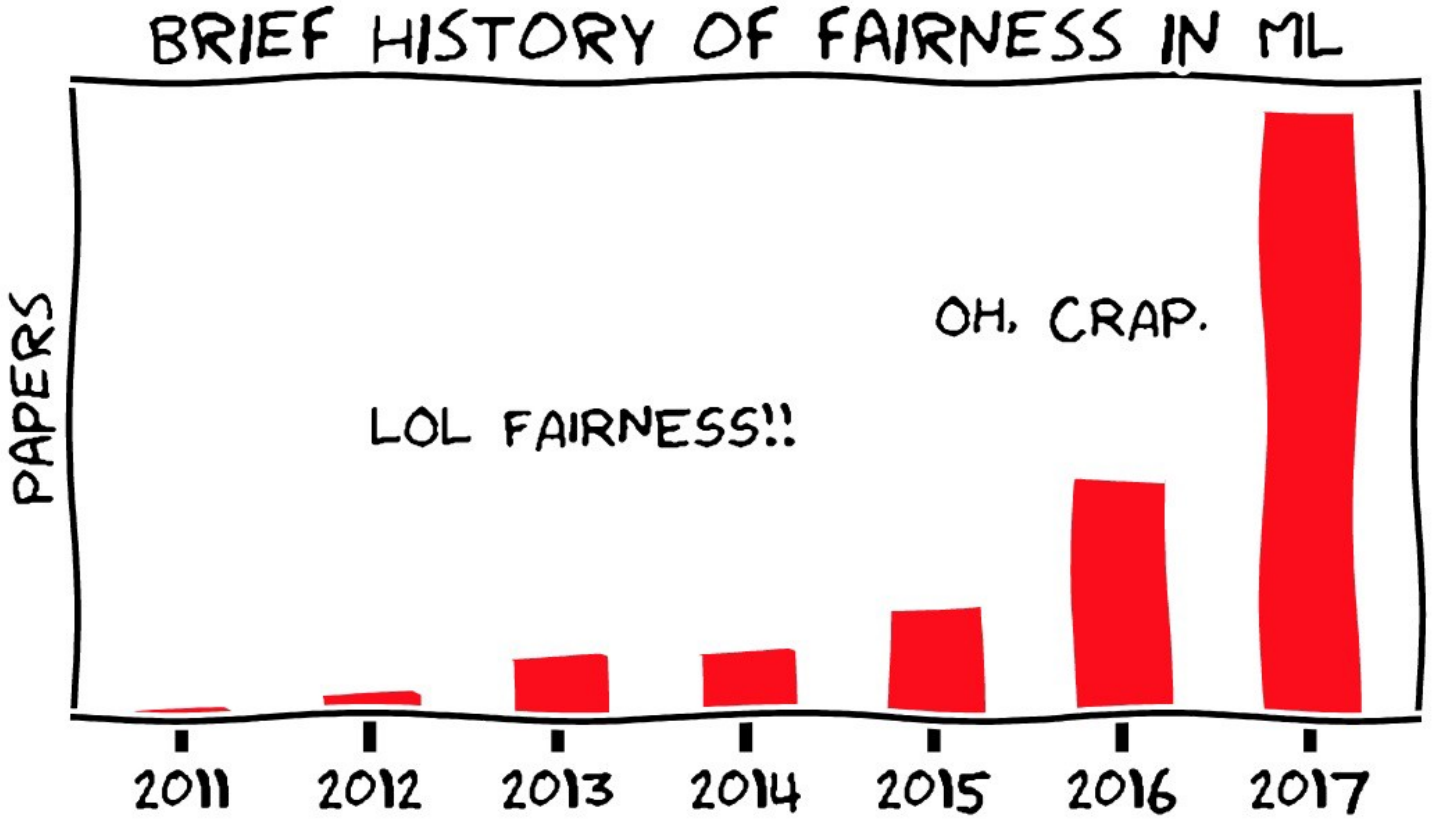
Recommender Systems Fairness

Professor Robin Burke
Spring 2019

Outline

- Fairness in machine learning
- Fairness in recommendation
- Sample algorithms
  - Fairness-aware re-ranking
  - Balanced neighborhood SLIM

Fairness in machine learning



Why do we care about this?

