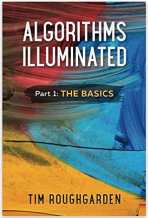
On Algorithms and Fairness

Jon Kleinberg

Cornell University

Includes joint work with
Sendhil Mullainathan, Manish Raghavan, and Maithra Raghu

Feuilleter



Algorithms Illuminated: Part 1: The Basics (Anglais)

Broché – 27 septembre 2017

de Tim Roughgarden (Auteur)

11 commentaires provenant des USA

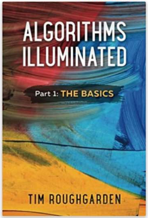
Voir les 2 formats et éditions

Format Kindle	Broché
EUR 8,99	**EUR 15,81**
Lisez avec notre **Appli gratuite**	1 neufs à partir de EUR 15,81

Voulez-vous le faire livrer le jeudi 18 jan.? Choisissez la **AmazonGlobal Eclair** au cours de votre commande. En savoir plus.

Note: Cet article est éligible à la livraison **en points de collecte.** Détails

Feuilleter



Algorithms Illuminated: Part 1: The Basics (Anglais)

Broché – 27 septembre 2017

de Tim Roughgarden (Auteur)

★★★★☆ 11 commentaires provenant des USA

Voir les 2 formats et éditions

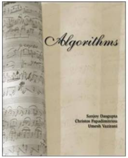
Format Kindle	Broché
EUR 8,99	**EUR 15,81**
Lisez avec notre **Appli gratuite**	1 neufs à partir de EUR 15,81

Voulez-vous le faire livrer le jeudi 18 jan.? Choisissez la **AmazonGlobal Eclair** au cours de votre commande. En savoir plus.

Note: Cet article est éligible à la livraison **en points de collecte.** Détails

---

Les clients ayant acheté cet article ont également acheté



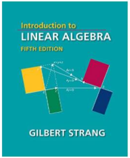
Algorithms

› Sanjoy Dasgupta

★★★★★ 2

Broché

EUR 34,64 prime



Introduction to Linear Algebra

› Gilbert Strang

★★★★★ 1

Relié

EUR 75,60 prime

Programmation Efficace Les 128 Algorithmes Qu'Il Faut Avoir Compris et…

› Christoph Dürr

★★★★☆ 14

Broché

EUR 26,00 prime

JOHN DOE

Full Address • City, State, ZIP • Phone Number • E-mail

OBJECTIVE: Design apparel print for an innovative retail company

EDUCATION:

UNIVERSITY OF MINNESOTA — City, State
College of Design — May 2011
- Bachelor of Science in Graphic Design
- Cumulative GPA 3.93, Dean's List
- Twin cities Iron Range Scholarship

WORK EXPERIENCE:

AMERICAN EAGLE — City, State
Sales Associate — July 2009 - present
- Collaborated with the store merchandiser creating displays to attract clientele
- Use my trend awareness to assist customers in their shopping experience
- Thoroughly scan every piece of merchandise for inventory control
- Process shipment to increase my product knowledge

PLANET BEACH — City, State
Spa Consultant — Aug. 2008 - present
- Sell retail and memberships to meet company sales goals
- Build organizational skills by single handedly running all operating procedures
- Communicate with clients to fulfill their wants and needs
- Attend promotional events to market our services
- Handle cash and deposits during opening and closing
- Received employee of the month award twice

HEARTBREAKER — City, State
Sales Associate — May 2008 – Aug. 2008
- Stocked sales floor with fast fashion inventory
- Marked down items allowing me to see unsuccessful merchandise in a retail market
- Offered advice and assistance to each guest

VICTORIA'S SECRET — City, State
Fashion Representative — Jan. 2006 – Feb. 2009
- Applied my leadership skills by assisting in the training of coworkers
- Set up mannequins and displays in order to entice future customers
- Provided superior customer service by helping with consumer decisions
- Took seasonal inventory

VOLUNTEER EXPERIENCE:

TARGET CORPORATION — City, State
Brand Ambassador — August 2009
- Represented Periscope Marketing and Target Inc. at a college event
- Engaged University of Minnesota freshman in the Target brand experience

Common Application FAQs

Here are step-by-step instructions and links to resources to help answer the questions you have asked most frequently about the Common Application. If you are having technical difficulties with completing and submitting your Common Application, you should seek assistance directly from the Common Application at "Ask A Question".

