Aalto University

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Accuracy and fairness in binary classification

Indrė Žliobaitė
Aalto University School of Science, Department of Computer Science
Helsinki Institute for Information Technology (HIIT)
University of Helsinki

11 July, 2015
FATML workshop, Lille

Problem setting

- Given a discriminating dataset the goal is to build a classifier
  - as accurate as possible, and
  - obey non-discrimination constraints
- Binary classification, binary protected characteristic

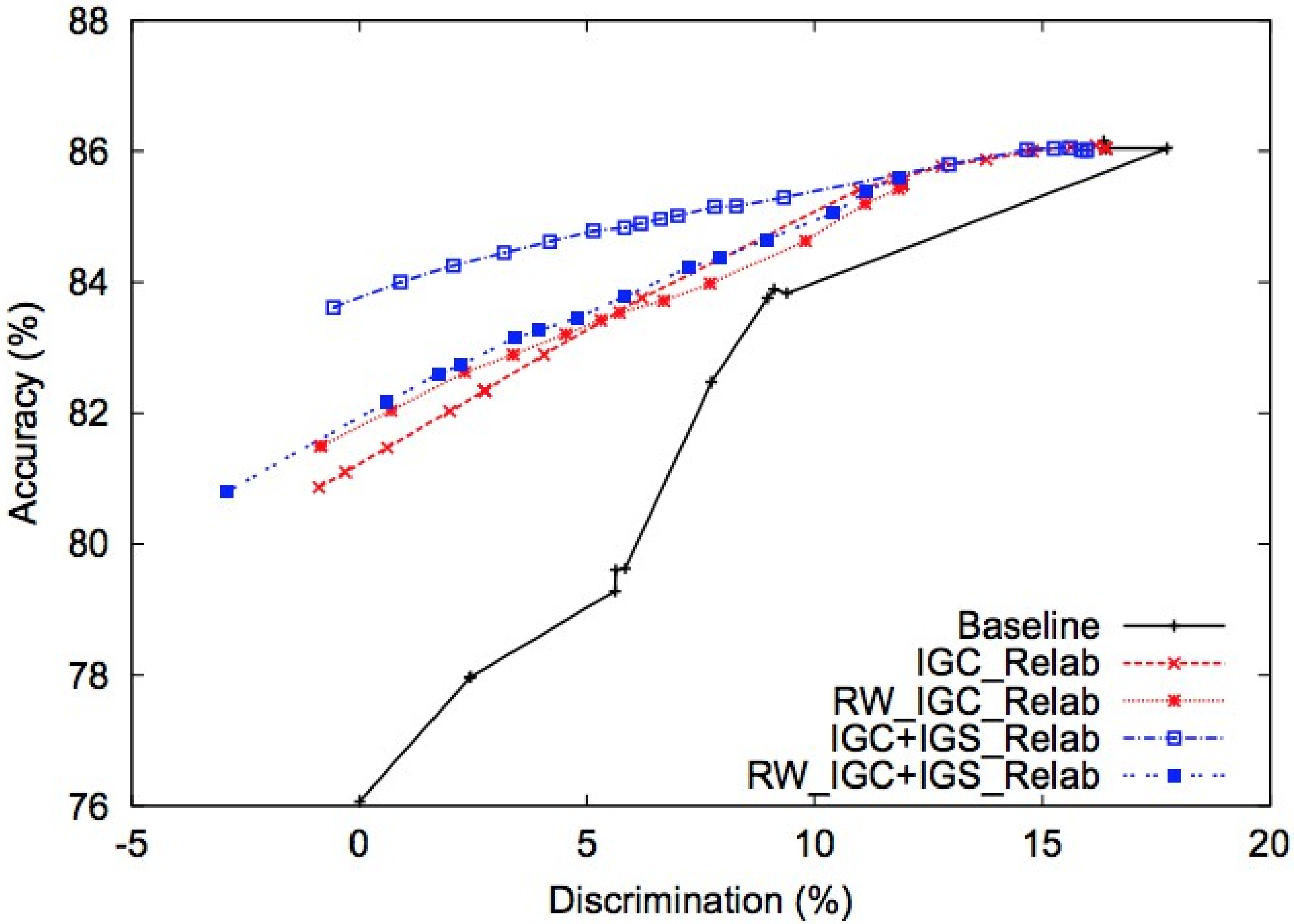
- Assumptions
  - Labels are objectively correct
  - Acceptance rates should be equal for the protected and general groups