- Machine learning is increasingly used for applications that impact people's lives
  - Credit
  - Housing
  - Employment
  - Criminal justice proceedings
- Danger that biases in inputs lead to biases in outputs

Dangers of unfair algorithms

- People treat algorithms as more "rational" than humans
  - Less likely to question or appeal decisions
- Algorithms are (usually) opaque
  - Often proprietary
  - May be impossible to know why a decision was made
  - Design decisions made without adequate stakeholder input
- Positive feedback effects
  - Algorithm says "hire these people"
  - Cannot later learn that the people not hired would have been just as good
  - No counter-factual input

Very important!

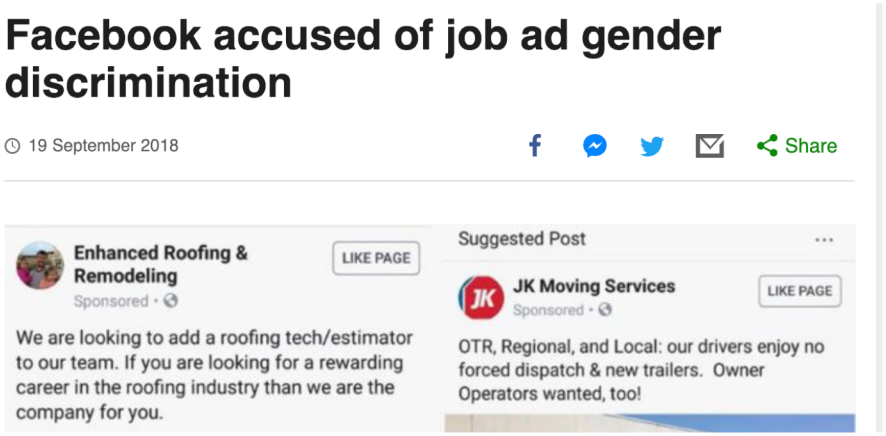
- If you plan to do machine learning for a living, you need to educate yourself
- Or if you just want to be an informed citizen in an algorithmic world

A police department uses a predictive system to decide where to send officers to patrol. The predictions are based on the number of people arrested in a given area the previous week. What might be some of the effects of such a system?

- A. Small discrepancies in arrest numbers will lead to large disparities in police presence
- B. Police presence will increase in poorer neighborhoods
- C. Police presence will increase in more densely-populated neighborhoods
- D. People who complain about policing disparities will be told that the system is objective
- E. All of the above

Fairness in recommendation

- What does it mean for recommendation to be fair?
  - ”Equals should be treated equally and unequals unequally.”
- Individuals have different preferences
  - should get different results
- But we have a sense that some kinds of recommendation outcomes can be unfair



Users don't always want fairness

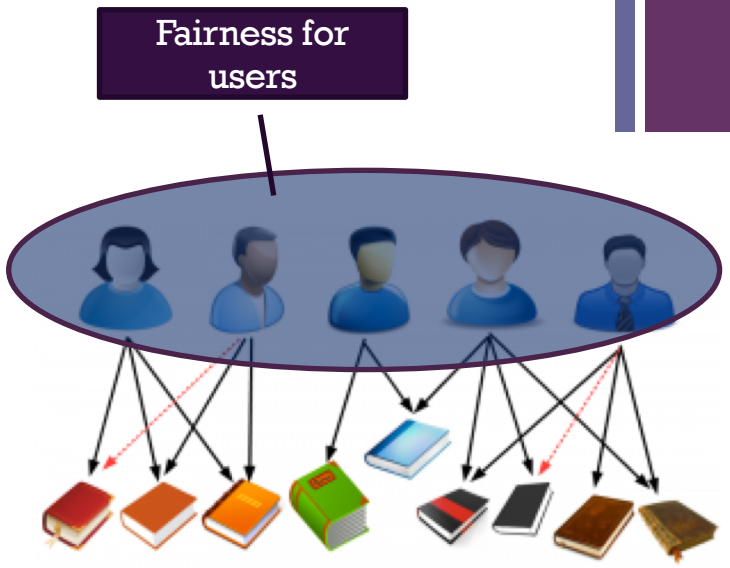
Bias ≠ Unfairness

- Biased (and discriminatory) rule
  - IF person has student ID, charge $10
  - IF person is over 65, charge $10
  - ELSE charge $20
- Discriminatory but (possibly) not unfair

