Maximizing Accuracy and Fairness using Fairness Constraint & Information Theoretic Measures for Fairness-aware Feature Selection

Group 6: Tianxiao He, Linda Lin, Xinming Pan, Namira Suniaprita, Han Wang, YiXun Xu

Presenter: Han Wang

Table of content

Data Preprocessing

Change categorical data to numerical

Encode categorical variables with numerical variables:

- `sex`: 1 for male and 0 for female
- `age_cat`: 2 for > 45, 0 for 25 - 45 and 1 for < 25
- `race`: 1 for caucasian and 0 for african-american
- `c_charge_degree`: 0 for F and 1 for M

```
features = df[['sex', 'age_cat', 'c_charge_degree', 'length_of_stay',"priors_count"]]
sensitive = df['race']
target = df['two_year_recid']
```

- **Protected**: Caucasians (i.e., `race == 1`)
- **Not protected**: African-Americans (i.e., `race == 0`)

Baseline Model

	Classifier	Set	Accuracy (%)	P-rule (%)	Protected (%)	Not protected (%)
0	LR	Train	63.743961	55.108121	21.962896	39.854192
1	LR	Test	62.028169	59.382529	24.186704	40.730337

	Classifier	Set	Accuracy (%)	P-rule (%)	Protected (%)	Not protected (%)
0	SVM	Train	62.850242	51.049930	15.858767	31.065209
1	SVM	Test	61.352113	53.767411	17.821782	33.146067

The p-rule function is commonly used in evaluate fairness in machine learning model, by checking whether the model's positive predictions are distributed similarly across different sensitive groups. The higher the p-rule, the better the fairness.

A2 Algorithm

- minimize the loss function L subject to fairness constraint
- c controls the tradeoff between fairness and accuracy

$$
\begin{aligned}
\min \quad & L(\theta) \\
\text{s.t.} \quad & \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \leq c \\
& \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \geq -c
\end{aligned}
$$

A2 Algorithm: Results

Accuracy drop by 15%

	Classifier	Set	Accuracy (%)	Calibration(%)	P-rule (%)	Protected (%)	Not protected (%)
0	C-SVM	Train	47.098592	14.390507	99.403947	94.733692	95.301742
1	C-SVM	Test	45.386473	11.280586	99.281184	95.190948	95.880150

	Classifier	Set	Accuracy (%)	P-rule (%)	Protected (%)	Not protected (%)
0	C-LR	Train	48.454106	99.403947	94.733692	95.301742
1	C-LR	Test	49.577465	99.281184	95.190948	95.880150

A7 Algorithm (FFS)

Definition 1 (Accuracy coefficient). For a subset of features $X_S \subseteq X^n$, the accuracy coefficient of X_S is given by

$$v^{Acc}(X_S) = I(Y; X_S|\{A, X_{S^c}\}) = UI(Y; X_S \backslash \{A, X_{S^c}\}) + CI(Y; X_S, \{A, X_{S^c}\}). \tag{2.5}$$

Definition 2 (Discrimination coefficient). For a subset of features $X_S \subseteq X^n$, the discrimination coefficient is

$$v^D(X_S) \triangleq SI(Y; X_S, A) \times I(X_S; A) \times I(X_S; A|Y). \tag{2.6}$$

Accuracy coefficient and Discrimination coefficient are two information–theoretic measures that separately quantify the accuracy and discriminatory impact of features. Using Shapley value, we can deduce the marginal impacts of each features. The Shapley value ensures each feature gains as much or more as they would have from acting independently.

A7 Algorithm (FFS): Results

	Feature	Accuracy Coefficient	Discrimination Coefficient
0	sex	0.003568	0.000008
1	age_cat	0.011273	0.000045
2	priors_count	0.024272	0.000045
3	c_charge_degree	0.001959	0.000007
4	length_of_stay	0.005168	0.000011

From the table above, we can see that **Discrimination Coefficient** of `priors_count` is highest, but **Accuracy Coefficient** is also high, so we can't ignore this feature. `length_of_stay` has a little high **Discrimination Coefficient**, and its **Accuracy Coefficient** is low, so we can ignore this feature. So finally, we can choose these four features:

- `priors_count`
- `age_cat`
- `sex`
- `c_charge_degree`

Logistic Using FFS

	Classifier	Set	Accuracy (%)	P-rule (%)	Protected (%)	Not protected (%)
0	FFS-LR	Train	63.405797	58.808292	23.937762	40.704739
1	FFS-LR	Test	61.802817	65.541470	26.449788	40.355805

SVM Using FFS

	Classifier	Set	Accuracy (%)	P-rule (%)	Protected (%)	Not protected (%)
0	SVM	Train	62.512077	51.714542	15.918612	30.781693
1	SVM	Test	61.183099	57.733015	18.811881	32.584270

P-rule increases by 5%, accuracy stays the same