Reduzindo discriminação em classificadores

Hilton Pintor (hpbl)
Victor Miranda (vmm)

github.com/if1015-datascience-ufpe/2018-2-ex3-p2p

Tópicos

- Decisores
  - O que são?
  - Naive Bayes
  - Árvores de decisão
- Problemática
  - Discriminação em bases de dados
    - O que é discriminação
    - Como afeta decisores
  - Impacto em casos reais
- Soluções
  - Dependency-aware tree construction
  - Leaf Relabeling
    - Resultados
  - Modifiying naive bays
  - Two naive Bayes models
  - Latent variable model
    - Resultados

O que são decisores?

Naive bayes

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Árvores de Decisão

Fisher's *Iris* Data [hide]

Dataset Order	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	*I. setosa*
2	4.9	3.0	1.4	0.2	*I. setosa*
3	4.7	3.2	1.3	0.2	*I. setosa*
4	4.6	3.1	1.5	0.2	*I. setosa*
5	5.0	3.6	1.4	0.3	*I. setosa*
6	5.4	3.9	1.7	0.4	*I. setosa*
7	4.6	3.4	1.4	0.3	*I. setosa*
8	5.0	3.4	1.5	0.2	*I. setosa*
9	4.4	2.9	1.4	0.2	*I. setosa*
10	4.9	3.1	1.5	0.1	*I. setosa*
11	5.4	3.7	1.5	0.2	*I. setosa*
12	4.8	3.4	1.6	0.2	*I. setosa*

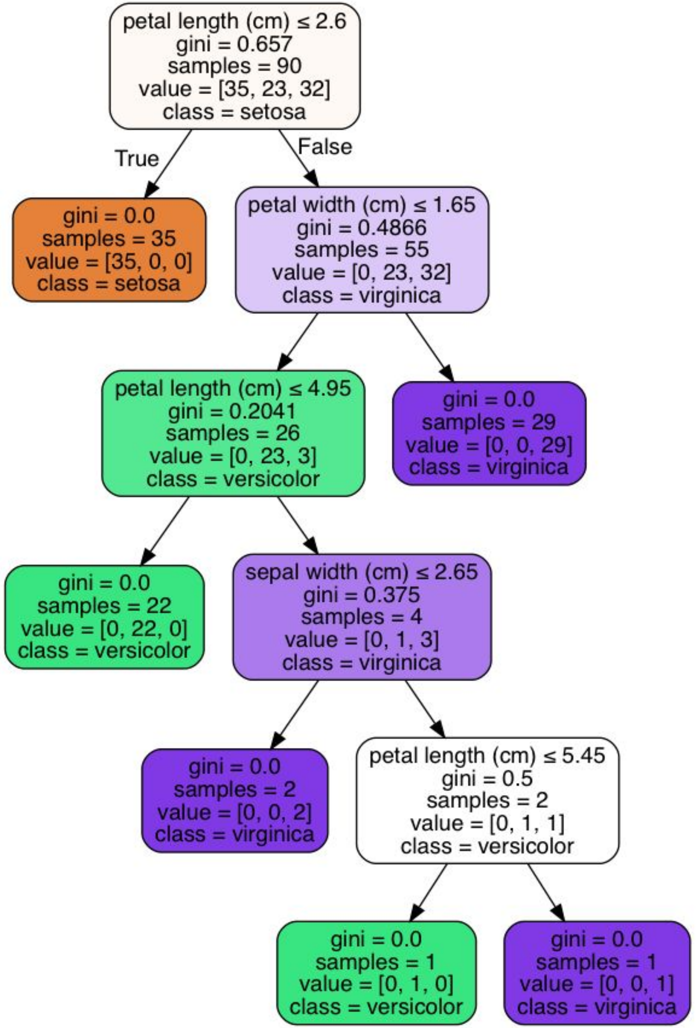
- Dados de treinamento
- Dados de teste
- Iris data
- Classificar espécie da flor

[setosa, versicolor, virginica]

Árvores de Decisão

- Questionamento sobre os atributos nos nós
- Particiona os dados
- Classifica nas folhas

[setosa, versicolor, virginica]



Discriminação em decisores

O que é discriminação?