Multiple sides

- Recommender systems may bring together multiple parties
- Consumers
  - People who get recommendations
- Providers
  - People who provide the items that get recommended
- Fairness may matter to each

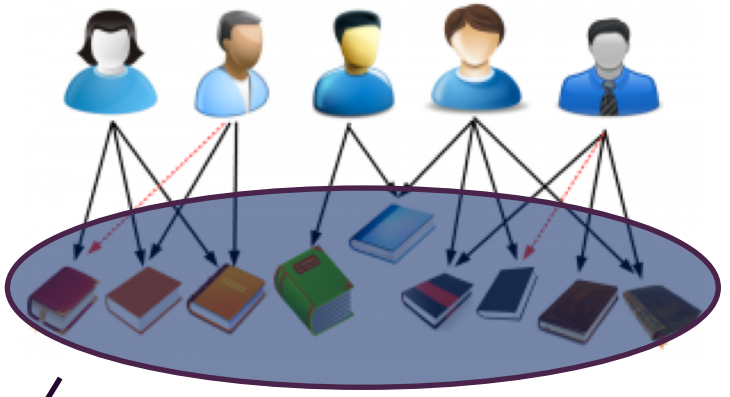
Consumer fairness case C-fairness

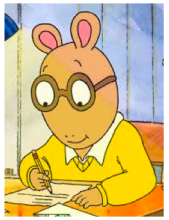


- Site may wish to be fair to the consumers of recommendations
  - Job seekers
- Example: male job seekers should not get better / different recommendations than female
  - Might be a legal requirement

Provider fairness case P-fairness

- Fairness relative to items being recommended
- Kiva cares about being fair to borrowers
- Does each loan have a fair chance of being recommended?
- Items linked to people who may be in protected groups





Fairness across items

Because of creators / owners

CP-fairness (PC-fairness?)

- Might need to combine both concerns
- Fairness for consumers and providers at the same time
- Example
  - Job recommendation
  - Protected groups in the user community
    - Female job seekers
  - Protected groups among the providers
    - Minority-owned businesses

A system is recommending rental apartments to students. State law requires that rental agencies not discriminate against users based on gender, marital status, race or religion. This application would require what kind of fairness:

- A. C-fairness
- B. P-fairness
- C. CP-fairness
- D. None, nobody cares if students are discriminated against.

Protected class

- Protected attribute
  - Gender, religion, race, sexual orientation, etc.
- Defines a protected class
  - Usually but not always a minority class
- Goal
  - Decisions should be independent of the protected attribute
  - Protected and unprotected cases treated the same if that's the only difference
- Defining what this means in practice is surprisingly tricky

Defining fairness

- Most definitions have to do with conditional probability
- Let A be the sensitive attribute that defines a protected group
  - Think male / female
- Let Y be the target classification variable
  - Think hire / don't hire
- Let R be the score returned by the classifier
- More than one definition

Independence

- $R \perp A$
  - The score is statistically independent of the sensitive attribute
- Another way to say this (for binary classifier)
  - $P(R = 1 \mid A = a) = P(R = 1 \mid A = b)$
  - The probability of "acceptance" is the same regardless of sensitive attribute
- $P(R = 1 \mid A = a) \geq P(R = 1 \mid A = b) - \in$
- $P(R = 1 \mid A = a) / P(R = 1 \mid A = b) \geq 1 - \in$

Impact ratio

Separation

- $R \perp A \mid Y$
  - Accepts that there may be some relationship between R and Y
  - For example, maybe there are fewer men with degrees in nursing
- For a binary classifier
  - $P(R = 1 \mid Y = 1, A = a) = P(R = 1 \mid Y = 1, A = b)$
  - $P(R = 1 \mid Y = 0, A = a) = P(R = 1 \mid Y = 0, A = b)$

True positive rate

False positive rate

Sufficiency

- $Y \perp A \mid R$
  - For any given score returned by the classifier, the target variable is independent of the sensitive attribute
- For binary classifier,
  - $P(Y = 1 \mid R = r, A = a) = P(Y = 1 \mid R = r, A = b)$
- Related to "calibration"

Your institution requires that every short list of job candidates has to have a least one woman on it. Which kind of fairness property is this? (You can think of being on the list as having R=1, and not on the list as R=0.)