Does submission of the Common Application end with payment of your application fee?

Does Cornell require a Writing Supplement?

How do I make the Writing Supplement appear?

Am I able to view a PDF of my Common Application and Writing Supplement?

What forms may I print and submit via mail?

How do I submit the required documentation for my fee waiver request?

Where are the help resources on the Common Application website?

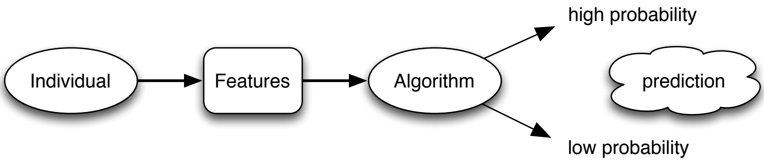
How do I submit a musical recording to be considered with my application?

QUESTIONS?

Forming Estimates of Future Performance

Estimating probability of a person's future outcome via algorithm.

- On-line content: engaging with content or an ad
- Employment: hiring decisions
- Education: admissions decisions
- Criminal justice: recidivism (future crime)



(1) Is the algorithm designed to focus on the right outcome?

(2) Does the algorithm have the right features for individuals?

(3) Are the algorithm's decisions *fair*?

Fairness in Risk Scores


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Angwin et al., ProPublica, 23 May 2016

COMPAS: An algorithm used in the U.S. criminal justice system to predict whether criminals will re-offend.

- Basic operation: assign a level of "risk" to each defendant.

Fairness in Risk Scores

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Angwin et al., ProPublica, 23 May 2016

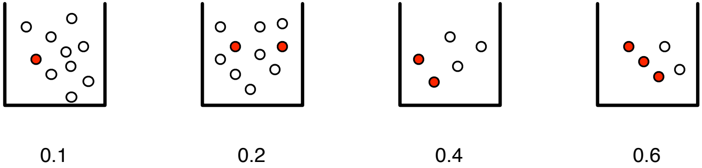
COMPAS: An algorithm used in the U.S. criminal justice system to predict whether criminals will re-offend.

- Basic operation: assign a level of "risk" to each defendant.

ProPublica's findings about COMPAS risk tool

- African-American defendants who didn't subsequently re-offend had higher average scores than white defendants who didn't re-offend.
- White defendants who subsequently re-offended had lower average scores than African-American defendants who re-offended.

Fairness in Risk Scores



How should we think about this concern?

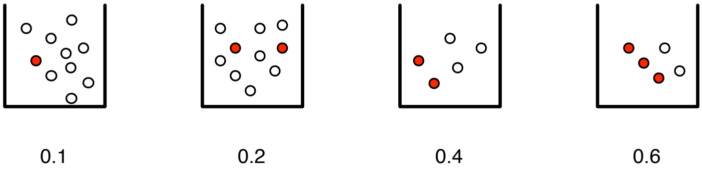
First, consider alternate definition [Dieterich et al, Flores et al 2016]:

COMPAS's scores are *well-calibrated* in each group.

- Consider all African-American defendants assigned a score of s.
- An s fraction of them go on to re-offend.
- The same is true for white defendants assigned a score of s.

A score of s means the same thing regardless of race.

Fairness in Risk Scores

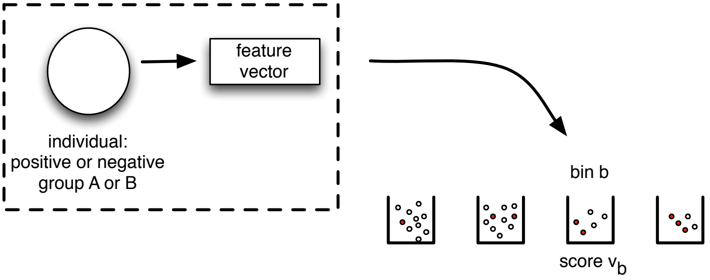


A concern about using an uncalibrated rule.

