Tell me Why? Tell me More!
Explaining Predictions, Iterated Learning Bias, and Counter-Polarization in Big Data Discovery Models

*CCS@Lexington, October 16, 2017*

**Olfa Nasraoui**

This work is a Collaboration with:
Behnoush Abdollahi, Mahsa Badami, Sami Khenissi, Wenlong Sun, Gopi Nutakki, Pegah Sagheb: @UofL
& Patrick Shafto: @Rutgers-Newark

Knowledge Discovery & Web Mining Lab
Computer Engineering & Computer Science Dept.
University of Louisville

http://webmining.spd.louisville.edu/
olfa.nasraoui@louisville.edu

Acknowledgements:
National Science Foundation:
NSF INSPIRE (IIS)- Grant #1549981
NSF IIS - Data Intensive Computing Grant # 0916489
Kentucky Science & Engineering Foundation: KSEF-3113-RDE-017
Outline

• What can go Wrong in Machine Learning?
  o Unfair Machine Learning
  o Iterated Bias & Polarization
  o Black Box models

• Tell me more: Counter-Polarization

• Tell me why: Explanation Generation
“Twitter and Facebook can’t predict the election, but they did predict what you’re going to have for lunch: a tuna salad sandwich. You’re having the wrong sandwich.”
What Can Go Wrong in Machine Learning?

- We are relying on Machine Learning (ML) algorithms to support decisions:
  - **Recommender Systems:**
    - They **guide** humans in discovering **only a few** choices from among a **vast space** of options
  - **Choose among options:** Reading the News, Watching movies, Reading books, Discovering friends, Dating, Marriage, etc
  - **Supervised Learning:**
    - **Predict class label** for given instance
      - Example of label: whether to approve a loan, etc
    - Credit Scoring, Criminal investigation, Justice, Healthcare, Education, Insurance risk modeling, etc
Real life data can include **biases** that can affect the predictions

- May result in **unfair** ML models
  - discriminative,
  - unreasonable,
  - biased...

- **worse** when models are opaque/black box!
What Can Go Wrong in Machine Learning?

- Increasing (unchecked) Human-ML algorithm interaction...
  
  - Think about **Recommender Systems**
    - They **guide** humans in discovering **only a few** choices from among a vast space of options
  
  - Why are they needed?
    - Information Overload ⇒ need **Relevance Filters**!
  
  - But ...
    - could result in **hiding** important information from humans
    - could exacerbate **polarization** around divisive issues
    - could **fail to explain why they recommend** a particular choice (Black Box models: e.g, Matrix Factorization, Deep Learning)
What Can Go Wrong in Machine Learning?

Increasing unchecked Human-ML algorithm interaction...

Need for:
- Understanding Impact of interaction
- Limiting or reversing biases
  ⇒ Tell Me More!

- Adding Transparency / Explanations
  - to scrutinize biased or incorrect predictions
  - ⇒ more trust in ML models!
  ⇒ Tell Me Why?
Outline

• What can go Wrong in Machine Learning?
  o Unfair Machine Learning
  o Iterated Bias & Polarization
  o Black Box models

• Tell me more: Counter-Polarization

• Tell me why: Explanation Generation
Iterated Bias

Diagram:
- Machine Learning Algorithms
- Humans
- Information

Arrows indicate the iterative process between these components.
In the **past**, Machine learning algorithms relied on **reliable** labels from experts to build predictive models.

- **Expert** users, **limited** data, **reliable** labels

**Today**, algorithms receive data from the **general population**

- Labeling, annotations, etc.
- **Everybody** is a user, **Big Data**, **subjective** labels

- **Labeled Data** (User **Relevance labels**)
  - Machine Learning **Models**
    - **Filtering** of information visible to the user
      - **Next Labeled Data**
        - **Next ML Model**
          - **etc**

  - **Bias!**

- **Iterated Learning Bias!**
Recommender Systems

Collaborative Filtering

Uses previous ratings of the user to predict future preferences
Recommender Systems $\Rightarrow$ Iterated Bias

Recommendation system based on Machine Learning

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>
Recommendation system based on Machine Learning

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>5</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

Loved it
Liked it
It was ok
Disliked it
Hated it

Iterated Bias: results from repeated interaction between user & algorithm
Impact of Iterated Bias on Predicted Ratings

- Collaborative Filtering Simulation: Item-based, U=100, N=200
- Gini Index of the rating distributions vs iterations between rater and algorithm

Open loop

Closed loop

- Feedback loop / interaction between rater and recommender
  ⇒ Increases the divergence between ratings (Likes / Dislikes)
  ⇒ We are witnessing the birth of polarization

Note: Existing public benchmark data sets are useless for studying this problem!
  (1) they do not record every interaction
  (2) they do not have the absolute user preference on each item!

