Double logistic regression approach to biased positive-unlabeled data

Jan Mielniczuk

- Institute of Computer Science, Polish Academy of Sciences
- Faculty of Mathematics and Information Sciences, Warsaw University of Technology

ML in PL'2023

Based on joint research with **K. Furmańczyk, P. Teisseyre and W. Rejchel**

Classification of Positive Unlabeled (PU) Data

Partially observable variant of classical classification model when $P_{X,Y}$, $X \in \mathcal{X}$ and $Y \in \{0, 1\}$ are observable.
Now, only $P_{X,S}$, $S \in \{0, 1\}$ will be observable, where $S = Y$ with a certain probability.
Assumption:

$$P(S = 1|X, Y = 0) = 0$$

Thus

$$Y = 0 \quad \Rightarrow \quad S = 0$$

$$Y = 1 \quad \Rightarrow \quad \text{S=0 or S=1}$$

Many instances of practical situations when such situation occurs ...

Citizens' assembly example

Random pick is performed for climate change citizens' assembly which would prepare suggestions for a government on this issue. One can opt in or opt out of becoming a member.

- $Y = 1$: climate change believer, $Y = 0$: climate change denialist;
- $S = 1$: one opts in, $S = 0$: one opts out;

We have

$$S = 1 \quad \Rightarrow \quad Y = 1$$

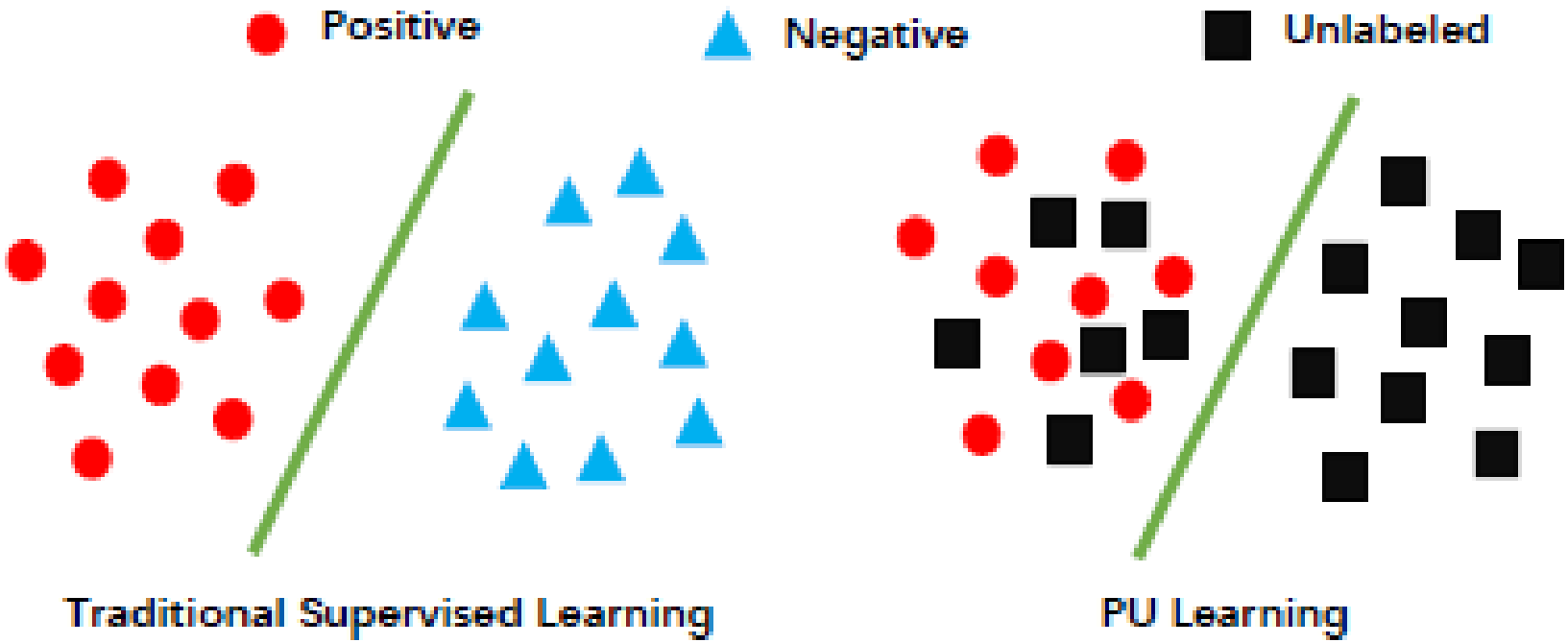
however

$$S = 0 \quad \Rightarrow \quad \mathrm{Y{=}0 \text{ or } S{=}0}$$

Many other examples in medicine, surveys, NLP, recomendation systems, ecology etc;

Supervised classification vs PU classification[a]

[a] Gong et al (2021)



Posterior probability and propensity score

Posterior probability of Y (unobservable)

$$y(x) = P(Y = 1|X = x)$$

Propensity score (unobservable)

$$e(x) = P(S = 1|Y = 1, X = x)$$

Posterior probability of S (observable)

$$s(x) = P(S = 1|X = x)$$

Selection biased PU data

We have

$$s(x) = y(x)e(x)$$

Objective: to model situations when $e(x) \not\equiv C$