- Discrimination measure: D = p(+|native) - p(+|foreign)

Measuring discrimination

Comparing non-discriminatory classifiers

Baseline Disc=19.3 Acc=76.3

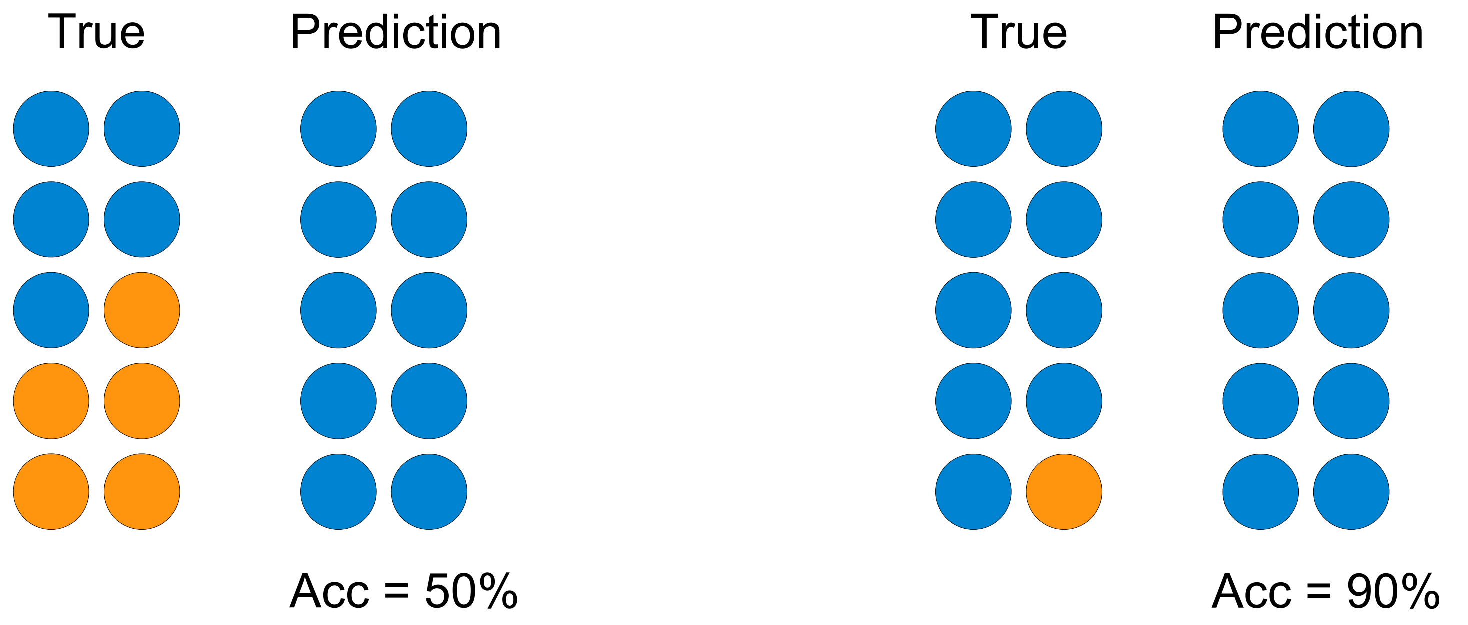


(a) Census Income Data

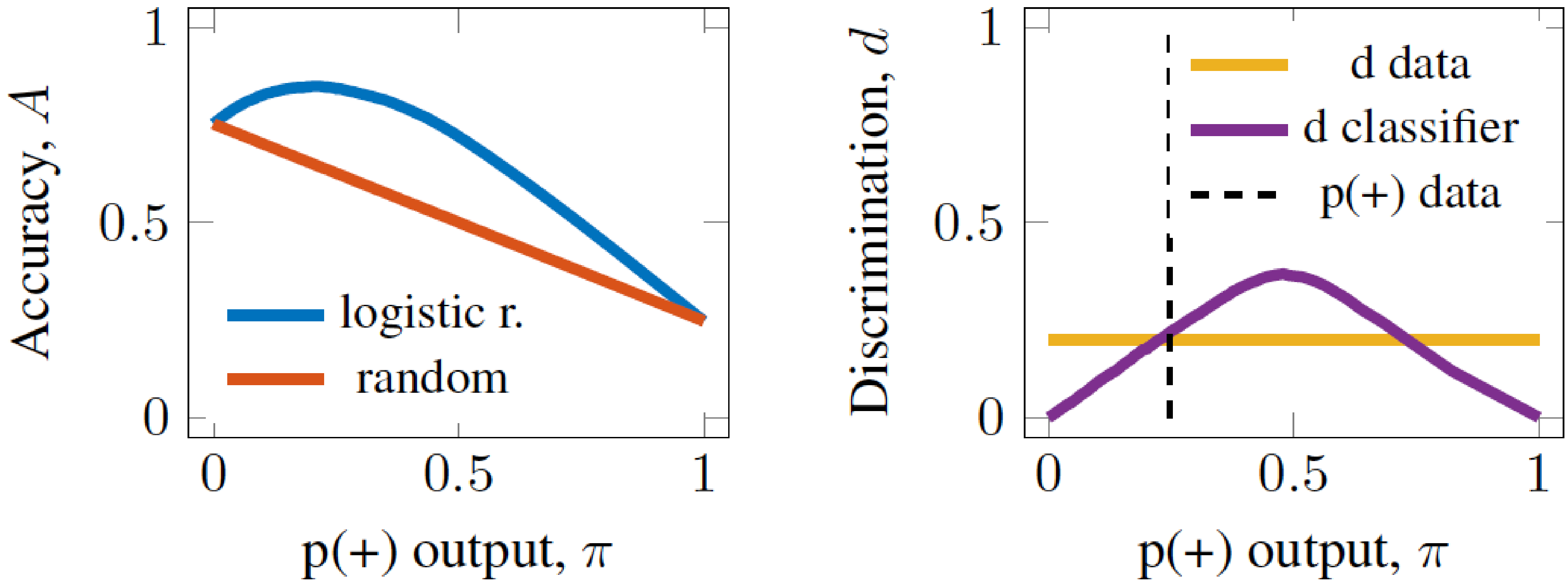
Kamiran et al 2010

Problem

- Baseline accuracy and baseline discrimination varies with varying overall acceptance rate