- A. Independence ($R \perp A$)
- B. Separation ($R \perp A \mid Y$)
- C. Sufficiency ($Y \perp A \mid R$)
- D. None of these

Which definition is right?

- There is no general answer to this question!
- Worse yet
  - The definitions are not mutually compatible
- Talk to your legal team, your resident sociologist, etc. to decide what kind of fairness is important in your application

Measuring fairness

- Metrics can derived from these definitions
- For example, disparate impact
  - Related to independence
  - A common legal standard is
    - No more than 20% disparate impact
  - That is, the results for protected and unprotected users shouldn't differ by more than 20%
  - $P(R = 1 \mid A = a) / P(R = 1 \mid A = b) \geq 0.8$
    - $\in = 0.2$
- Similar metrics possible for other definitions

Note: intersectionality

- In many cases, there are multiple groups that might be protected
  - Race, age, gender, religion, marital status, etc.
- Quickly becomes complicated to apply these definitions
- "Rich sub-group fairness"
  - Active area of research right now

Note: binary features

- The model so far assumes that protected features are binary (or discrete)
- The real world isn't like that
  - No sudden change when you turn 65
  - Gender is a spectrum
  - Race is socially-defined construct, different in different social contexts

The reality

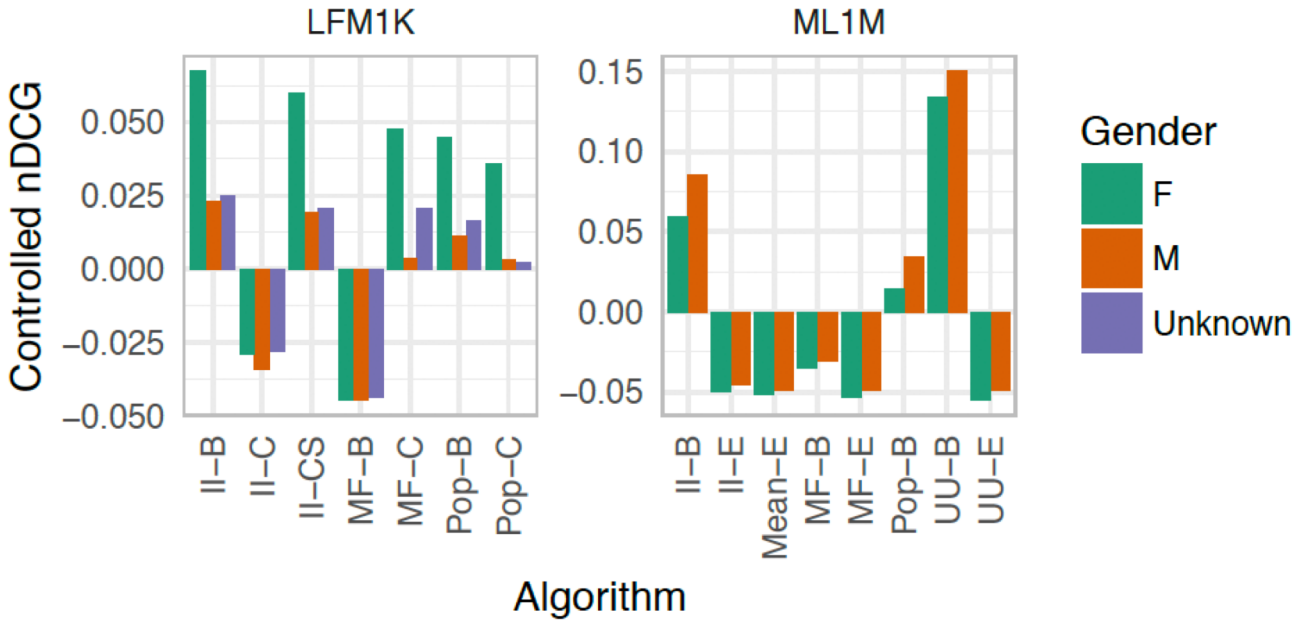
- Discrimination / bias is complex field with deep theoretical aspects
  - Law
  - Sociology
  - Psychology
- You are not going to fix this by writing the right objective function
  - Problems can creep in throughout the application development cycle
  - Important to be asking the right questions
  - Having the right people involved in decision making
  - Considering impact of automated systems

