5243 Project 4 Group 5

A2 & A6

Background and Motivation

- Supervised learning uses historical data to infer a relation between an instance and its label

- Design discrimination free classifires

A2 :Maximizing accu. under fairness constraints

$$\begin{aligned} \text{minimize} \quad & -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} \quad & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \geq -\mathbf{c}, \end{aligned}$$

C-LR: Objective is to minimize the log likelihood subject to cross-covariance between sensitive variables and the distance to the hyperplane. C Controls the trade-offs between accuracy and fairness.

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{b}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\mathbf{b}^T[-1\ \mathbf{x}_i]) \geq 1 - \xi_i, \forall i \in \{1, \ldots, n\} \\ & \xi_i \geq 0, \forall i \in \{1, \ldots, n\}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c}. \end{aligned}$$

C-SVM: Objective is to minimize the margin between support vectors under penalty where b is the weight vectors. It is subject to cross-covariance between sensitive variables and distance to the hyperplane. C Controls the trade-offs between accuracy and fairness.

Results (A2): Logistic Regression

LR

	Classifier	Set	P-rule (%)	Accuracy (%)	Calibration (%)	Protected (%)	Not protected (%)
0	LR	Train	54.662021	66.462384	1.417935	29.752501	54.429933
1	LR	Test	62.270080	65.426881	3.369650	33.611691	53.977273

C-LR

	Classifier	Set	P-rule (%)	Accuracy (%)	Calibration (%)	Protected (%)	Not protected (%)
0	C-LR	Train	99.947862	48.013525	13.261964	99.842022	99.894105
1	C-LR	Test	100.000000	46.661031	10.000415	100.000000	100.000000

Results (A2): Support Vector Machine

SVM

	Classifier	Set	P-rule (%)	Accuracy (%)	Calibration (%)	Protected (%)	Not protected (%)
0	SVM	Train	52.057837	66.208791	0.818440	26.276988	50.476527
1	SVM	Test	63.109384	65.511412	3.227605	31.106472	49.289773

C-SVM

	Classifier	Set	P-rule (%)	Accuracy (%)	Calibration (%)	Protected (%)	Not protected (%)
0	C-SVM	Train	99.930458	47.950127	13.331984	99.789363	99.858807
1	C-SVM	Test	100.000000	46.661031	10.000415	100.000000	100.000000

A6 :Handling Cindutional Discrimination

Goal is to obtain a fair dataset: $P(Y = 1 \mid Race = 1) = P(Y = 1 \mid Race = 0)$