(**Selection Bias**, **S**elected **C**ompletely **A**t **R**andom (**SCAR**) assumption does not necessarily hold)

Joint modelling of $y(x)$ and $e(x)$

Estimation of $s(x)$ does not yield direct conclusions about $y(x)$.
Can we draw conclusions about $y(x)$ & $e(x)$ simultaneously ?

$$s(x) = y(x)e(x) = e(x)y(x)$$

Obvious problem: $y(x)$ and $s(x)$ may be swapped ...
Is it possible to identify $y(x)$ and $e(x)$ up to a swap ?

Joint modelling of $y(x)$ and $e(x)$

Estimation of $s(x)$ does not yield direct conclusions about $y(x)$.
Can we draw conclusions about $y(x)$ & $e(x)$ simultaneously ?

$$s(x) = y(x)e(x) = e(x)y(x)$$

Obvious problem: $y(x)$ and $s(x)$ may be swapped ...
Is it possible to identify $y(x)$ and $e(x)$ up to a swap ?
The answer is yes, at least in some parametric models.

Double logistic model: identifiability

Let $\sigma(s) = 1/(1 + e^{-s})$ is a logistic function and assume that both $y(x)$ and $e(x)$ are governed by logistic model

$$y(x) = \sigma(\beta_0^* + \beta^{*T} x) \qquad e(x) = \sigma(\gamma_0^* + \gamma^{*T} x) \quad (*)$$

$\tilde{\beta} = (\beta_0, \beta^T)^T$....

Double logistic model: identifiability

Let $\sigma(s) = 1/(1 + e^{-s})$ is a logistic function and assume that both $y(x)$ and $e(x)$ are governed by logistic model

$$y(x) = \sigma(\beta_0^* + \beta^{*T} x) \qquad e(x) = \sigma(\gamma_0^* + \gamma^{*T} x) \quad (*)$$

$\tilde{\beta} = (\beta_0, \beta^T)^T$....

Theorem

In double logistic model parameters $\tilde{\beta}^*$ and $\tilde{\gamma}^*$ are uniquely defined up to interchange of $y(x)$ and $e(x)$ i.e. if for some $\tilde{\beta}$ and $\tilde{\gamma}$, $s(x) = \sigma(\beta_0 + \beta^T x)\sigma(\gamma_0 + \gamma^T x)$ for all $x \in R^p$, then $(\tilde{\beta}, \tilde{\gamma}) = (\tilde{\beta}^*, \tilde{\gamma}^*)$ or $(\tilde{\beta}, \tilde{\gamma}) = (\tilde{\gamma}^*, \tilde{\beta}^*)$.