- Classifiers with different overall acceptance rates are not comparable



Experiment



Adult dataset from UCI

Baseline discrimination

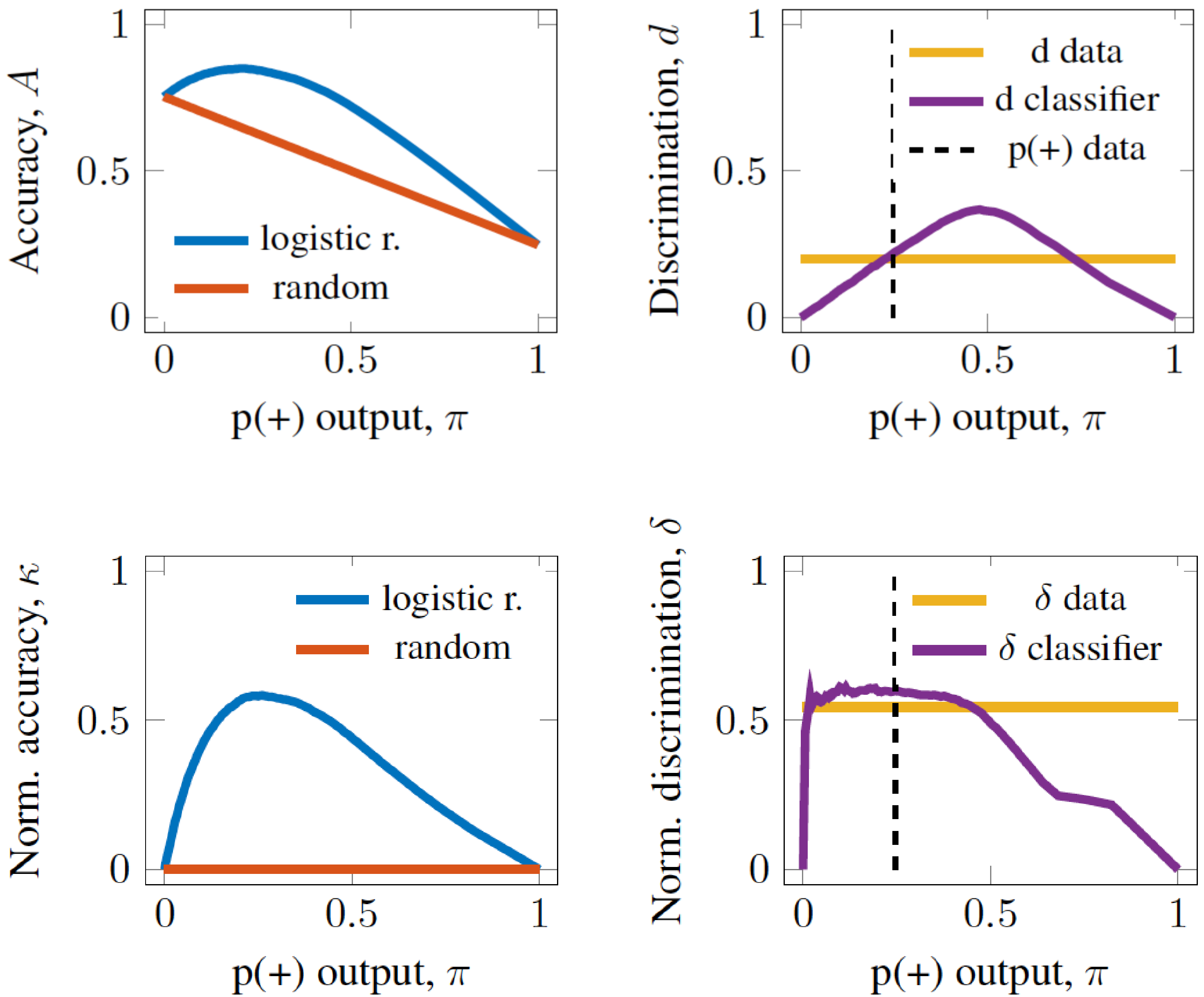
- Maximum discrimination when first everyone from the favored community is accepted, only then members of protected community start to get accepted