⇒ Need Benchmark human choice and rating cognitive models!
  (Shafto & Nasraoui, ‘Human-Recommender System’ RecSys 2016)
Polarization & Counter-Polarization in Recommender Systems
Machine Learning Algorithm

Interface (Output: predicted rating)

Interface (Input: rating data)

Human (user)
Positive Feedback Loop

Limited Historical Data  Temporal Discounting
Positive Feedback Loop
Positive Feedback Loop
Filter Bubble
Does this bore the user enough to leave the Recommender System?

Filter Bubble
Or rather...

Self-fulfilling Identity
Consequences

- Over Specialization
- User Unsatisfaction
- Misperceiving Facts
- Polarization
- Extreme Attitudes
- Low Sales Rates

Deconstructing non-prevailing views, opinions and behaviors
It gets worse in a *Polarized* environment!

**Definition of Polarization**

1. the action of polarizing or state of being or becoming polarized: such as
   a. (1): the action or process of affecting radiation and especially light so that the vibrations of the wave assume a definite form (2): the state of radiation affected by this process
   b. an increase in the resistance of an electrolytic cell often caused by the deposition of gas on one or both electrodes
   c. magnetization

2. a. division into two opposites
   b. concentration about opposing extremes of groups or interests formerly ranged on a continuum
Our survey ⇒

The field of polarization is rather not unified in

- how polarization is defined?

and

- what is done after recognizing it?

almost nothing...
1. **Social Polarization**: how people *congregate* with one another,

2. **Written Polarization**: how people *write* about topics,

3. **Rated and Recommended Polarization**: how people *behave*, *consume* and express their preferences,
   How they *interact with algorithms*. 
Basic Polarization Taxonomy

1. **Social Polarization**: how people congregate with one another,
2. **Written Polarization**: how people write about topics,
3. **Rated and Recommended Polarization**: how people behave, consume and express their preferences:
   - How they interact with algorithms

*What can we do about it?*
Polarization Detection Classifier - PDT

Data Science Pipeline:

- Data-driven problem formulation
- Feature engineering
- Modeling
  - Training a classifier using rating data
  - **Polarization Score** = predicted probability of belonging to the polarized class
- Evaluation
- Interpretation
Recommender System Counter Polarization Methods: RS-CP

During Recommendation

Pre-recommendation

Post-recommendation
Pre-recommendation Countering Polarization - PrCP

Why do we need it?

- Changing the Recommender System algorithm may not be always feasible
  - Black box
  - or too complex to modify ...

What do we do?

- **Transform the source data** to mitigate extreme ratings that make an item polarized.
- Take into account the **user's relative preferences**, yet **reduce extreme recommendation** that can be generated from a standard recommender system algorithm.
Pre-recommendation-based Countering Polarization - PrCP

Mapping Function:

\[ f : (U, I, R) \rightarrow (U, I, R') \text{ with probability of } p \]

\[ r'_{ij} = r_{ij} - \lambda_i \times (\bar{r} + g_i) \times \Phi_j^{\lambda_i + r_{ij}} \quad \text{if } r_{ij} \text{ is } \geq \delta \]

\[ r'_{ij} = r_{ij} + \lambda_i \times (\bar{r} - g_i) \times \Phi_j^{\lambda_i + r_{ij}} \quad \text{if } r_{ij} \text{ is } < \delta \]
Polarization-aware Recommender Interactive System - PaRIS

**Goal:**
Design a recommendation system which not only recommends *relevant items* but also may include *opposite views* in case the user is *interested to discover new items*.
Goal: Design a recommendation system which not only recommends relevant items but also includes opposite views in case the user is interested to discover new items.