"Discrimination is a sociological term that refers to the **unfair and unequal treatment** of individuals of a certain group based solely on their affiliation to that particular group, category or class.

Such discriminatory attitude **deprives** the members of one group **from the benefits and opportunities** which are accessible to other groups"

Discriminação nos decisores

$$disc_B(C,D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

For $\epsilon \in [0,1]$, the formula $disc_B(C,D) \leq \epsilon$ is called a *non-discriminatory constraint*.

Exemplos práticos

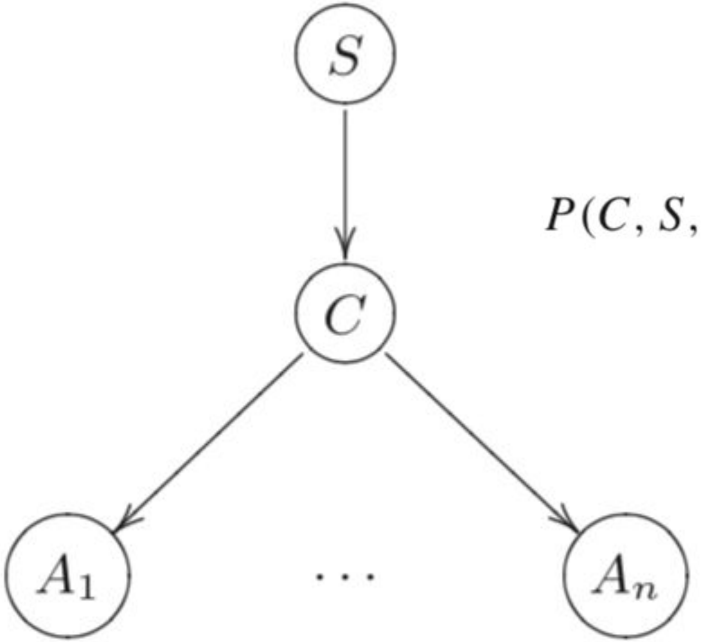
- Seguros
- Liberdade Condicional
- Bancos

Soluções propostas

Modelos tradicionais

- Remover atributos sensíveis
- Massaging
- Reweighing
- Red-lining

Modifiying naive Bayes



$$P(C, S, A_1, \ldots, A_n) = P(C)P(S|C)P(A_1|C)\ldots P(A_n|C)$$

Modifiying naive Bayes

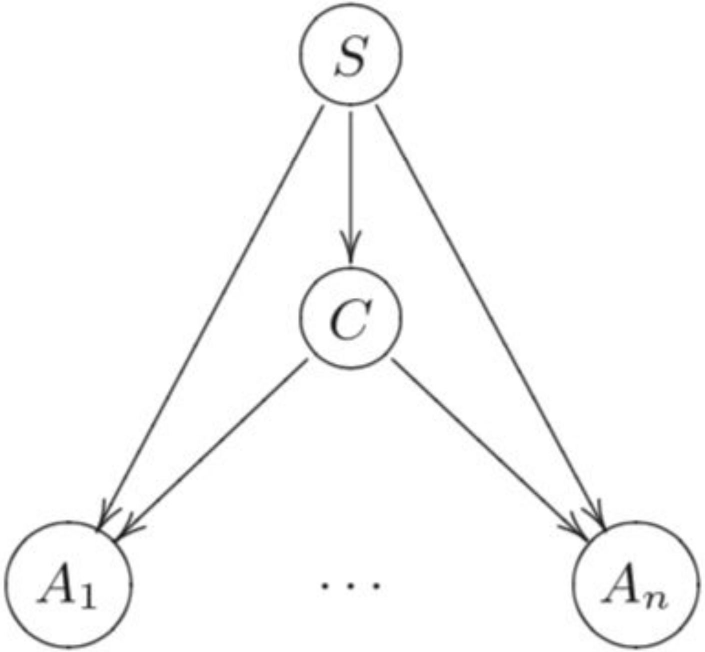
Algorithm 1 Modifying naive Bayes

Require: a probabilistic classifier M that uses distribution $P(C|S)$ and a data-set D