- Suppose hospitals hire doctors using a score that is not calibrated with respect to gender.
- Simple example: hire the candidates with the highest score s^*.
- Suppose female doctors with score s^* are more likely to be good doctors than male doctors with same s^*. (Failure of calibration.)
- Then: as a patient, it would be rational to choose your doctor (at least in part) based on their gender.

Criminal risk scores: we have calibration, but ProPublica's objections still remained. Could we achieve all the desired properties at once?

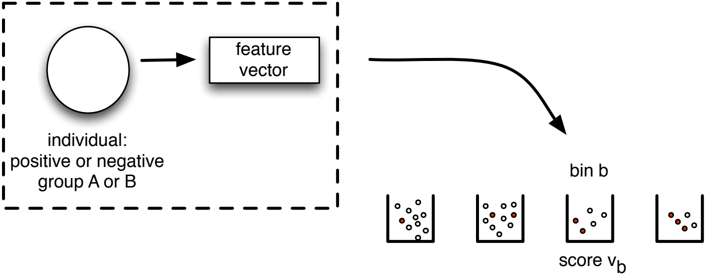
A Model of Risk Scores



Basic model for assigning scores as probability estimates.

- Individuals are either positive or negative (exhibit the behavior or not).
- Each individual belongs to group A or B.
- Each individual has a set of features, with the data we have access to.
- A risk score is a function mapping individuals to discrete "bins," where everyone in bin b is assigned a score of v_b.

A Model of Risk Scores



Desired properties:

- Calibration within groups: For each group, a v_b fraction of people in bin b are positive.
- Balance for the positive class: Average score of positive members in group A equals average score of positive members in group B.
- Balance for the negative class: Average score of negative members in group A equals average score of negative members in group B.

When are the Properties Achievable?

Can achieve all three properties in two simple cases.

- Perfect prediction: for each feature set, either everyone is in the negative class or everyone is in the positive class.
  (Then we can assign scores of 0 or 1 to everyone.)
- Equal base rates: the groups have the same fraction of positive instances.
  (Then there's a trivial risk score equal to this base rate for everyone.)

When are the Properties Achievable?

Can achieve all three properties in two simple cases.

- Perfect prediction: for each feature set, either everyone is in the negative class or everyone is in the positive class.
  (Then we can assign scores of 0 or 1 to everyone.)
- Equal base rates: the groups have the same fraction of positive instances.
  (Then there's a trivial risk score equal to this base rate for everyone.)

Theorem [Kleinberg-Mullainathan-Raghavan 2016]: In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rates.

When are the Properties Achievable?

Can achieve all three properties in two simple cases.

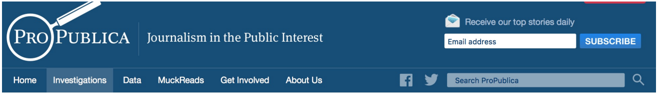
- Perfect prediction: for each feature set, either everyone is in the negative class or everyone is in the positive class.
  (Then we can assign scores of 0 or 1 to everyone.)
- Equal base rates: the groups have the same fraction of positive instances.
  (Then there's a trivial risk score equal to this base rate for everyone.)

Theorem [Kleinberg-Mullainathan-Raghavan 2016]: In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rates.

Notes:

- Not a theorem about computational power or inference power.
  It's a more basic limitation on assigning estimates to equalize averages.
- As such it applies to any decision procedure — algorithmic or human.

Concurrent Work on Related Themes



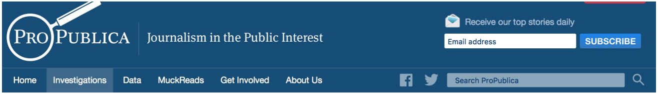
Machine Bias

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

---

Theorem [K-M-R 2016]: In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rates.

Concurrent Work on Related Themes



Machine Bias

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

---

Theorem [K-M-R 2016]: In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rates.

Chouldechova 2016 and CorbettDavies-Pierson-Feller-Goel 2016

- Classification using just "yes" / "no" rather than probability.

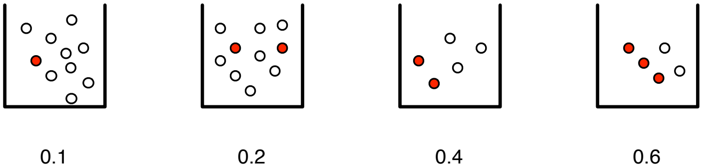
Hardt-Price-Srebro 2016

- Equalize false positive and false negatives, without calibration.