Our Baseline: Non-negative Matrix Factorization (NMF)-based recommender systems:

- Good scalability
- High predictive accuracy
- Flexibility for modeling various real-life situations
- Easy incorporation of additional information
**Input**: Rating matrix

<table>
<thead>
<tr>
<th>user u</th>
<th>item v</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r_{uv}$</td>
</tr>
</tbody>
</table>

Rating from user $u$ to item $v$

**Idea**: Learn $p$ and $q$ to predict all values of the rating matrix
- $p$ and $q$ are the representation of the user $u$ and item $v$ in a latent space.

$$r_{uv} = q_v^T * p_u$$

**Learning process**:

$$\min_{P,Q} = \sum_{(u,v) \in R} (r_{uv} - q_v^T p_u)^2 + \lambda (\|q_v\|^2 + \|p_u\|^2)$$
PaRIS - Intuition
PaRIS - Intuition
PaRIS - Intuition
PaRIS - Intuition
Polarization-aware Recommender Interactive System - PaRIS

\[
\min (1 - \lambda_i) \times ||r_{ij} - p_i q_j||^2 + \lambda_i \times ||r'_{ij} - p_i q_j||^2
\]

\[
r'_{ij} = r_{ij} - (\bar{r} + g_i) \times \Phi_j^{\lambda_i + r_{ij}} \quad \text{if } r_{ij} \text{ is } \geq \delta
\]

\[
r'_{ij} = r_{ij} + (\bar{r} - g_i) \times \Phi_j^{\lambda_i + r_{ij}} \quad \text{if } r_{ij} \text{ is } < \delta
\]

- Initial rating
- Average rating
- Item gap ratio
- User Discovery Factor
- User Preference Threshold
- Polarization Score
Definition 3: Let the number of users, $|U| = n$ and number of items, $|I| = m$. A recommender system algorithm takes environment $G$ as input along with a user $u \in U$, and outputs a set of items $i_1, ..., i_{k_t} \in I$. 
NMF: Fully Polarized Environment

- It is **easy** and **fast** to learn discriminating models in a polarized environment!
  - The result: Keep each user in the safety of their preferred viewpoint
Effect of Increasing Polarization on NMF

<table>
<thead>
<tr>
<th>Gap = 0</th>
<th>Gap = 2</th>
<th>Gap = 4</th>
<th>Gap = 6</th>
<th>Gap = 8</th>
</tr>
</thead>
</table>

- **Gap = 0**: Initial distribution showing a mix of blue and red points indicating some level of polarization.
- **Gap = 2**: More pronounced separation, with blue and red points becoming distinct clusters.
- **Gap = 4**: Further increased separation, with clear clustering.
- **Gap = 6**: Strongly polarized distribution, with minimal overlap between clusters.
- **Gap = 8**: Extreme polarization, with completely separated clusters.

The graphs illustrate the change in error over iterations as the polarization increases, showing a decrease in error with increased gap values.
Effect of Polarization on NMF

Can monitor convergence trend to detect emergence of polarization!!
Counter Polarization Methods: Recommend **More** Items from Opposite View

<table>
<thead>
<tr>
<th></th>
<th>Opposite View Ratio</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OVHR_u</td>
<td>OVHR_t</td>
</tr>
<tr>
<td></td>
<td>mean, std</td>
<td>mean, std</td>
</tr>
<tr>
<td><strong>Classic NMF</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0% ± 0.00</td>
<td>0.0% ± 0.00</td>
</tr>
<tr>
<td>( \lambda_t = 0.2 )</td>
<td>5.4% ± 0.073</td>
<td>12.32 ± 0.31</td>
</tr>
<tr>
<td>( \lambda_t = 0.5 )</td>
<td>6.0% ± 0.08</td>
<td>18.1% ± 0.21</td>
</tr>
<tr>
<td>( \lambda_t = 0.7 )</td>
<td>61.0% ± 0.17</td>
<td>31.0% ± 0.167</td>
</tr>
<tr>
<td>( \lambda_t = 1.0 )</td>
<td>67.0% ± 0.24</td>
<td>68.0% ± 0.24</td>
</tr>
<tr>
<td>( \lambda_t = 0.2 )</td>
<td>5.4% ± 0.73</td>
<td>4.9% ± 0.021</td>
</tr>
<tr>
<td>( \lambda_t = 0.5 )</td>
<td>6.2% ± 0.075</td>
<td>5.2% ± 0.042</td>
</tr>
<tr>
<td>( \lambda_t = 0.7 )</td>
<td>7.0% ± 0.075</td>
<td>5.4% ± 0.033</td>
</tr>
<tr>
<td>( \lambda_t = 1.0 )</td>
<td>6.8% ± 0.064</td>
<td>5.8% ± 0.03</td>
</tr>
</tbody>
</table>
Conclusion

