Sentiment Analysis

Sentiment Analysis

this movie was great! would watch again +

the movie was gross and overwrought, but I liked it +

this movie was not really very enjoyable —

- Bag-of-words doesn't seem sufficient (discourse structure, negation)
- There are some ways around this: extract bigram feature for "*not* X" for all X following the *not*

Pang et al. (2002)

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

- Simple feature sets can do pretty well!
- Learning alg. doesn't matter too much

- ME = "Maximum Entropy" = what we call Logistic Regression

Bo Pang, Lillian Lee, Shivakumar Vaithyanathan (2002)

Wang and Manning (2012)

- 10 years later — revisited basic BoW classifiers vs. other methods

Method	RT-s	MPQA
MNB-uni	77.9	85.3
MNB-bi	**79.0**	**86.3**
SVM-uni	76.2	86.1
SVM-bi	77.7	**86.7**
NBSVM-uni	**78.1**	85.3
NBSVM-bi	**79.4**	**86.3**
RAE	76.8	85.7
RAE-pretrain	77.7	**86.4**
Voting-w/Rev.	63.1	81.7
Rule	62.9	81.8
BoF-noDic.	75.7	81.8
BoF-w/Rev.	76.4	84.1
Tree-CRF	77.3	86.1

Before neural nets had taken off — results weren't that great

Kim (2014) CNNs **81.5** **89.5**

Wang and Manning (2012)

Multiclass Examples

"Now! ... *That* should clear up a few things around here!"

Entailment

- Three-class task over sentence pairs
- Not clear how to do this with simple bag-of-words features

A soccer game with multiple males playing.

ENTAILS

Some men are playing a sport.

A black race car starts up in front of a crowd of people.

CONTRADICTS

A man is driving down a lonely road

A smiling costumed woman is holding an umbrella.

NEUTRAL

A happy woman in a fairy costume holds an umbrella.

Bowman et al. (2015)