Pleiss-Raghavan-Wu-Kleinberg-Weinberger 2017

- Can achieve calibration together with any one linear function of scores on positive and negative classes, but not two in general.

Proof Sketch



Let N_t be the number of people in group t.
Let k_t be the number of people in the positive class in group t.

The calibration condition implies:

- The total score of all group-t people in bin b equals the expected number of group-t people in the positive class in bin b.

Summing over all bins:

- The total score of all group-t people equals the expected number of group-t people in the positive class.

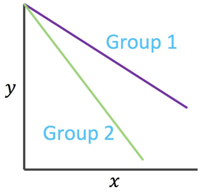
Proof Sketch

Let N_t be the number of people in group t.
Let k_t be the number of people in the positive class in group t.
(By calibration, k_t is also the total score in group t.)

Let x be the average score of a person in the negative class.
Let y be the average score of a person in the positive class.
(Note: independent of which group t we're talking about.)

Total score in group t is
$(N_t - k_t)x + k_t y = k_t$.

Rearranging:
$x = (1-y)\dfrac{k_t}{N_t - k_t}$.



Unless slopes are the same, only intersect at (0, 1)

Can We Achieve Approximate Guarantees?

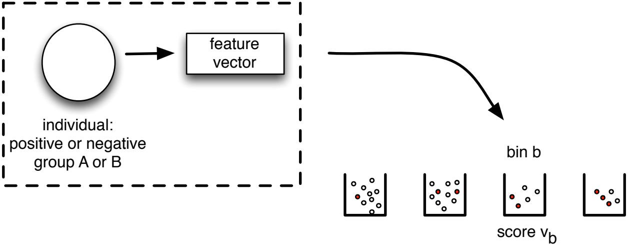
Approximate versions of our properties:

- Calibration within groups: For each group, approx. v_b fraction of people in bin b are positive.
- Balance for the positive class: Average score of positive members in group A is approx. average score of positive members in group B.
- Balance for the negative class: Average score of negative members in group A is approx. average score of negative members in group B.

Theorem [Kleinberg-Mullainathan-Raghavan 2016]: In any instance where all three properties can be approximately achieved, we must have either approximately perfect prediction or approximately equal base rates.

- Approximate versions of the conditions only hold in approximate versions of the two structured special cases.

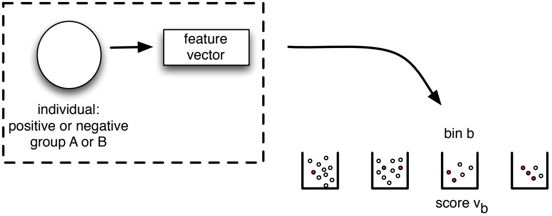
The Case of Equal Base Rates



If both groups have the same base rate p

- Can always create a single bin for each group with score p.
- Is there a *non-trivial* risk score assignment, where not every individual gets a score of p?
- Computationally hard for deterministic assignments, where everyone with the same feature set x must go to the same bin.

Reflections



- Inherent trade-offs between natural definitions of fairness.
- Many contexts in which we take complex data about individuals and rank them using a rule for assigning scores.
- May care about an objective formulated for a set, not an individual (e.g. if we are evaluating the diversity of the set [Page 2008])
  Finding the right score can be crucial [Kleinberg-Raghu 2015]

Second Theme: Objectives over Sets, not Individuals

CORNELLCHRONICLE

September 30, 2015

Search Chronicle

Bookmarks

Science, Tech & Medicine | Arts & Humanities | Business, Law & Society | Campus Life | Global Outreach | Archive

March 31, 2015

Admitted Class of 2019 is most diverse in Cornell history

By Daniel Aloi

Cornell's newly admitted class of freshmen is the most diverse and international in its 150-year history, with prospective undergraduates representing 100 nations from around the world, based on citizenship.

Cornell received 41,907 applications for admission to the Class of 2019, the second highest applicant pool in university history. A total of 6,234 applicants were admitted, and 3,590 were offered a place on a wait list.