1. Local Massaging
2. Local Preferential Sampling

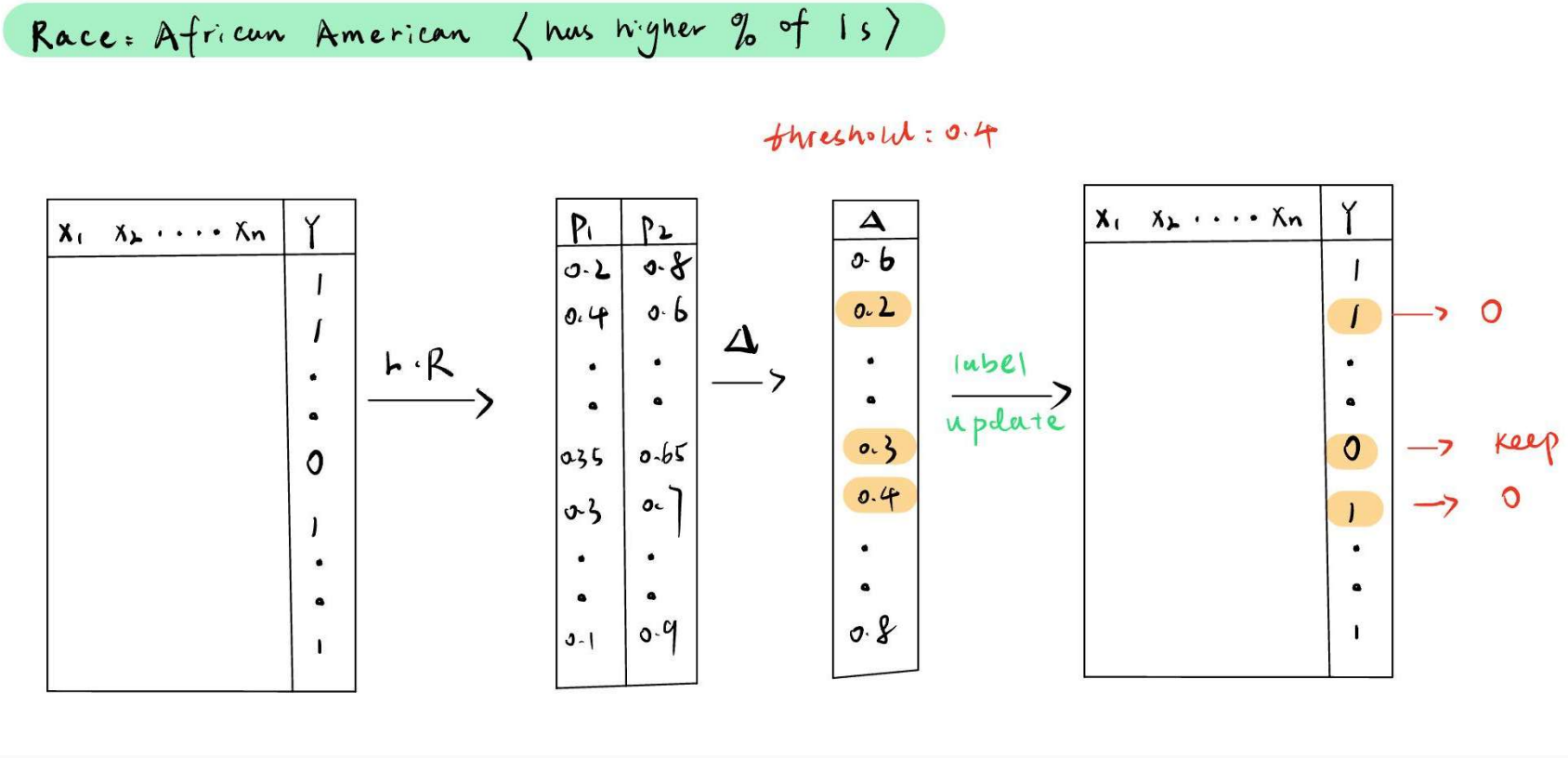
A6.1 Local Massaging

Algorithm 1: Local massaging

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (Algorithm 3);
for *each partition* $X^{(i)}$ **do**
    learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
    rank males using $\mathcal{H}_i$;
    `relabel DELTA` (male) males that are the closest to the decision boundary from $+$ to $-$ (Algorithm 4);
    rank females using $\mathcal{H}_i$;
    `relabel DELTA` (female) females that are the closest to the decision boundary from $-$ to $+$
end

A6.1 Local Massaging Illustration



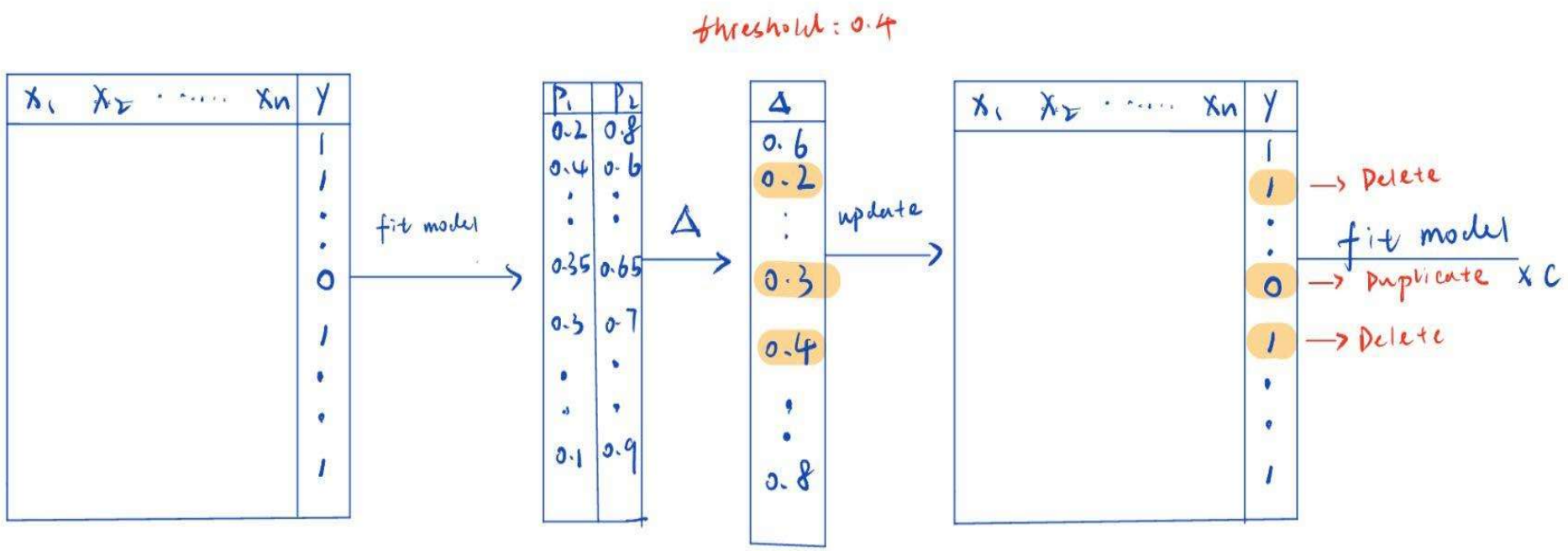
A6.2. Local Preferential sampling

Algorithm 2: Local preferential sampling

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: resampled dataset (a list of instances)

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (see Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  `rank` males using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (male) (see Algorithm 4) males $+$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (male) males $-$ that are the closest to the decision boundary;
  `rank` females using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (female) females $-$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (female) females $+$ that are the closest to the decision boundary;
end

A6.2. Local Preferential sampling



Percentage of Committing a Crime between Races

(Percentage of 1s)	Original	Local massaging	Local preferential sampling
African American	0.51	0.46	0.47
Caucasian	0.39	0.44	0.44

Results comparison in A6

	Local sampling	Local massaging	No techniques
Overall Acc	**0.959**	**0.9496**	0.97156
Acc for African American	**0.961**	**0.95**	0.962
Acc for Caucasian	**0.9572**	**0.9496**	0.988
Difference(calibration)	**0.0035**	**0.001**	0.0256