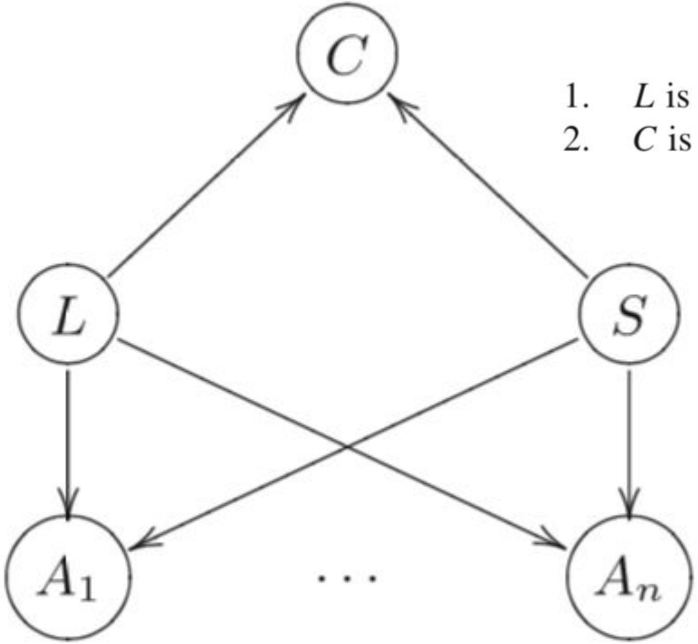
Ensure: M is modified such that it is (almost) non-discriminating, and the number of positive labels assigned by M to items from D is (almost) equal to the number of positive items in D

Calculate the discrimination *disc* in the labels assigned by M to D
while $disc > 0.0$ **do**
  numpos is the number of positive labels assigned by M to D
  if $numpos <$ the number of positive labels in D **then**
    $N(C_+, S_-) = N(C_+, S_-) + 0.01 \times N(C_-, S_+)$
    $N(C_-, S_-) = N(C_+, S_-) - 0.01 \times N(C_-, S_+)$
  else
    $N(C_-, S_+) = N(C_-, S_+) + 0.01 \times N(C_+, S_-)$
    $N(C_+, S_+) = N(C_-, S_+) - 0.01 \times N(C_+, S_-)$
  end if
  Update M using the modified occurrence counts N for C and S
  Calculate *disc*
end while

2 naive bayes models



Latent variable model



1. L is independent from S, i.e., the actual labels are discrimination-free;
2. C is determined by discriminating the L labels using S uniformly at random.

Resultados

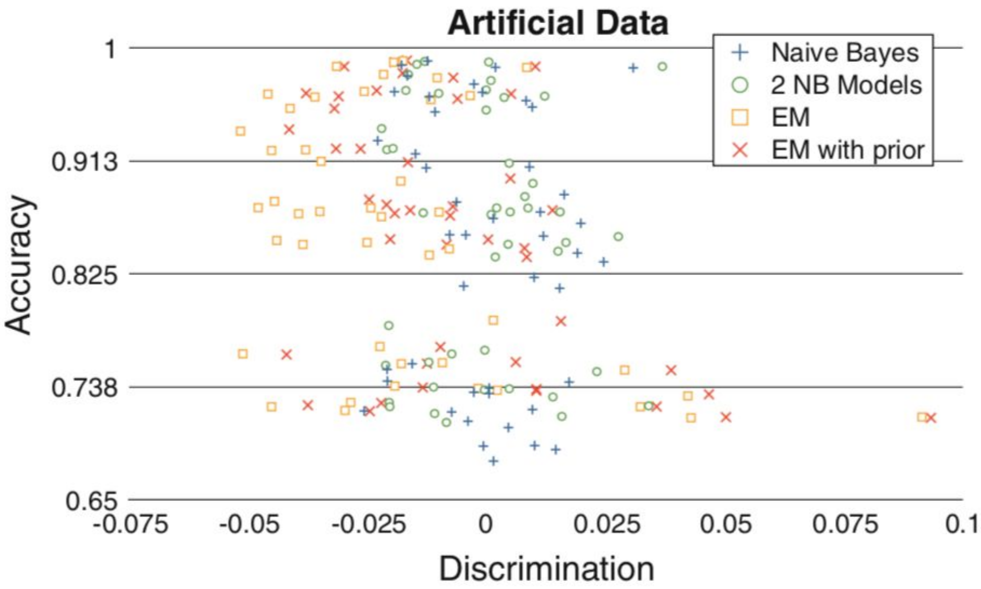


Fig. 2 The resulting discrimination and accuracy values of the trained classifiers on the discrimination-free test-set