Normalized measures

- We propose to normalize discrimination by dmax
  - d = D/dmax, where $d_{max} = \min\left(\frac{\pi}{\alpha}, \frac{1-\pi}{1-\alpha}\right)$ acceptance rate p(+) proportion of natives p(native)
    - 1 max, 0 no discrimination, <0 reverse discrimination

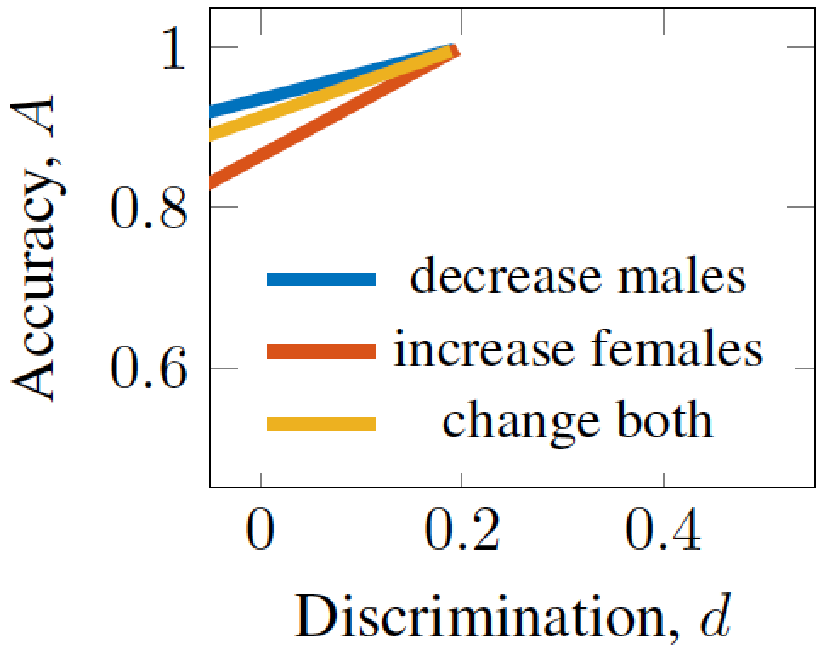
- We recommend normalizing accuracy – Cohen's kappa
  - k = (Acc – RAcc) / (1 – Racc)
    - 1 max, 0 like random, <0 very bad