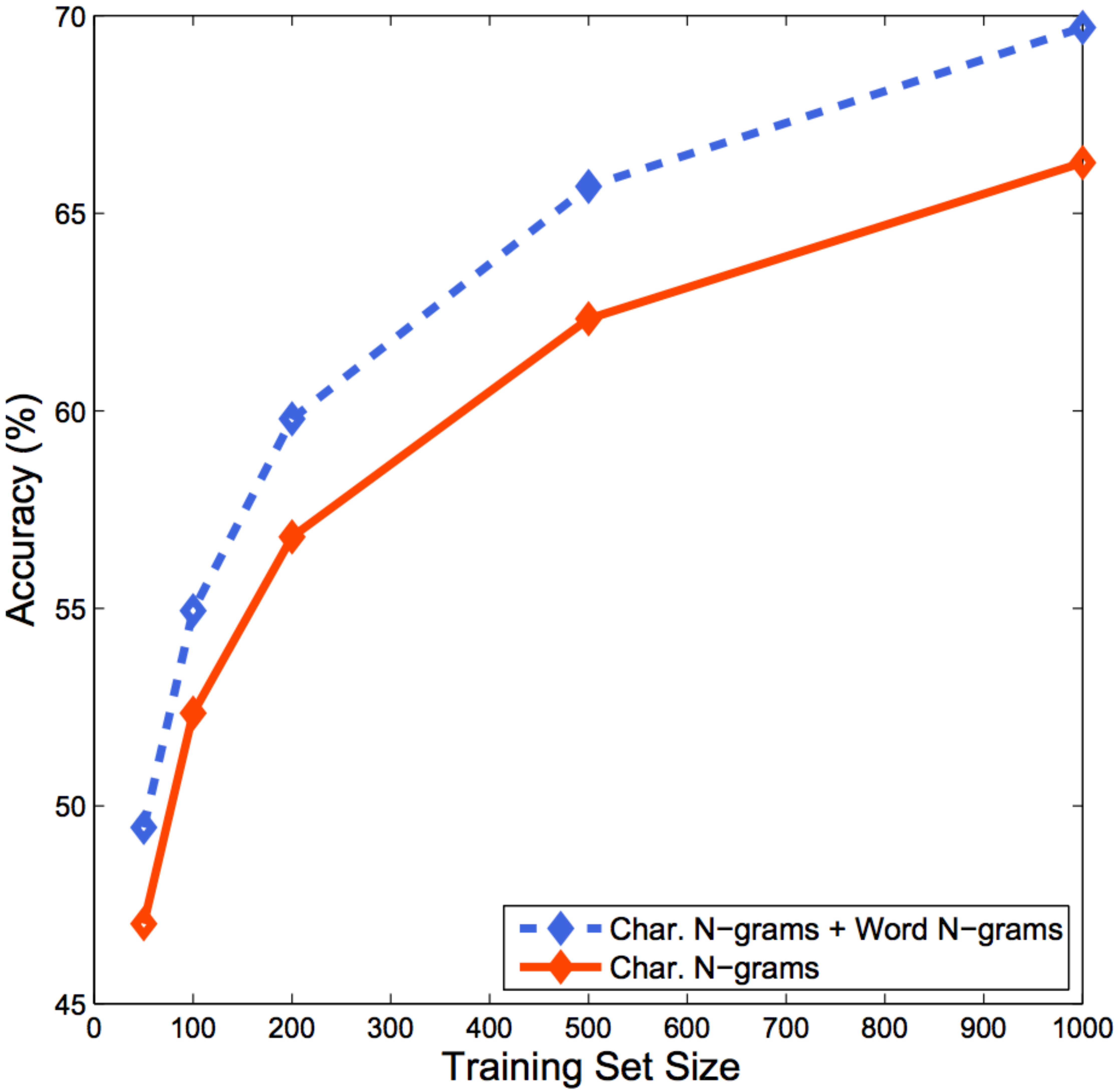
Authorship Attribution

- Statistical methods date back to 1930s and 1940s
  - Based on handcrafted heuristics like stopword frequencies
  - Early work: Shakespeare's plays, Federalist papers (Hamilton v. Madison)
- Twitter: given a bunch of tweets, can we figure out who wrote them?
  - Schwartz et al. EMNLP 2013: 500M tweets, take 1000 users with at least 1000 tweets each
- Task: given a held-out tweet by one of the 1000 authors, who wrote it?

Authorship Attribution

- SVM with character 4-grams, words 2-grams through 5-grams
- 1000 authors, 200 tweets per author => 30% accuracy
- 50 authors, 200 tweets per author => 71.2% accuracy

50-author case



Schwartz et al. (2013)

Authorship Attribution

- k-signature: n-gram that appears in k% of the authors tweets but not appearing for anyone else — suggests why these are so effective

Signature Type	10%-signature	Examples
Character n-grams	‘ ^_^’	REF oh ok ^_^ Glad you found it!
		Hope everyone is having a good afternoon ^_^
		REF Smirnoff lol keeping the goose in the freezer ^_^
	‘**yew** ’	gurl **yew** serving me tea nooch
		REF about wen **yew** and ronnie see each other
		REF lol so **yew** goin to check out tini’s tonight huh???

Schwartz et al. (2013)

Fairness

Fairness in Classification

- Classifiers can be used to make real-world decisions:
  - Who gets an interview?
  - Who should we lend money to?
  - Is this online activity suspicious?
  - Is a convicted person likely to re-offend?
- Humans making these decisions are typically subject to anti-discrimination laws; how do we ensure classifiers are *fair* in the same way?
- Many other factors to consider when deploying classifiers in the real world (e.g., impact of a false positive vs. a false negative) but we'll focus on fairness here

Fairness Response (SUBMIT ON CANVAS)

Consider having each data instance *x* associated with a **protected attribute A** when making a prediction. For example, suppose for sentiment analysis we also had information about the **ethnicity of the director** of the movie being reviewed.