Resultados

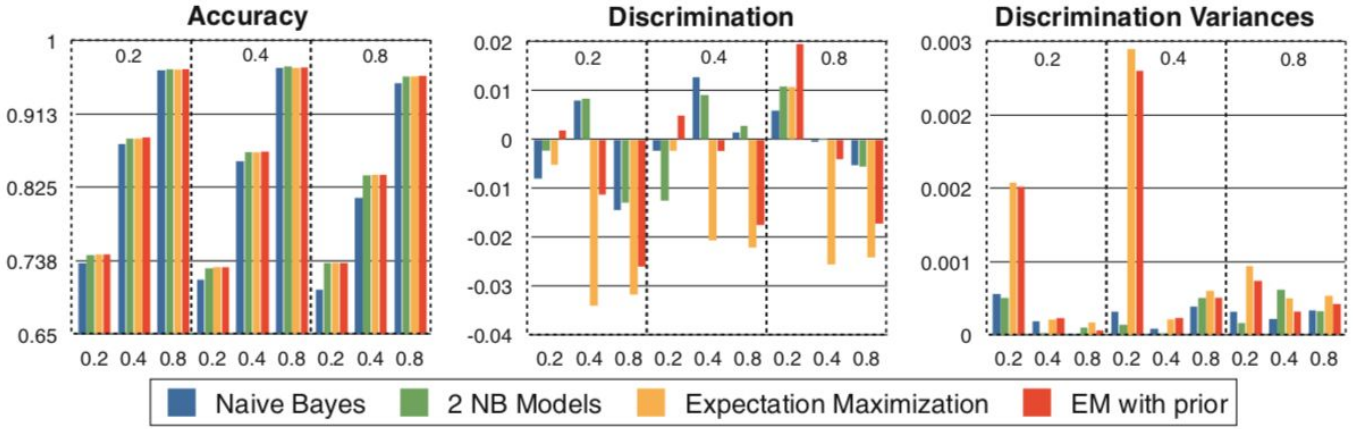


Fig. 3 The results of Fig. 2 (accuracy, discrimination, and discrimination variance) grouped per maximal difference value. The charts show the average values achieved by all methods for all combinations of the maximum bound values 0.2, 0.4, and 0.8. The values on the x-axis are the maximum bounds on $|P(A|L_+) - P(A|L_-)|$, the values in the x-axis boxes (at the top) are the maximum bounds on $|P(A|S_+) - P(A|S_-)|$

Resultados

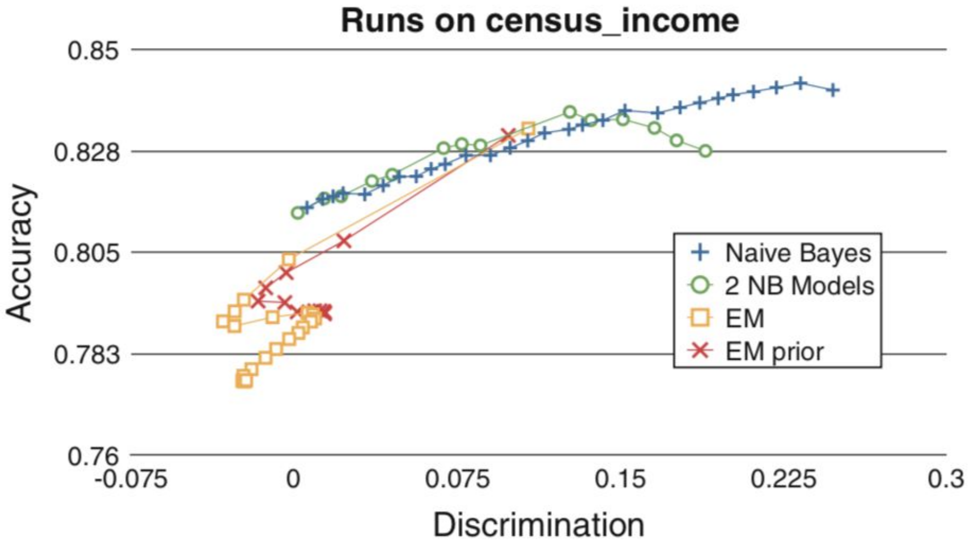
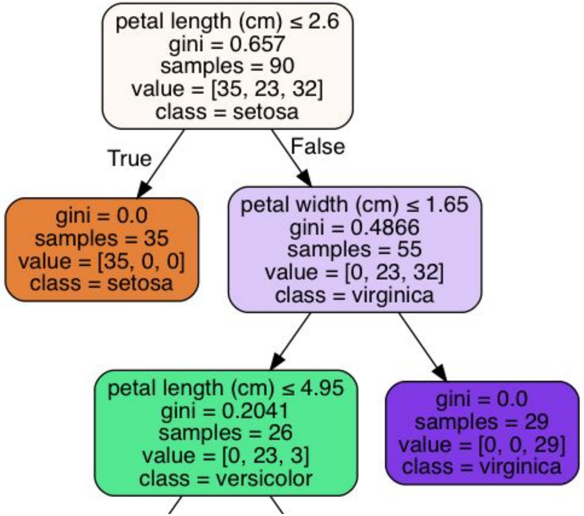


