
  - 30,000 Movies
  - 100,000,000 Rankings
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...
- Important Problem: Collaborative Filtering
- Large Data Set
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...  
- Important Problem: Collaborative Filtering
- Large Data Set
Why Interesting?

1 Million Dollar

Interesting Questions

- Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...

Important Problem: Collaborative Filtering

Large Data Set
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...
- Important Problem: Collaborative Filtering
- Large Data Set
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...

- Important Problem: Collaborative Filtering
- Large Data Set
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - …
- Important Problem: Collaborative Filtering
- Large Data Set
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...
- Important Problem: Collaborative Filtering
- Large Data Set
Why Interesting?

- 1 Million Dollar
- Interesting Questions
  - Is it possible to improve by 10%?
  - Is it a difficult problem?
  - How many rankings are needed to make “good” predictions
  - ...
- Important Problem: Collaborative Filtering
- Large Data Set
Goal

- Interesting Problems/Models
- Possible to Find Answers
- Many Interesting/Important Open Problems
Goal

- Interesting Problems/Models
- Possible to Find Answers
- Many Interesting/Important Open Problems
Goal

- Interesting Problems/Models
- Possible to Find Answers
- Many Interesting/Important Open Problems
Model

- $N$ users
- $I_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$
- Correlation
  - $C$ classes, $c = 1, \ldots, C$
  - Ranking vector: $r_c = (r_c(1), \ldots, r_c(I_N))$
- Ranking vectors can “overlap”
Model

- $N$ users
- $I_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$
- Correlation
  - $C$ classes, $c = 1, \ldots, C$
  - Ranking vector: $r_c = (r_c(1), \ldots, r_c(I_N))$
- Ranking vectors can “overlap”
Model

- \( N \) users
- \( l_N \) items to be ranked
  - Each user ranks \( m_N \) items (chosen at random)
  - Ranking: \([0, 1]\)
- Correlation
  - \( C \) classes, \( c = 1, \ldots, C \)
  - Ranking vector: \( r_c = (r_c(1), \ldots, r_c(l_N)) \)
- Ranking vectors can “overlap”
Model

- $N$ users
- $I_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$ 
- Correlation
  - $C$ classes, $c = 1, ..., C$
  - Ranking vector: $r_c = (r_c(1), ..., r_c(I_N))$
- Ranking vectors can “overlap”
Model

- $N$ users
- $I_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$
- Correlation
  - $C$ classes, $c = 1, ..., C$
  - Ranking vector: $r_c = (r_c(1), ..., r_c(I_N))$
- Ranking vectors can “overlap”
Model

- $N$ users
- $l_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$
- Correlation
  - $C$ classes, $c = 1, \ldots, C$
  - Ranking vector: $r_c = (r_c(1), \ldots, r_c(l_N))$
- Ranking vectors can “overlap”
**Model**

- $N$ users
- $I_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$
- Correlation
  - $C$ classes, $c = 1, \ldots, C$
  - Ranking vector: $r_c = (r_c(1), \ldots, r_c(I_N))$
- Ranking vectors can “overlap”
Model

- $N$ users
- $I_N$ items to be ranked
- Each user ranks $m_N$ items (chosen at random)
- Ranking: $[0, 1]$
- Correlation
  - $C$ classes, $c = 1, ..., C$
  - Ranking vector: $r_c = (r_c(1), ..., r_c(I_N))$
- Ranking vectors can “overlap”
Model

- \( N \) users
- \( l_N \) items to be ranked
- Each user ranks \( m_N \) items (chosen at random)
- Ranking: \([0, 1]\)
- Correlation
  - \( C \) classes, \( c = 1, \ldots, C \)
  - Ranking vector: \( r_c = (r_c(1), \ldots, r_c(l_N)) \)
- Ranking vectors can “overlap”
“What is the minimal number of $m_N$ of rankings (as a function of $N$ and $l_N$) required in order to correctly associate all users with their corresponding class, in the limit as $N$ approaches infinity?”

- Trivial $m_N = l_N$
- Impossible if $m_N = 1$
- Threshold?
- Lower-bound on $m_N$
“What is the minimal number of $m_N$ of rankings (as a function of $N$ and $I_N$) required in order to correctly associate all users with their corresponding class, in the limit as $N$ approaches infinity?”

- Trivial $m_N = I_N$
- Impossible if $m_N = 1$
- Threshold?
- Lower-bound on $m_N$
“What is the minimal number of $m_N$ of rankings (as a function of $N$ and $I_N$) required in order to correctly associate all users with their corresponding class, in the limit as $N$ approaches infinity?”

- Trivial $m_N = I_N$
- Impossible if $m_N = 1$
- Threshold?
- Lower-bound on $m_N$
Question

“What is the minimal number of $m_N$ of rankings (as a function of $N$ and $l_N$) required in order to correctly associate all users with their corresponding class, in the limit as $N$ approaches infinity?”