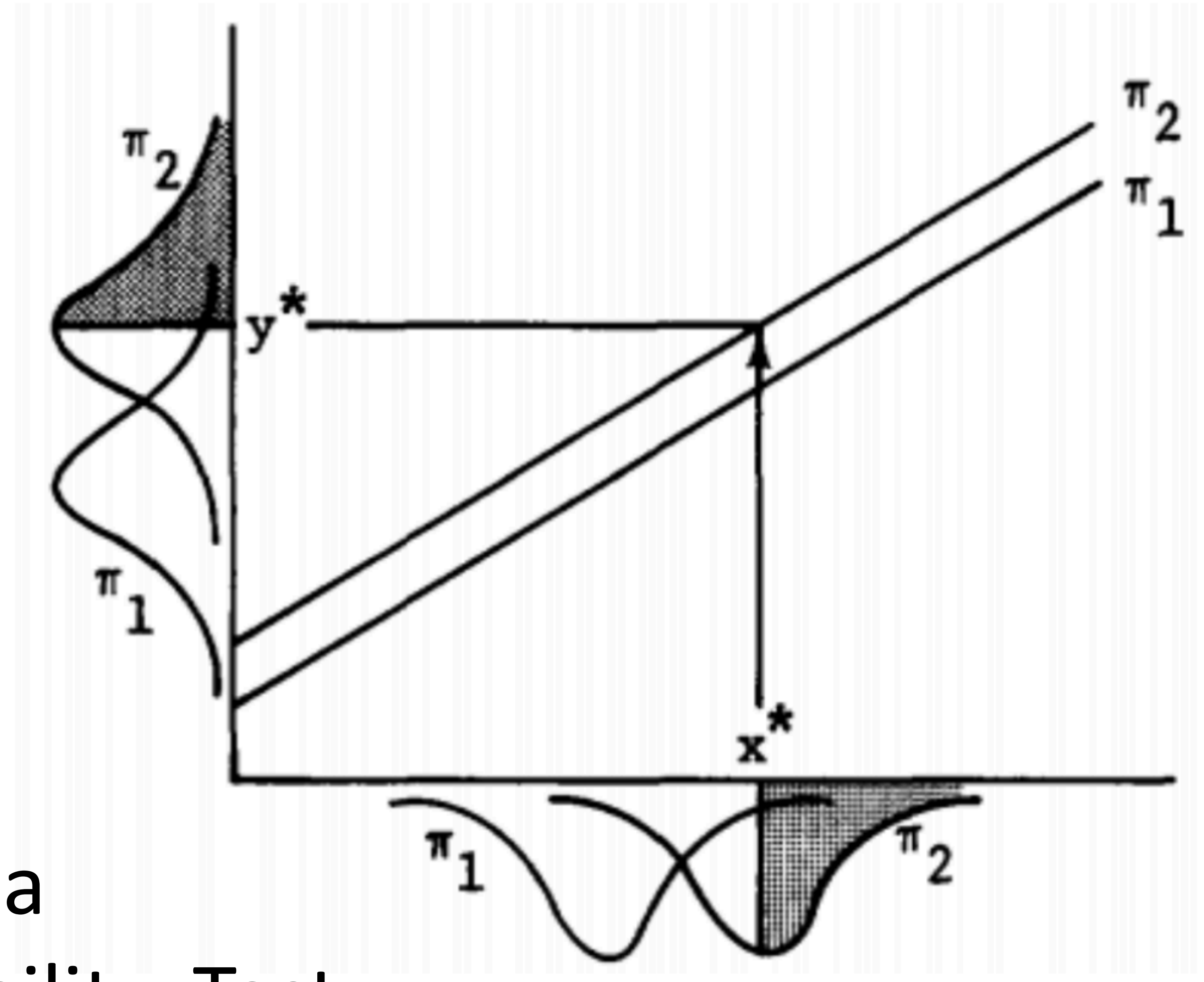
- What do **you** think it would mean for a classification model to be discriminatory in this context? Try to be as precise as you can!
- Do you think our **unigram bag-of-words** model might be discriminatory according to your criterion above? Why or why not?
- Suppose we add A as an additional "word" to each example, so our bag-of-words can use it as part of the input. Do you think the unigram model might be discriminatory according to your criterion? Why or why not?
- Suppose we enforce that the model must predict at least *k*% positives across every value of A; that is, if you filter to only the data around a particular ethnicity, the model must predict at least *k*% positives on that data slice. Is this fair? Why/why not?

Fairness in Classification

Idea 1: Classifiers need to be evaluated beyond just accuracy

- T. Anne Cleary (1966-1968): a test is biased if prediction on a subgroup makes *consistent* nonzero prediction errors compared to the aggregate

- Individuals of X group could still score lower on average. But the *errors* should not be consistently impacting X



- Member of π_1 has a test result higher than a member of π_2 for the same ground truth ability. Test penalizes π_2

Hutchinson and Mitchell (2018)

Fairness in Classification

Idea 1: Classifiers need to be evaluated beyond just accuracy

- Thorndike (1971), Petersen and Novik (1976): fairness in classification: ratio of predicted positives to ground truth positives must be approximately the same for each group ("**equalized odds**")
  - Group 1: 50% positive movie reviews. Group 2: 60% positive movie reviews
  - A classifier classifying 50% positive in both groups is unfair, regardless of accuracy
- Allows for different criteria across groups: imposing different classification thresholds actually can give a fairer result
- There are many other criteria we could use as well — this isn't the only one!

Petersen and Novik (1976)
Hutchinson and Mitchell (2018)

Discrimination

Idea 2: It is easy to build classifiers that discriminate even *without meaning to*

- A feature might correlate with minority group X and penalize that group:
  - Bag-of-words features can identify non-English words, dialects of English like AAVE, or code-switching (using two languages). (Why might this be bad for sentiment?)
  - ZIP code as a feature is correlated with race
- Reuters: "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization, women's colleges were negative-weight features
  - Accuracy will not catch these problems, very complex to evaluate depending on what humans did in the **actual** recruiting process

Credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Takeaways

- What marginalized groups in the population should I be mindful of? (Review sentiment: movies with female directors, foreign films, ...)

- Can I check one of these fairness criteria?

- Do aspects of my system or features it uses introduce potential correlations with protected classes or minority groups?