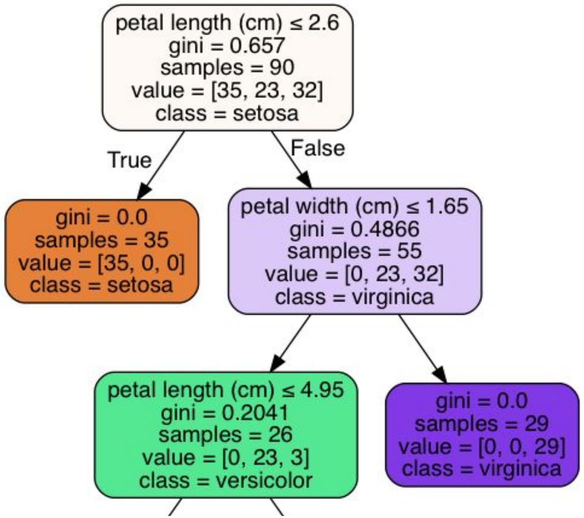
Fig. 4 Lines showing the the consecutive values reached by the runs of each of our algorithms. The accuracy and discrimination values are determined using the data-set

Discrimination-aware tree construction



$$IGC \ := \ H_{Class}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{Class}(D_i)$$

Discrimination-aware tree construction



$$IGC \;:=\; H_{Class}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{Class}(D_i)$$

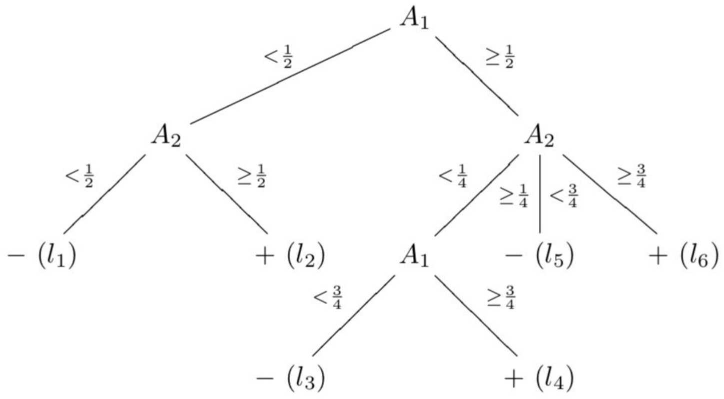
$$IGS \;:=\; H_B(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_B(D_i)$$

Discrimination-aware tree construction

- IGC - IGS
- IGC / IGS
- IGC + IGS

Leaf relabeling

Trocar o label (classe) de um conjunto de folhas visando **diminuir a discriminação** com **menor perda de acurácia**