- Trivial $m_N = l_N$
- Impossible if $m_N = 1$
  - Threshold?
  - Lower-bound on $m_N$
Question

“What is the minimal number of $m_N$ of rankings (as a function of $N$ and $I_N$) required in order to correctly associate all users with their corresponding class, in the limit as $N$ approaches infinity?”

- Trivial $m_N = I_N$
- Impossible if $m_N = 1$
- Threshold?
- Lower-bound on $m_N$
“What is the minimal number of $m_N$ of rankings (as a function of $N$ and $l_N$) required in order to correctly associate all users with their corresponding class, in the limit as $N$ approaches infinity?”

- Trivial $m_N = l_N$
- Impossible if $m_N = 1$
- Threshold?
- Lower-bound on $m_N$
Random Graph Model

Users  Items

. . .  . . .

. . .  . . .
Random Graph Model

Users

Items
Motivation
Problem Formulation
Results

Random Graph Model

“Edge does not mean same class!”
Random Graph Model

“Edge does not mean same class!”
Lower-Bound on $m_N$

- Complete Separation: $r_{c'}(i) \neq r_{c''}(i)$, $i = 1, \ldots, I_N$
- “Edge does mean same class!”
- “Correct Classification” means “Full Connectivity”
- Is “complete separation” too strong an assumption?
Complete Separation: $r_{c'}(i) \neq r_{c''}(i), i = 1, \ldots, l_N$

“Edge does mean same class!”

“Correct Classification” means “Full Connectivity”

Is “complete separation” too strong an assumption?
Lower-Bound on $m_N$

- Complete Separation: $r_{c'}(i) \neq r_{c''}(i), i = 1, \ldots, I_N$
- “Edge does mean same class!”
- “Correct Classification” means “Full Connectivity”
- Is “complete separation” too strong an assumption?
Lower-Bound on $m_N$

- Complete Separation: $r_{c'}(i) \neq r_{c''}(i), i = 1, \ldots, I_N$
- “Edge does mean same class!”
- “Correct Classification” means “Full Connectivity”
- Is “complete separation” too strong an assumption?
Lower-Bound on $m_N$

- Complete Separation: $r_{c'}(i) \neq r_{c''}(i), i = 1, \ldots, I_N$
- “Edge does mean same class!”
- “Correct Classification” means “Full Connectivity”
- Is “complete separation” too strong an assumption?
Lower-Bound on $m_N$

- **Notation**
  - Fixed class $c$
  - $N$
  - Probability that an edge exists between two users

$$p(l_N, m_N) \approx \frac{m_N^2}{l_N} = \frac{1}{l_N} \cdot m_N \cdot m_N$$

- Not a Erdos-Renyi graph
Lower-Bound on $m_N$

- **Notation**
  - Fixed class $c$
  - $N$

  Probability that an edge exists between two users

  $$p(l_N, m_N) \approx \frac{m_N^2}{l_N} = \frac{1}{l_N} \cdot m_N \cdot m_N$$

- Not a Erdos-Renyi graph
Lower-Bound on $m_N$

- **Notation**
  - Fixed class $c$
  - $N$

- Probability that an edge exists between two users

$$p(I_N, m_N) \approx \frac{m_N^2}{I_N} = \frac{1}{I_N} \cdot m_N \cdot m_N$$

- Not a Erdos-Renyi graph
Lower-Bound on $m_N$

- **Notation**
  - Fixed class $c$
  - $N$

- Probability that an edge exists between two users

\[
p(l_N, m_N) \approx \frac{m_N^2}{l_N} = \frac{1}{l_N} \cdot m_N \cdot m_N
\]

- Not a Erdos-Renyi graph
Lower-Bound on $m_N$

- Notation
  - Fixed class $c$
  - $N$

- Probability that an edge exists between two users

\[
p(I_N, m_N) \approx \frac{m_N^2}{l_N} = \frac{1}{l_N} \cdot m_N \cdot m_N
\]

- Not a Erdos-Renyi graph
Random Graph Model
Results

- $P_N$: Probability of Full Connectivity

Note

$$\frac{Nm^2_N}{l_N} = N \frac{m^2_N}{l_N} \approx Np(I_N, m_N)$$

If

$$\frac{Nm^2_N}{l_N} = \omega(\log N)$$

then $\lim_{N \to \infty} P_N = 1$,

if

$$\frac{Nm^2_N}{l_N} = \log N + a + o(1)$$

then $\lim_{N \to \infty} P_N \leq e^{-e^{-a}}$. 