Expected loglikelihood: JOINT method

Logistic model is fitted for both $y(x)$ and $e(x)$. Expected log-likelihood:

$$Q(\tilde{\beta}, \tilde{\gamma}) = \tag{1}$$
$$E_{X,s}[S \log s_{\tilde{\beta},\tilde{\gamma}}(X) + (1 - S) \log(1 - s_{\tilde{\beta},\tilde{\gamma}}(X))],$$

where $s_{\tilde{\beta},\tilde{\gamma}}(x) = \sigma(\beta_0 + \beta^T x)\sigma(\gamma_0 + \gamma^T x)$.
$Q(\tilde{\beta}, \tilde{\gamma})$ is *not* concave function.

Theorem

Let assumptions of Theorem 1 hold and $|\tilde{\beta}^*|_1 > |\tilde{\gamma}^*|_1$. *Then (i)*

$$(\tilde{\beta}^{*T}, \tilde{\gamma}^{*T})^T = \arg \max_{(\tilde{\beta},\tilde{\gamma}):|\tilde{\beta}|_1 > |\tilde{\gamma}|_1} Q(\tilde{\beta}, \tilde{\gamma})$$

and $(\tilde{\beta}^{*T}, \tilde{\gamma}^{*T})^T$ *is the only maximiser of* $Q(\tilde{\beta}, \tilde{\gamma})$.

This leads to JOINT method. Maximiser of $\hat{Q}(\tilde{\beta}, \tilde{\gamma})$ is proved to be consistent.

First proposal: alternate maximisation of two Fisher-consistent score functions for $P_{X,Y}$

Let $y(x, \tilde{\beta}) = \sigma(\beta_0 + \beta^T x)$ and $e(x, \tilde{\gamma}) = \sigma(\gamma_0 + \gamma^T x)$.
We will look for solutions of empirical counterparts of two concave optimisation problems solved alternately.

The first one based on:

$$W(x,\tilde{\beta}) = y(x,\tilde{\beta}^*)\log y(x,\tilde{\beta}) + (1-y(x,\tilde{\beta}^*))\log(1-y(x,\tilde{\beta}))$$

we have

$$\mathrm{argmax}_{\tilde{\beta}} E_X W(X,\tilde{\beta}) = \tilde{\beta}^*$$

and

$$W(x,\tilde{\beta}) = E_{S|X=x}(w_1(S,x)\log y(x,\tilde{\beta}) + w_0(S,x)\log(1-y(x,\tilde{\beta}))$$

for

$$w_1(s,x) = I\{S=1\} + I\{S=0\} \times P(Y=1|S=0,X=x) =$$

$$= I\{S=1\} + I\{S=0\}\frac{(1-e(x))}{e(x)}/\frac{(1-s(x))}{s(x)}$$

$$= I\{S=1\} + I\{S=0\}OR(x)$$

and $w_0(s,x) = P(Y=0|S=0,X=x)$.

Second equation

Analogously

$$\operatorname{argmax}_{\tilde{\gamma}} E_{X|Y=1}[e(X,\tilde{\gamma}^*)\log e(X,\tilde{\gamma}) + (1-e(X,\tilde{\gamma}^*))\log(1-e(X,\tilde{\gamma}))]$$

$$= \tilde{\gamma}^*$$

But $E_{S|Y=1,X}S = e(X,\tilde{\gamma}^*)$

$$E_{X|Y=1}[e(X,\tilde{\gamma}^*)\log e(X,\tilde{\gamma}) + (1-e(X,\tilde{\gamma}^*))\log(1-e(X,\tilde{\gamma}))] =$$

$$= E_{S,X|Y=1}K(S,X,\tilde{\gamma})$$

where $K(S,X,\tilde{\gamma}) = S\log e(X\tilde{\gamma}) + (1-S)\log(1-e(X,\tilde{\gamma}))$.