Leaf relabeling

Problem 2 (RELAB). *Given a decision tree T, a bound $\epsilon \in [0, 1]$, and for every leaf l of T, Δacc_l and $\Delta disc_l$, find a subset L of the set of all leaves $\mathcal{L}$ satisfying*

$$rem\_disc(L) \;:=\; disc_T + \sum_{l \in L} \Delta disc_l \leq \epsilon$$

that minimizes

$$lost\_acc(L) \;:=\; -\sum_{l \in L} \Delta acc_l \;\; .$$

Leaf relabeling

Algorithm 1: *Relabel*

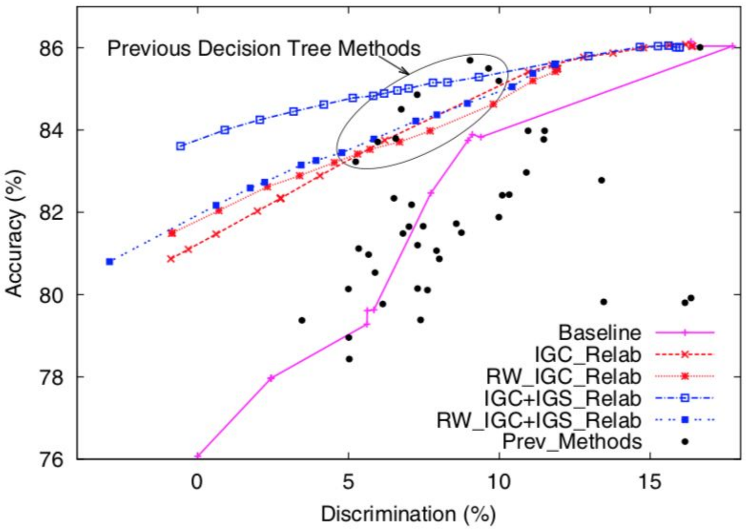
Input Tree T with leaves $\mathcal{L}$, $\Delta acc(l), \Delta disc(l)$ for every $l \in \mathcal{L}, \epsilon \in [0, 1]$

Output Set of leaves L to relabel

```
I := { l ∈ 𝓛 | Δdisc_l < 0 }
L := {}
while rem_disc(L) > ε do
    best_l := arg max_{l ∈ I \ L} (disc_l / acc_l)
    L := L ∪ {l}
end while
return L
```

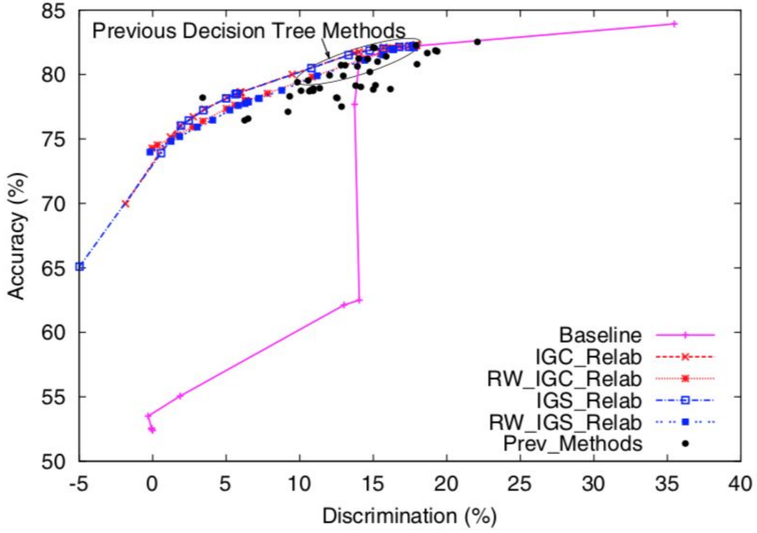
Resultados

(a) Census Income Data



Baseline Disc=19.3 Acc=76.3

(b) Dutch Census 2001 Data



Baseline Disc=29.85 Acc=52.39

Referências

- **Discrimination Aware Decision Tree Learning**.
  Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. In Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM '10). IEEE Computer Society, Washington, DC, USA, 869-874.
  http://dx.doi.org/10.1109/ICDM.2010.50

- **Three naive Bayes approaches for discrimination-free classification.**
  Toon Calders, Sicco Verwer. Data Min Knowl Disc (2010) 21: 277.
  https://doi.org/10.1007/s10618-010-0190-x

Reduzindo discriminação em classificadores

Hilton Pintor (hpbl)
Victor Miranda (vmm)

github.com/if1015-datascience-ufpe/2018-2-ex3-p2p