Consumer-side fairness

- How to determine "benefit"
  - Context- and domain-specific
  - Are some jobs higher-paying? More likely to lead to advancement?
- If you can define this,
  - Then you have Y (the "benefit")
  - Can apply these ideas

Error distribution

- Ekstrand et al. found that different recommendation algorithms had different nDCG (ranking accuracy) values for male and female users



Provider-side fairness

- You can treat being recommended as the benefit
  - Y = recommendation frequency
- But not every user is a good match with every recommendation
  - Might calculate "hits" instead
  - Recommendations that match user interest
- Again, context- and domain-specific

Algorithms

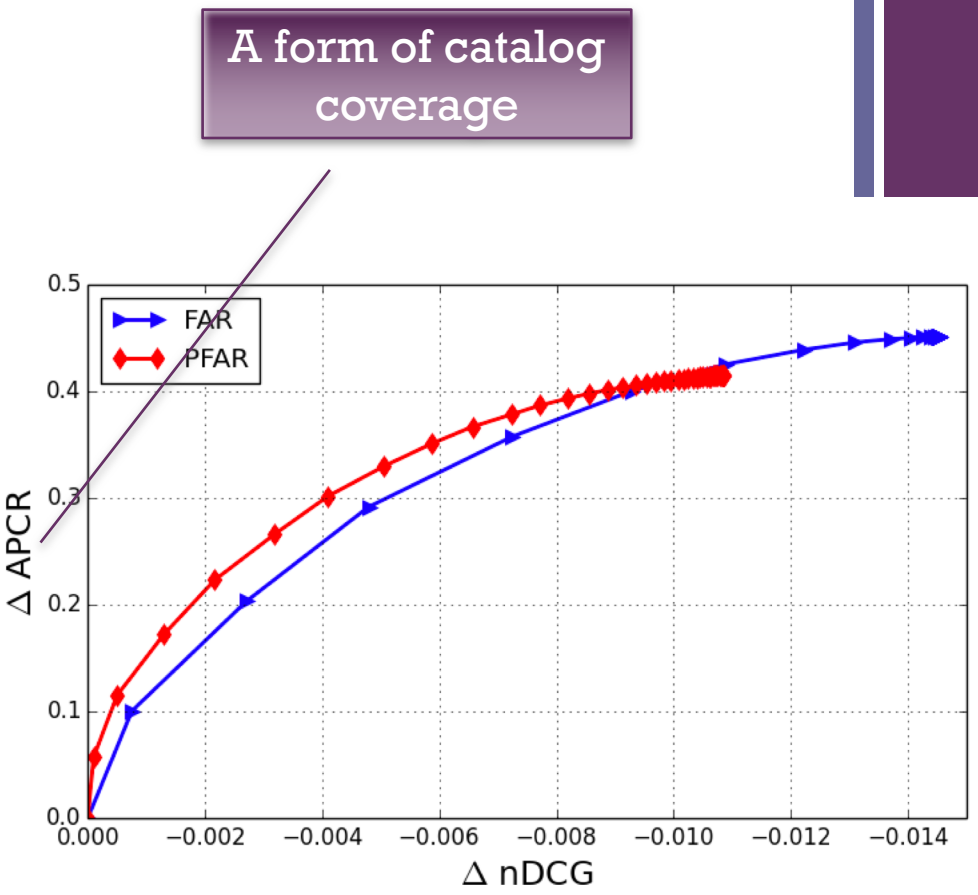
- P-fairness
- Post-processing = re-ranking
  - Take recommendation results and make them more fair
- Model-based
  - Build fairness into the algorithm
  - Balanced neighborhood SLIM

Fairness-oriented re-ranking

- We can treat sensitive groups as "aspects" for the purpose of re-ranking
- Similar to diversity-oriented re-ranking
- $\text{Score}(u,i) = P_o(i \mid u) + \lambda P(i|u, L)$
- Where L is the (re-ranked) list so far
  - And $P_o(i \mid u)$ is a function of the original recommender's score
- We compute $P(i|u, L)$ by making the assumption that the probability of the item should be higher if its inclusion will make the list more fair

Example

- Kiva.org
- Fairness across regions
- FAR = Fairness Aware Re-ranking
- PFAR = personalized version
  - Amount of fairness depends on user interest in regional diversity