Recommender Systems Using Collaborative Filtering
Results

- $P_N$: Probability of Full Connectivity

$$\frac{Nm_N^2}{I_N}$$

- Note

$$\frac{Nm_N^2}{I_N} = N\frac{m_N^2}{I_N} \approx Np(I_N, m_N)$$

- If

$$\frac{Nm_N^2}{I_N} = \omega(\log N)$$

then $\lim_{N \to \infty} P_N = 1$,

- if

$$\frac{Nm_N^2}{I_N} = \log N + a + o(1)$$
$P_N$: Probability of Full Connectivity

Note

$$\frac{Nm^2}{I_N} = N\frac{m^2}{I_N} \approx Np(I_N, m_N)$$

If

$$\frac{Nm^2}{I_N} = \omega(\log N)$$

then $\lim_{N \to \infty} P_N = 1$,

if

$$\frac{Nm^2}{I_N} = \log N + a + o(1)$$

then $\lim_{N \to \infty} P_N \leq e^{-e^{-a}}$. 
P_N: Probability of Full Connectivity

Note

\[ \frac{Nm^2_N}{I_N} = N \frac{m^2_N}{I_N} \approx Np(I_N, m_N) \]

If

\[ \frac{Nm^2_N}{I_N} = \omega(\log N) \]

then \( \lim_{N \to \infty} P_N = 1 \),

if

\[ \frac{Nm^2_N}{I_N} = \log N + a + o(1) \]

then \( \lim_{N \to \infty} P_N \leq e^{-e^{-a}} \).
Motivation

Problem Formulation

Results

$P_N$: Probability of Full Connectivity

Note

$$\frac{Nm^2_N}{I_N} = N\frac{m^2_N}{I_N} \approx Np(I_N, m_N)$$

If

$$\frac{Nm^2_N}{I_N} = \omega(\log N)$$

then $\lim_{N \to \infty} P_N = 1$,

if

$$\frac{Nm^2_N}{I_N} = \log N + a + o(1)$$

then $\lim_{N \to \infty} P_N \leq e^{-e^{-a}}$. 

Recommender Systems Using Collaborativ Filtering
Analysis

- Many-User Case

\[ \lim_{N \to \infty} \frac{N}{l_N \log N} = \infty \]

- Balanced Case

\[ \lim_{N \to \infty} \frac{N}{l_N} = b \]

- Many-Item Case

\[ \lim_{N \to \infty} \frac{Nm_N}{l_N} = 0 \]
Many-User Case

\[ \lim_{N \to \infty} \frac{N}{I_N \log N} = \infty \]

Balanced Case

\[ \lim_{N \to \infty} \frac{N}{I_N} = b \]

Many-Item Case

\[ \lim_{N \to \infty} \frac{Nm_N}{I_N} = 0 \]
Analysis

- Many-User Case
  \[\lim_{N \to \infty} \frac{N}{I_N \log N} = \infty\]

- Balanced Case
  \[\lim_{N \to \infty} \frac{N}{I_N} = b\]

- Many-Item Case
  \[\lim_{N \to \infty} \frac{Nm_N}{I_N} = 0\]
Lower-Bound

\[
\frac{Nm^2_N}{I_N} \approx \log N
\]

Netflix
- \( N = 480,000 \)
- \( I = 30,000 \)
- \( m \approx 200 \)

For Netflix

\[
\frac{Nm^2}{I} \approx 1.3N
\]
Back to Netflix

- Lower-Bound
  \[ \frac{Nm^2_N}{I_N} \approx \log N \]

- Netflix
  - \( N = 480,000 \)
  - \( I = 30,000 \)
  - \( m \approx 200 \)

  For Netflix
  \[ \frac{Nm^2}{I} \approx 1.3N \]
Lower-Bound

\[
\frac{Nm^2}{l^N} \approx \log N
\]

Netflix
- \( N = 480,000 \)
- \( l = 30,000 \)
- \( m \approx 200 \)

For Netflix

\[
\frac{Nm^2}{l} \approx 1.3N
\]
**Conclusions**

- Collaborative Filtering
- Random Graph Model
- Lower-Bound
- Algorithm?
- Classify correctly a large fraction of the users
- ...

Recommender Systems Using Collaborative Filtering
Conclusions

- Collaborative Filtering
- Random Graph Model
- Lower-Bound
- Algorithm?
- Classify correctly a large fraction of the users
- ...

Recommender Systems Using Collaborative Filtering
Conclusions

- Collaborative Filtering
- Random Graph Model
- Lower-Bound
  - Algorithm?
  - Classify correctly a large fraction of the users
  - ...

Recommender Systems Using Collaborative Filtering
Conclusions

- Collaborative Filtering
- Random Graph Model
- Lower-Bound
- Algorithm?

Classify correctly a large fraction of the users

...
Conclusions

- Collaborative Filtering
- Random Graph Model
- Lower-Bound
- Algorithm?
- Classify correctly a large fraction of the users

...
Conclusions

- Collaborative Filtering
- Random Graph Model
- Lower-Bound
- Algorithm?
- Classify correctly a large fraction of the users
- ...
Thank You!