The empirical counterpart of $E_X W(X, \tilde{\beta})$:

$$W_n(\tilde{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \Big\{ \hat{w}_1(S_i, X_i) \log y(X_i, \tilde{\beta}) + \hat{w}_0(S_i, X_i) \log(1 - y(X_i, \tilde{\beta})) \Big\}, \tag{2}$$

and counterpart of $E_{S,X|Y=1} K(S, X, \tilde{\gamma})$ as

$$K_n(\tilde{\gamma}) = \sum_{i=1}^{n} I\{i \in \hat{\mathcal{P}}\} \big[S_i \log e(X_i, \tilde{\gamma}) + (1 - S_i) \log(1 - e(X_i, \gamma)) \big],$$

where $\mathcal{P} = \{i : Y_i = 1\}$, $\hat{\mathcal{P}} = \{i : \widehat{Y_i = 1}\}$.

Two Models/Equations algorithm

We repeat the following two steps until convergence:

1. **Model 1.** Solve $\hat{\beta}_n = \arg\max_{\tilde{\beta}} W_n(\tilde{\beta})$.
2. Calculate $\hat{y}(X_i) = y(X_i, \hat{\beta}_n)$.
3. **Model 2.** Solve $\hat{\gamma}_n = \arg\max_{\gamma} \hat{R}_n(\tilde{\gamma})$, where

$$\hat{K}_n(\tilde{\gamma}) = \sum_{i=1}^{n} I(i \in \hat{\mathcal{P}}) K(S_i, X_i, \tilde{\gamma}),$$

4. Calculate $\hat{e}(X_i) = e(X_i, \hat{\gamma})$.
5. Update $\hat{s}(X_i) = \hat{e}(X_i)\hat{y}(X_i)$ and

$$\widehat{OR}(X_i) = \frac{1 - \hat{e}(X_i)}{\hat{e}(X_i)} \Big/ \frac{(1 - \hat{s}(X_i)}{\hat{s}(X_i)}$$

.

Estimation of $\mathcal{P} = \{i : Y_i = 1\}$

$$\hat{\mathcal{P}} = \{S_i = 1 \text{ or } \hat{y}(x_i) > t\},$$

where t is empirical quantile of order α of the set

$$\{\hat{y}(X_i), S_i = 1\},$$

$$\alpha = \hat{P}(S = 1)$$

Second proposal: JOINT method

$(X_i, Y_i, S_i), i = 1, \ldots, n$ iid sample drawn from $P_{(X,Y,S)}$.
Observed data $(X_i, S_i), i = 1, \ldots, n$ (single training sample scenario).
Empirical likelihood for (X_i, S_i):

$$Q_n(\tilde{\beta}, \tilde{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} S_i \log s_{\tilde{\beta},\tilde{\gamma}}(X_i) + (1 - S_i) \log(1 - s_{\tilde{\beta},\tilde{\gamma}}(X_i))$$

JOINT estimators: maximisers of $Q_n(\cdot, \cdot)$.

Regret of JOINT estimator

For any $r > 0$ and $\theta := (\beta, \gamma), \theta^* := (\beta^*, \gamma^*)$

$$\hat{\theta} = \arg\min_{\theta:|\theta-\theta^*|\leq r} Q_n(\theta)$$

Theorem

If X_i are subgaussian with parameter μ then for any $s \in (0,1)$ we have

$$P(\underbrace{Q\left(\hat{\theta}\right) - Q(\theta^*)}_{\text{regret of } \hat{\theta}} \leq \frac{32\mu r}{s}\sqrt{\frac{\log p}{n}}) \geq 1 - s$$

Labelling scenarios

1. **Scenario 1.** We consider constant propensity score function $e(x) = P(S = 1|Y = 1, X = x) = c$, where c is label frequency which varies in simulations.
2. **Scenario 3.** Propensity score function is defined as $e(x) = \prod_{j=1}^{k} [sc(x(j), p^-, p^+)]^{1/k}$, where $x(j)$ is j-th coordinate of x and $sc(x(j), p^-, p^+) := p^- + \frac{x(j) - \min x(j)}{\max x(j) - \min x(j)}(p^+ - p^-)$.