TRENDING

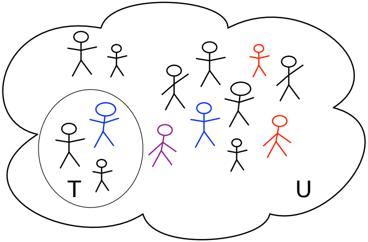
EDITOR'S PICKS | MOST EMAILED | MOST READ

Chemist Will Dichtel earns MacArthur 'Genius Award'

Coalition seeks to enhance lives of working families

Premarital births no longer predict breakups

- Cases — for example in hiring or school admissions — where we may care about the *set* we choose, not just individuals.
- Example 1: Diversity in admissions [Page 2008]
- Example 2: Evaluating full portfolio of loans, rather than one at a time.
- Example 3: Selecting a team to maximize its performance [Kleinberg-Raghu 2015]
- Using individual scores to achieve a group objective is challenging [Hong-Page 2004].



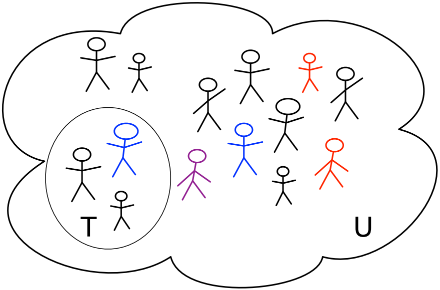
Given a set of n applicants, we want to choose a set T consisting of k of them.

- Future performance of each applicant i described by a random variable X_i. (Assume independence.)
- We give each applicant i a standardized test f, producing a numerical score $f(X_i)$.
- We select the k individuals with the highest test scores.

How good is the set T we select? It depends on:

- How we evaluate sets.
- How we define the test f.

Selecting for Maximum Performance



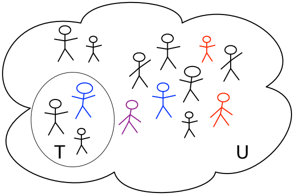
If the objective is the sum of expected individual performance:

- The standardized test for i should measure expected value of X_i.
- Choosing the k people with the highest test scores optimizes this objective.

Selecting for Maximum Performance

"Contest" objective:
the expected maximum of
the k random variables.

- E.g. we're scored by the best individual performance.



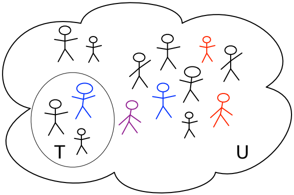
Example:

- 1000 candidates each produce value 500 with probability .001. 1000 candidates produce value 1 with probability 1.
- A test that evaluates the expectation chooses all the latter candidates; objective function is 1.
- If you choose all the former candidates, one of them achieves value 500 with probability approximately 40%.
- So if we choose these candidates, we get $\approx (.40) \cdot 500 = 200$.

Selecting for Maximum Performance

Is this the end; do tests not work for this problem?

- We just need a better test.



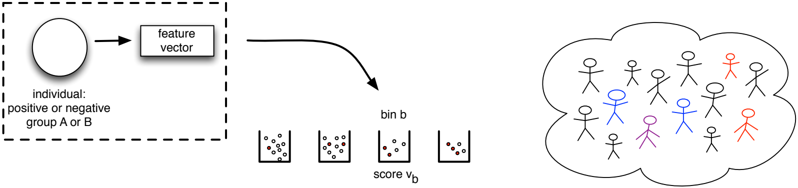
Test score of i is: the expected maximum of k independent draws from i's performance X_i.

- Theorem [Kleinberg-Raghu 2015]: Selecting the top k people according to this test produces performance that is approximately optimal.

Essentially, the test is evaluating i on "potential":
what's the expected best-case outcome if we choose i?

- Sometimes the problem is just that you're using the wrong test.
- Other performance measures where we can prove no test can yield near-optimal sets.

Reflections



- Inherent trade-offs between natural definitions of fairness.
- Many contexts in which we take complex data about individuals and rank them using a rule for assigning scores.
- May care about an objective formulated for a set, not an individual (e.g. if we are evaluating the diversity of the set [Page 2008])

  Finding the right score can be crucial [Kleinberg-Raghu 2015]