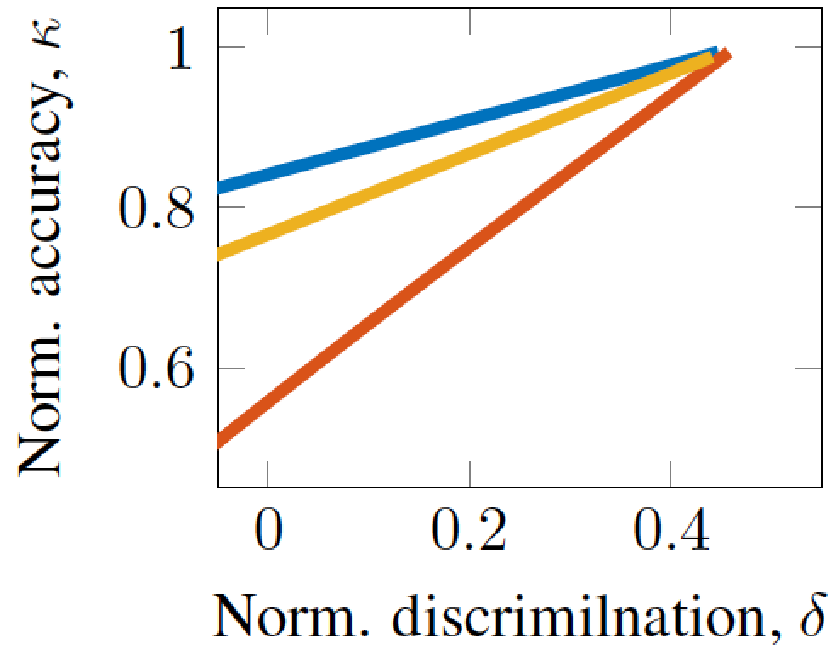
Experiment cont.



Discrimination prevention

What is the best we can do?





If data is correct, reducing discrimination will always reduce accuracy

Performance of discrimination prevention strategies

	p(+) π	Acc. A	Disc. d	N. acc. κ	N. disc. δ
Data/oracle	24.7	100	19.9	100	54.4
Logistic with s	20.2	84.9	18.3	56.7	**61.4**
Logistic no s	20.1	84.9	17.6	56.6	**59.6**
Logistic massage	22.1	83.5	6.9	53.9	21.3
NB with s	15.4	81.9	13.5	44.2	**59.7**
NB no s	14.4	81.4	10.9	41.7	51.3
NB massaged	15.4	81.5	6.8	43.3	29.7
Tree J48 with s	19.6	85.1	17.9	56.9	**61.9**
Tree J48 no s	19.6	85.0	17.9	56.7	**61.8**
Tree massage	22.9	83.5	6.1	54.6	18.1

Removing s does not solve the problem
Decreasing acceptance rates may show lower nominal discrimination

Performance of discrimination prevention strategies

	p(+) π	Acc. A	Disc. d	N. acc. κ	N. disc. δ
Data/oracle	24.7	100	19.9	100	54.4
Logistic with s	20.2	84.9	18.3	56.7	**61.4**
Logistic no s	20.1	84.9	17.6	56.6	**59.6**
Logistic massage	22.1	83.5	6.9	53.9	21.3
NB with s	15.4	81.9	13.5	44.2	**59.7**
NB no s	14.4	81.4	10.9	41.7	51.3
NB massaged	15.4	81.5	6.8	43.3	29.7
Tree J48 with s	19.6	85.1	17.9	56.9	**61.9**
Tree J48 no s	19.6	85.0	17.9	56.7	**61.8**
Tree massage	22.9	83.5	6.1	54.6	18.1