8 data sets from UCI repository artificially labeled using scenarios above.

Methods considered

1. JOINT
2. Two Models (TM)
3. TM SIMPLE
   $\hat{e}(x) = \hat{e}_{naive}(x) = (1 + \hat{s}(x))/2$, $\hat{y}(x)$ from $\operatorname{argmax}_{\tilde{\beta}} W_n(\tilde{\beta})$.
4. SAR-EM (Bekker Davis (2019))
5. LBE (Gong et al (2021))
6. Oracle (Y known)
7. NAIVE
   $\hat{y}(x)$ based on maximisation loglikelihood for (X_i, S_i)

Table: Accuracy[1] for scenario 3, $k = 5$, $p^- = 0.2$ and $p^+ = 0.6$

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**
Artif1	0.664	0.777	0.830	**0.857**	0.816
Artif2	0.643	0.743	0.780	**0.805**	0.762
diabetes	0.682	0.726	**0.732**	0.714	0.710
BCancer	0.814	0.877	0.903	**0.908**	0.905
heart-c	0.636	0.654	0.636	**0.679**	0.622
credit-a	0.624	0.688	0.691	**0.765**	0.676
adult	0.767	0.795	**0.828**	0.778	0.809
vote	0.735	0.750	0.757	**0.837**	0.733
wdbc	0.766	0.825	0.838	**0.852**	0.806
spambase	0.633	0.648	0.690	**0.810**	0.775
avg. rank	5.8	4.2	3.1	**2.5**	4.4

[1] Accuracy of classifer based on $\hat{y}(x)$

Table: Approximation error[2] for scenario 3, $k = 5$, $p^- = 0.2$ and $p^+ = 0.6$.

	NAIVE	**TM simple**	**EM**	**TM**	**JOINT**
Artif1	0.292	0.227	0.147	**0.118**	0.159
Artif2	0.300	0.243	0.177	**0.139**	0.194
diabetes	0.206	**0.128**	0.144	0.162	0.171
BCancer	0.180	0.139	0.093	**0.083**	0.098
heart-c	0.270	0.222	0.249	**0.219**	0.262
credit-a	0.287	0.218	0.229	**0.173**	0.283
adult	0.150	0.093	**0.071**	0.134	0.091
vote	0.272	0.228	0.197	**0.148**	0.275
wdbc	0.219	0.186	0.147	**0.117**	0.197
spambase	0.308	0.250	0.237	**0.088**	0.194
avg. rank	5.9	4.0	3.2	**2.5**	4.4