★ **Iterated Learning Bias**: theory and simulations

★ **Counter-polarization**
  - Empower the users who are increasingly entrapped in algorithmic filters
  - Allows humans to regain control of algorithm-induced filter bubble traps,
  - Impact on information filtering / recommender systems
    - News, social media, e-commerce, e-learning, etc

★ We uncovered **patterns** that are characteristic of environments where polarization emerges
  - Can monitor objective function optimization trend
  - ⇒ detect and quantify the evolution of polarization

★ ⇒ allow users to break free from their algorithmic chains!
Outline

• What can go Wrong in Machine Learning?
  o Unfair Machine Learning
  o Iterated Bias & Polarization
  o Black Box models

• Tell me more: Counter-Polarization

• Tell me why: Explanation Generation
Why is Explainability So Important?

Transparency is crucial to scrutinize:

- incorrect predictions
- biased predictions

More trustworthy ML models!
Black Box vs. White Box

- Black Box (opaque) predictors such as Deep learning and matrix factorization are accurate,
  - but lack interpretability and ability to give explanations
- White Box models such as rules and decision trees are interpretable (explainable)
  - but lack accuracy
- Explanations provide a rationale behind predictions
  → help the user gauge the validity of a prediction
  → may reveal prediction errors and reasons behind errors
  → increase trust between human and machine

Our Focus: Explanations in Recommender Systems
Recommender Systems

Collaborative Filtering

Uses previous ratings of the user to predict future preferences
Input Data

<table>
<thead>
<tr>
<th>User</th>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>I1</td>
<td>⭐⭐</td>
</tr>
<tr>
<td>U2</td>
<td>I2</td>
<td>⭐⭐⭐</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

ML Model

Black-Box Recommender

Recommendation

Input Data

<table>
<thead>
<tr>
<th>User</th>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>I1</td>
<td>⭐⭐</td>
</tr>
<tr>
<td>U2</td>
<td>I2</td>
<td>⭐⭐⭐</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

ML Model

Explainable Recommender

Recommendation/Explanation

“100 people with similar interests to you rated this show “5” out of “5”.

GAME OF THRONES
Tradeoff between Accuracy and Explainability

- Using Explanations, we can increase the transparency of the model.

- However there may be a downside:
  - Explainable models should also remain accurate!

Goal: a moderate tradeoff between accuracy and explainability

Explainability  Accuracy
MF: Matrix Factorization (Koren et al - 2009)

**Input Data**: Rating matrix

<table>
<thead>
<tr>
<th>user u</th>
<th>$r_{uv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>item v</td>
<td></td>
</tr>
</tbody>
</table>

Rating from user $u$ to item $v$

**Idea**: Learn $p$ and $q$ to predict all missing values of the rating matrix. $p$ and $q = \text{representation of user } u \text{ and item } v \text{ in a latent space.}$

$$r_{uv} = q_v^T \cdot p_u$$

**Learning process**: 

$$\min_{P, Q} \sum_{(u, v) \in R} (r_{uv} - q_v^T \cdot p_u)^2 + \lambda (\|q_v\|^2 + \|p_u\|^2)$$

**Main Problem**: Matrix Factorization is a **Black Box Model**
**EMF: Explainable Matrix Factorization ( Abdollahi & Nasraoui, 2016)**

**Idea:** Provide neighborhood style **Explanations** along with recommendations and **learn** a model that is **explainable**

**Recommendation:**

**Justification:**

80% of users who share similar interests with you liked this movie

**New objective function:**

\[ J = \sum_{(u, v) \in R} (r_{uv} - q_v^T p_u)^2 + \frac{\beta}{2} (\|p_u^2\| + \|q_v^2\|) + \frac{\lambda}{2} (p_u - q_v)^2 W_{uv} \]

\( W_{uv} \) = Explainability score calculated for user \( u \) and item \( v \).