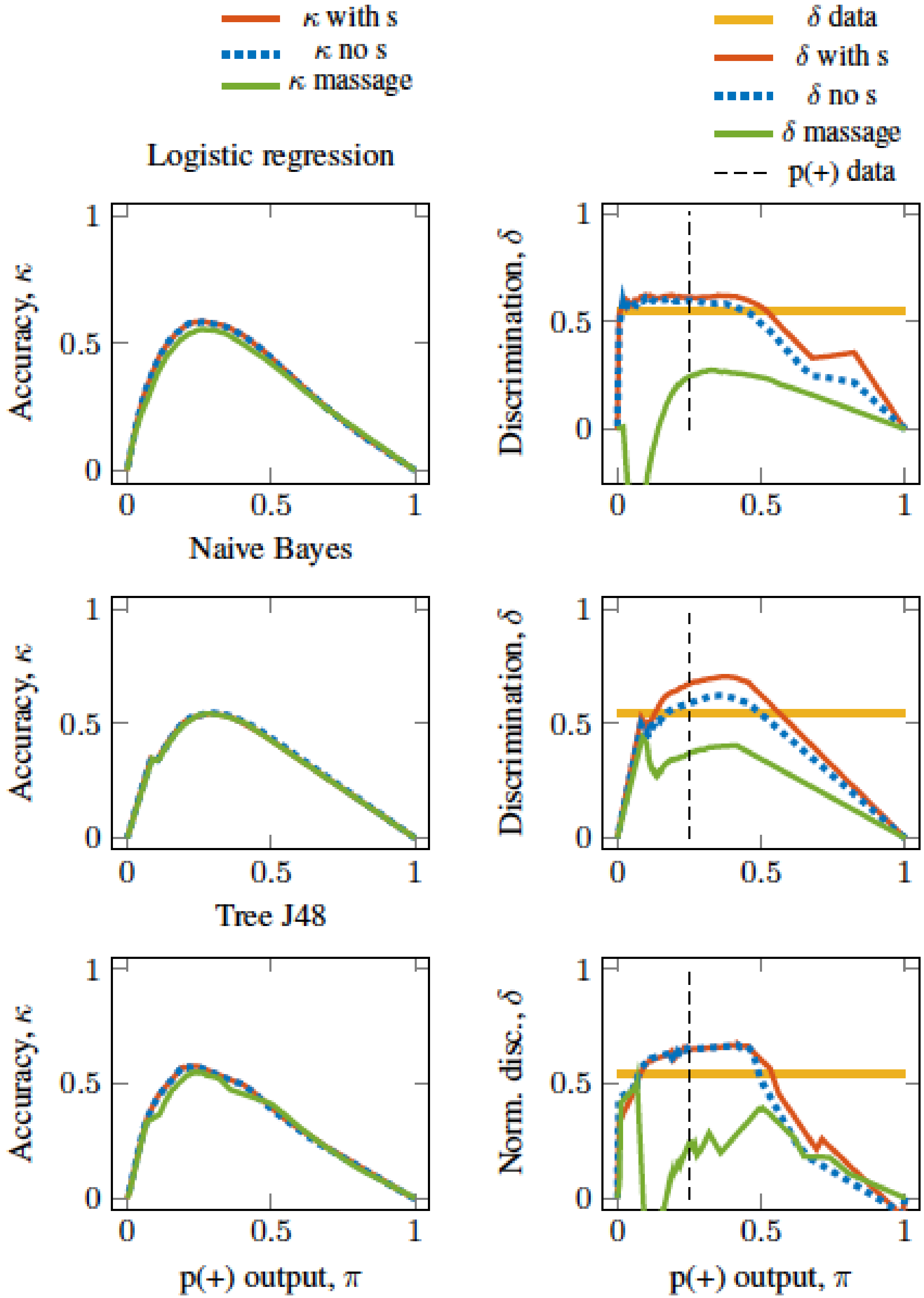
Removing s does not solve the problem
Decreasing acceptance rates may show lower nominal discrimination

Performance of discrimination prevention strategies

	p(+) π	Acc. A	Disc. d	N. acc. κ	N. disc. δ
Data/oracle	24.7	100	19.9	100	54.4
Logistic with s	20.2	84.9	18.3	56.7	**61.4**
Logistic no s	20.1	84.9	17.6	56.6	**59.6**
Logistic massage	22.1	83.5	6.9	53.9	21.3
NB with s	15.4	81.9	13.5	44.2	**59.7**
NB no s	14.4	81.4	10.9	41.7	51.3
NB massaged	15.4	81.5	6.8	43.3	29.7
Tree J48 with s	19.6	85.1	17.9	56.9	**61.9**
Tree J48 no s	19.6	85.0	17.9	56.7	**61.8**
Tree massage	22.9	83.5	6.1	54.6	18.1

Removing s does not solve the problem
Decreasing acceptance rates may show lower nominal discrimination

Massaging



κ with s
κ no s
κ massage
δ data
δ with s
δ no s
δ massage
p(+) data
Logistic regression
Naive Bayes
Tree J48
Accuracy, κ
Discrimination, δ
Norm. disc., δ
p(+) output, π
0
0.5
1

Concluding remark

- Evaluation of non-discriminatory classifiers needs to take into account acceptance rates, otherwise the results of different classifiers (or even different parameter settings) are not comparable

Thanks!