Model-based method

- Kamishima, 2012
- P-fairness
- Loss function = $(R-\hat{R})^2 - \lambda_1 D(\hat{R}, A) + \lambda_2(\|U\|^2 + \|V\|^2)$
- Where D is the dependence term
  - Penalize solutions in which predicted ratings depend on the protected feature

Dependence terms

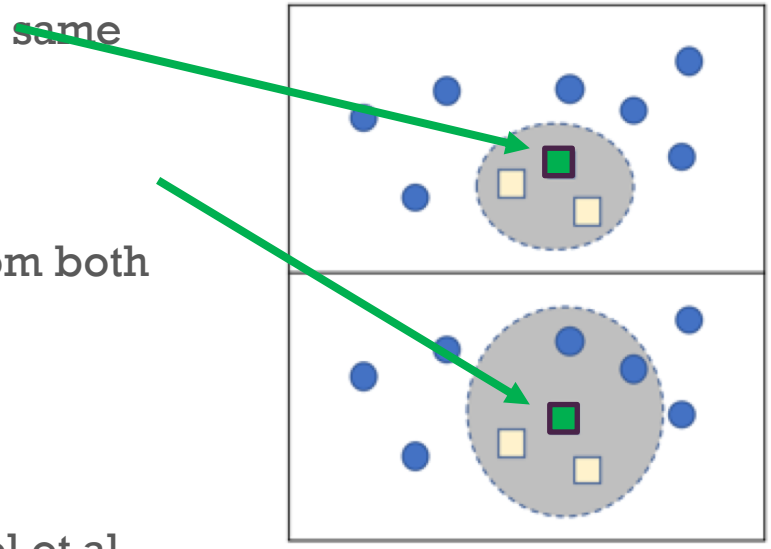
- $(\text{mean}(\hat{R}_{A=a})-\text{mean}(\hat{R}_{A=b}))^2$
  - Difference in mean between protected and unprotected groups
- Other dependence terms
  - Bhattacharyya distance
  - Mutual information between ratings and protected / unprotected distribution
- Familiar solution techniques
  - Compute gradients and update rules
  - Solve with gradient descent

Issue

- Criteria from classification may be too strict for recommendation
- We know that people like different items

Balanced Neighborhood SLIM

- Unbalanced neighborhood
  - Has only peers from the same (protected) group
- **Balanced neighborhood**
  - Has equal # of peers from both groups
- Inspiration
  - "Learning Fair Representations" (Zemel et al. 2013)



SLIM stands for

- A: Stochastic Linear Interpolation Method
- B: Sparse LInear Method
- C: Separable Latent Inference Matrices
- D: Simple Lecture Instruction Method

SLIM refresher

- Generalization of nearest neighbor
- Instead of discrete neighborhoods
  - We predict based on personalized regression equations
  - The coefficients define "near" and "far" items
- Turn into an optimization problem
  - Find the best weights
  - Use regularization to ensure sparsity
    - Many zeros

$$\hat{s}_{ij} = \sum_{k \in U} w_{ik} r_{kj},$$

$$\min_W \frac{1}{2} \|R - WR\|^2 + \lambda_1 \|W\|^1 + \frac{\lambda_2}{2} \|W\|^2,$$

Balanced neighborhoods

- How do we know a neighborhood is balanced?
  - Equal weights for protected and unprotected items
- Putting it all together
  - Loss function
- To minimize this
  - Use same technique as SLIM
  - Alternating least squares

$$b_i = \left(\sum_{w^+ in W_i^+} w^+ - \left| \sum_{w^- in W_i^-} w^- \right| \right)^2$$

$$L = \frac{1}{2} \|R - WR\|^2 + \lambda_1 \|W\|^1 + \frac{\lambda_2}{2} \|W\|^2 + \frac{\lambda_3}{2} \sum_{i \in U} \left(\sum_{k \in U} p_i w_{ik} \right)^2$$