- \( N' \): total number of neighbors of user \( u \) who rated item \( v \)
- \( N_k' \): total number of neighbors of user \( u \)

\[ W_{uv} \begin{cases} \frac{|N'(u)|}{|N'(u)|} > \theta; \\ 0 \quad \text{Otherwise;} \end{cases} \]
Classical Framework

1. Data for Recommendation Task
2. Recommender System
3. Explanation Generation
4. Output (Items+Explanation)
Classical Framework

- possible mismatch between (1) and (2)
- generally need to generate explanations at recommendation time (not efficient)
Classical Framework vs Proposed Framework

Data for Recommendation Task

(1) Recommender System

(2) Explanation Generation

Output (Items+Explanation)

Data for Recommendation/Explanation Task

Explanation-based Recommender System

Output (Items+Explanation)
Intuition
Intuition
Intuition: Bring explainable items *closer* to the user in latent space
Intuition: Now explainable item is more likely to be recommended
Active Learning

What If we make the algorithm **choose** the most useful training data?
1. Select items from an unlabeled pool of items using an **Active Learning selection strategy**
2. Obtain the true ratings of the selected item from the new user
3. Adjust the parameters of the model using the new ratings
4. Repeat the process until meeting a stopping criterion
Explainable Active learning Strategy Algorithm (ExAL)

New User

Recommender System

Active Learning

Select items that will improve explainability
Explainable Active learning Strategy Algorithm (ExAL)

New User

Ask for ratings for selected items

Active Learning

Select items that will improve explainability

Recommender System
Explainable Active learning Strategy Algorithm (ExAL)

- New User

- Ask for ratings for selected items
- Provide Ratings
- Active Learning

- Recommender System

Select items that will improve explainability
Explainable Active learning Strategy Algorithm (ExAL)

New User

Ask for ratings for selected items

Provide Ratings

Recommender System

Adjust the model using new ratings

Select items that will improve explainability

Active Learning
Explainable Active learning Strategy Algorithm (ExAL)

- Provide Ratings
- Adjust the model using new ratings
- Select items that will improve explainability
- New User
- Provide more explainable recommendations
- Ask for ratings for selected items
- Active Learning

Recommender System
Explainable Active learning Strategy Algorithm (ExAL)

Active Learning to improve explainability in MF

Problem:
How are we going to select the best items to be queried to the user?

Selection Criterion
Explainable Active learning Strategy Algorithm (ExAL)

Active Learning to improve explainability in MF

Proposition: A selection criterion for EMF to minimize testing error and increase explainability for user $u$:

$$i^* \text{ such that:}$$

$$i^*_u \simeq \arg\min_{i \in I^u_{pool}} \sum_{j \in I^u_{test}} \left| 1 - r_{uj} + 2\alpha ((r_{ui} - \bar{R}_i) \sum_{f=1}^{k} q_{if} q_{jf} + \lambda W_{ui} (r_{uj} - \sum_{f=1}^{k} q_{if} q_{jf})) \right|$$

- Index of the item that will be queried from the user
- Expected change in the accuracy of the testing error
- Explainability term that takes into consideration explainability as a selection criterion
Explainability F-score

Predictive Error (MAE)
Summary of Explainable Recommender Systems

- **EMF:** Explainable Matrix Factorization
  - Explainable Latent Factor Model

- **ERBM:** Explainable Restricted Boltzmann Machines for Recommender Systems
  - **Explainable Deep Learning Approach** for Collaborative Filtering

- Both EMF and ERBM:
  - improve explainability
  - without significant loss in accuracy

- **ExAL:** An *Active learning* approach to Explainable Recommendations
  - improves explainability **and** accuracy
References


Thank You!
Tell me Why? Tell me More!
Explaining Predictions, Iterated Learning Bias, and Counter-Polarization in Big Data Discovery Models

*CCS@Lexington, October 16, 2017*

**Olfa Nasraoui**
This work is a Collaboration with:
Behnoush Abdollahi, Mahsa Badami, Sami Khenissi, Wenlong Sun, Gopi Nutakki, Pegah Sagheb: @UofL
& Patrick Shafto: @Rutgers-Newark

Knowledge Discovery & Web Mining Lab
Computer Engineering & Computer Science Dept.
University of Louisville
http://webmining.spd.louisville.edu/
olfa.nasraoui@louisville.edu

Acknowledgements:
National Science Foundation:
NSF INSPIRE (IIS)- Grant #1549981
NSF IIS - Data Intensive Computing Grant # 0916489
Kentucky Science & Engineering Foundation: KSEF-3113-RDE-017