[2] $n^{-1} \sum_{i=1}^{n} |\hat{y}(X_i) - \hat{y}_{oracle}(X_i)|$

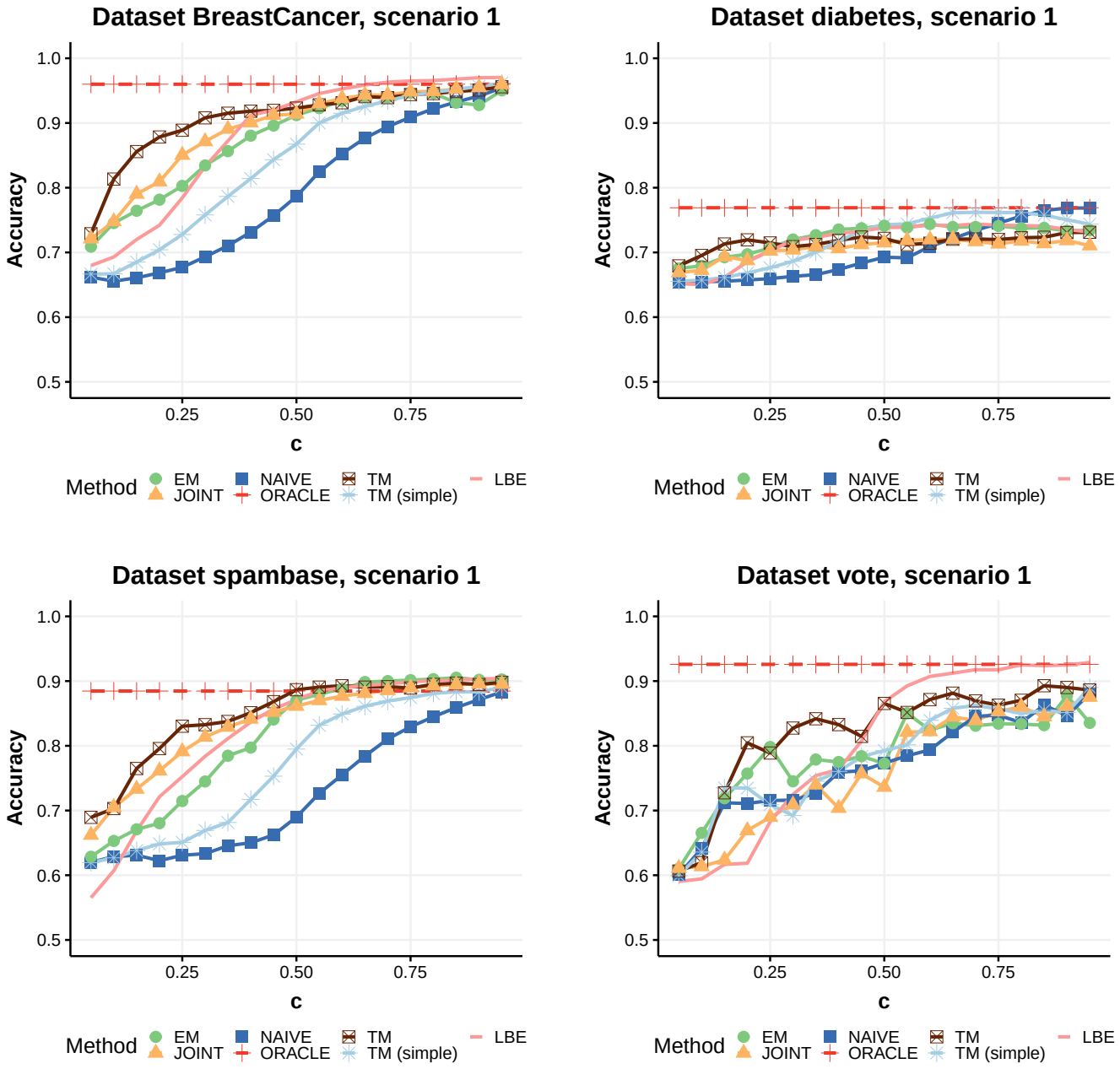


Figure: Accuracy for benchmark datasets for scenario 1 and different values of

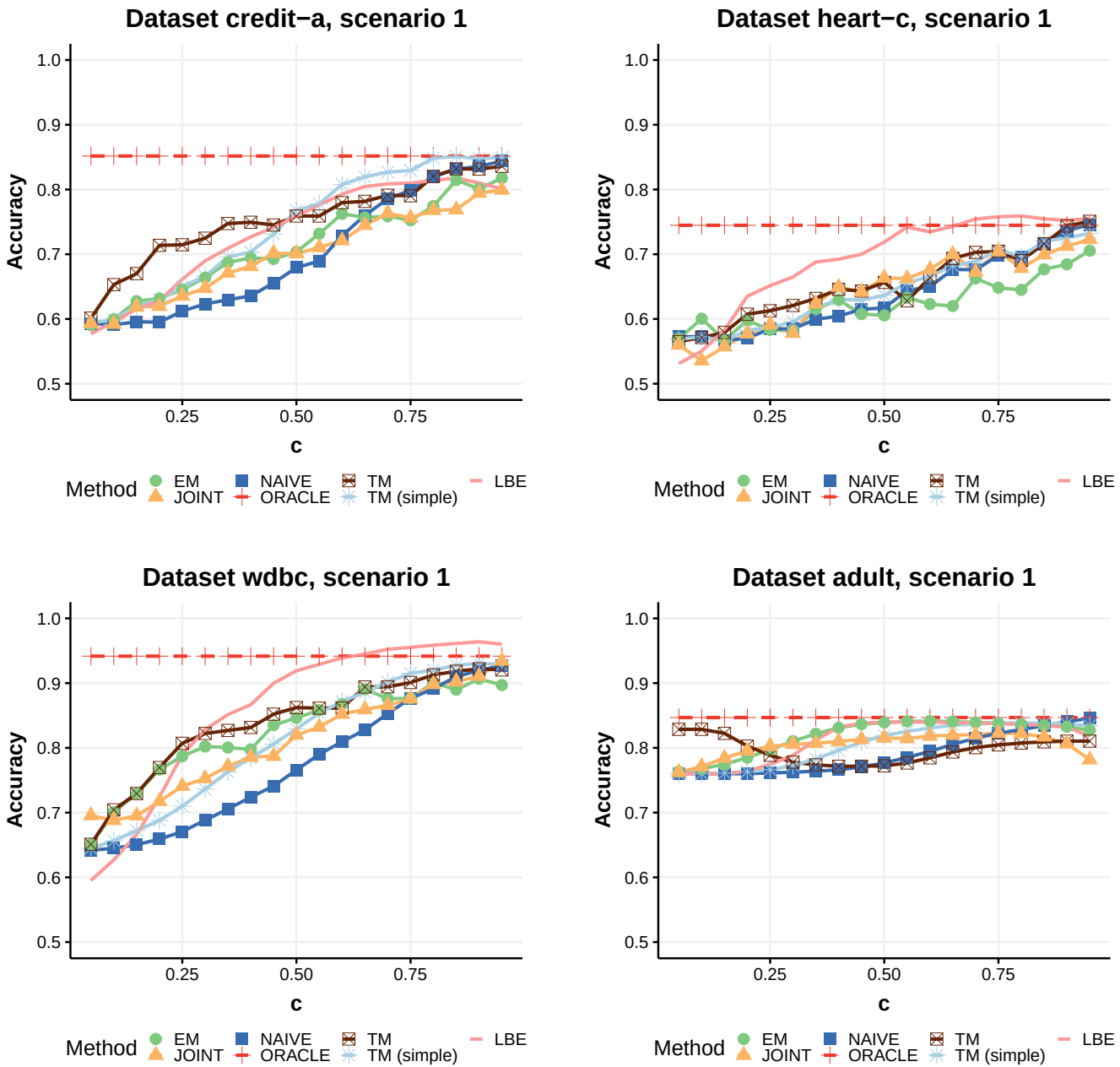


Figure: Accuracy for benchmark datasets for scenario 1 and different values of

Issues

- Estimation of the set $\mathcal{P} = \{i : Y_i = 1\}$ crucial for performance of TM algorithm
- Feature selection under SCAR/ for an arbitrary propensity score ?
- Testing SCAR assumption $H_0 : \tilde{\gamma} = (\gamma_0, 0^T)^T$
- Under-performance of JOINT method (unlike in SCAR scenario) is due to maximisation issues (MM algorithm applied at present)

References

- J. Bekker, J. Davis, Beyond the selected completely at random assumption for learning from positive and unlabeled data, ECML'19 **SAR-EM**
- C. Gong et al, Instance-Dependent Positive and Unlabeled Learning with Labeling Bias Estimation, IEEE PAMI, 2022 **LBE**
- K. Furmańczyk,JM, P. Teisseyre, W. Rejchel, Double logistic regression approach to biased positive-unlabeled data, ECAI'2023 **TWO MODELS, JOINT**
- A. Wawrzeńczyk, JM, One-class classification approach to variational learning from biased positive unlabeled data, ECAI'2023 **VAE-PU-OCC**
- M. Platek, JM, Enhancing naive classifier positive unlabeled data based on logistic regression approach, FedCsis'2023 **Enhanced-Naive**