Alternating least squares

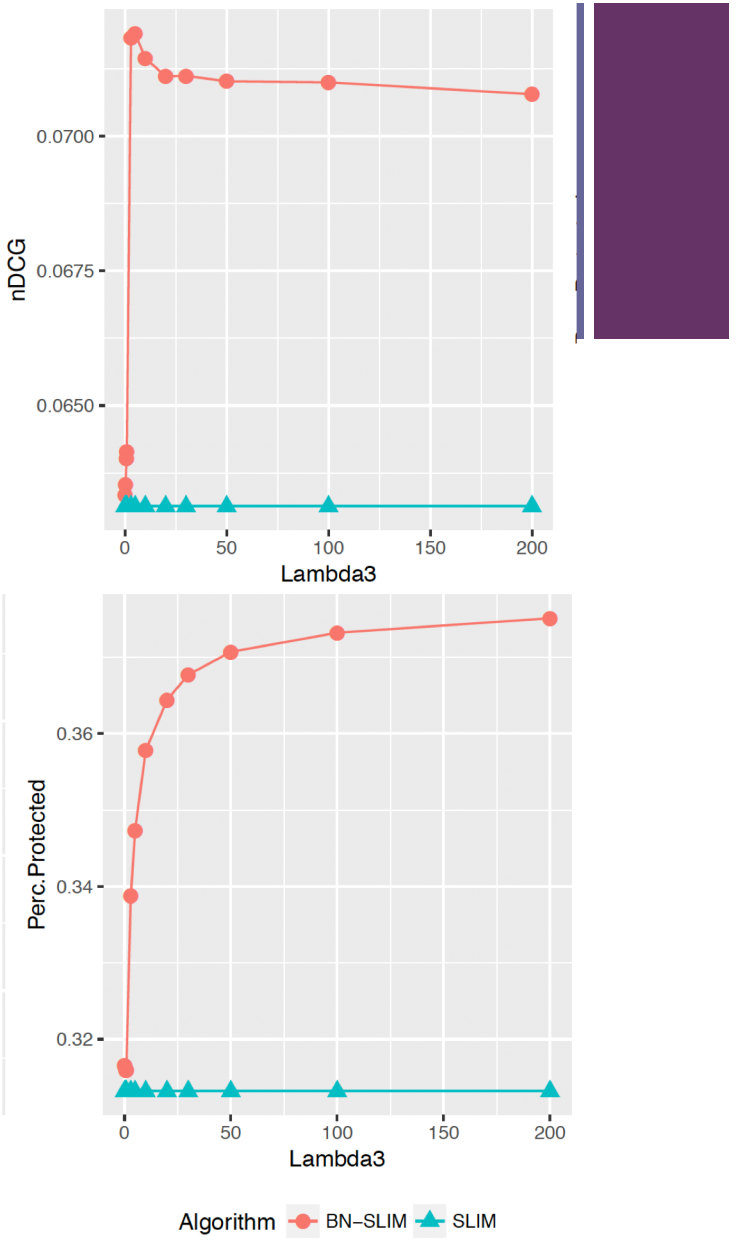
- Have to adapt because our loss function is different
- Iterative process
  - Pick a dimension
  - Set gradient to zero
  - Compute minimizing weight
  - Repeat

$$\frac{\partial L_i}{\partial w_{ik}} = \sum_{j \in I} (r_{ij} - \sum_{l \in U'} w_{il} r_{lj}) + w_{ik} \sum_{j \in I} r_{kj}^2 + \lambda_1 + \lambda_2 w_{ik} + \lambda_3 p_k \sum_{l \in U'} p_l w_{il}$$

$$w_{ik} \leftarrow \frac{S\left(\sum_{j \in I} (r_{ij} - \sum_{l \in U'} w_{il} r_{lj}) + \lambda_3 p_k \sum_{l \in U'} p_l w_{il}, \lambda_1\right)_+}{\sum_{j \in I} r_{kj}^2 + \lambda_2 + \lambda_3}$$

Results

- Improved fraction of protected items
  - Loans from underfunded countries
- Increased recommendation accuracy
  - Small, but statistically significant
- Usually this is framed as a tradeoff
  - Here synergy
- Possible reason
  - Reduction of overfitting
  - More diverse item neighborhoods



Fairness in recommendation

- Relatively new subfield in recommender systems
- Some tricky issues
  - Defining fairness
  - Allowing for personalization
  - Appropriate tradeoffs
  - Data sets hard to come by


LOONEY TUNES
"That's all Folks!"

Tuesday

- Project group work time
- I will have some clicker questions
- So show up!