Positve Unlabelled data -different scenarios

Jan Mielniczuk

(based joint work with Paweł Teisseyre and Małgorzata Łazęcka)

Traditional binary classification

X_1	X_2	...	X_p	Y
1.0	2.2	...	4.2	1
2.4	1.3	...	3.1	1
0.9	1.4	...	3.2	1
0.6	1.2	...	3.2	1
1.2	3.5	...	7.2	0
1.7	3.2	...	3.2	0
...	...	...	...	...

- Y- target variable.
- $X = (X_1, \ldots, X_p)^T$- vector of features.

TASK: Model the relationship between Y and X.

Positive and unlabelled data

X_1	X_2	...	X_p	Y	S
1.0	2.2	...	4.2	1	1
2.4	1.3	...	3.1	1	1
0.9	1.4	...	3.2	1	?
0.6	1.2	...	3.2	1	?
1.2	3.5	...	7.2	0	?
1.7	3.2	...	3.2	0	?
...	...	...	...	...	...

- Y- TRUE target variable (**NOT OBSERVED DIRECTLY**)
- S- SURROGATE target variable (**OBSERVED**).
- $X = (X_1, \ldots, X_p)^T$- vector of explanatory variables (features).

TASK: Model the relationship between Y and X USING ONLY S and X.

Positive and unlabelled data

X_1	X_2	...	X_p	Y	S
1.0	2.2	...	4.2	1	1
2.4	1.3	...	3.1	1	1
0.9	1.4	...	3.2	1	0
0.6	1.2	...	3.2	1	0
1.2	3.5	...	7.2	0	0
1.7	3.2	...	3.2	0	0
...	...	...	...	...	...

Surrogate variable S:

- $S = 1$ (observation is labelled); $S = 0$ (observation is unlabelled)
- $S = 1 \implies Y = 1$ (labelled examples are positive)
- For $S = 0$, the example can be either positive ($Y = 1$) or negative ($Y = 0$)

Positive Unlabeled (PU): two scenarios

- Single sample scenario ('single sample') $PU - ss$
- Case - control scenario $PU - cc$

Single sample scenario:
Distribution $P_{X,Y,S}$ such that

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

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

We have

$$S \perp X|Y$$

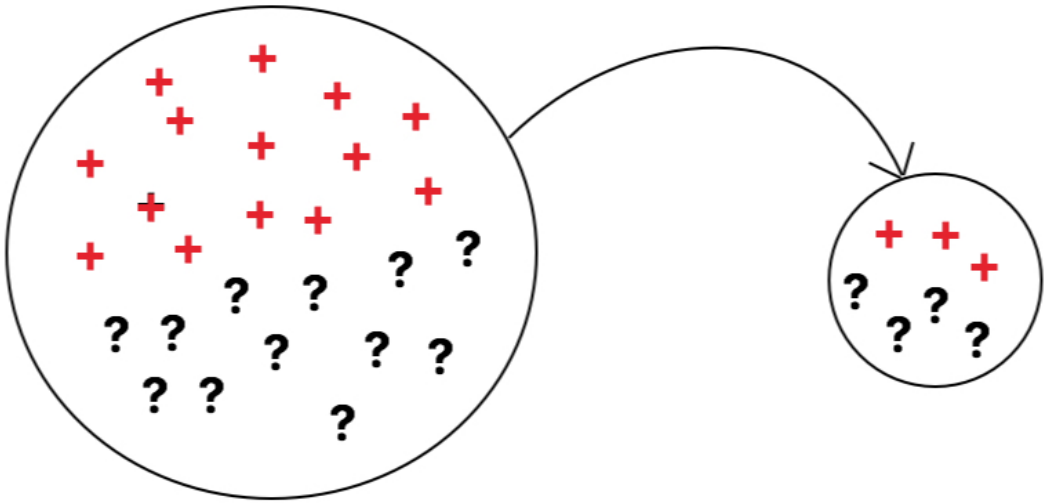
Selected Completely at Random (SCAR). Sample $(X_i, S_i), i = 1, \ldots, n$ from $P_{X,S}$.

$$n_l = \#\{i : S_i = 1\} \qquad n_u = \#\{i : S_i = 0\}$$

n_l, n_u- random variables

Single training data scenario

- We assume that there is some unknown distribution $P(Y, X, S)$ such that $(y_i, x_i, s_i), i = 1, \ldots, n$ is iid sample drawn from it.
- Only data $(x_i, s_i), i = 1, \ldots, n$, is observed.
- Distribution of X is a mixture of distributions $X|S = 1$ and $X|S = 0$.



Positive and unlabelled data: single sample scenario

Example (survey: under reporting)

Sensitive question concerning e.g. smoking during pregnancy

True ($Y = 1$ smoking; $Y = 0$ no smoking)

Answer ($S = 1$ admitting smoking; $S = 0$ not admitting smoking)

We can define 3 groups:

1. Women admitting smoking ($Y = 1$ and $S = 1$)
2. Women not admitting smoking who smoked ($Y = 1$ and $S = 0$)
3. Women not admitting smoking who really did not smoke ($Y = 0$ and $S = 0$)

Case control-scenario: cc

- Fix n_l i n_u;
- We sample n_l observations from $P_{X|Y=1}$ and n_u observations from P_X.
- Most PU data relate to this scenario.

Can we build a classifier based on such data ?
Naive classifier treats all unlabelled data ($S = 0$) as $Y = 0$
-heavily biased

Positive and unlabelled data: c-c scenario

Example (medicine: undiagnosed diseases)

Occurrence of disease ($Y = 1$ disease; $Y = 0$ no disease)

Diagnosis of disease ($S = 1$ diagnosed disease; $S = 0$ undiagnosed disease)

Two data bases available: one for patients with diagnosed disease, second for a general population (healthy and ill).
We sample n_l observations from the first base and n_u observations from the second.

Positive and unlabelled data: c-c scenario

Example II (ecology: predicting occurrence of the species (habitat determination))

Data consist of a sample of locations with observed presences and a separate group of locations sampled from the full landscape, with unknown presences.

Occurrence of the species ($Y = 1$ present ; $Y = 0$ absent)
Reported occurrence ($S = 1$ reported presence; $S = 0$ not reported)

We can define 3 groups:

1. Reported occurence of the species ($Y = 1$ and $S = 1$)
2. Occurrence of species not reported ($Y = 1$ and $S = 0$)
3. No species ($Y = 0$ and $S = 0$)

PU learning- basics

Two important quantities:

- Label frequency $c := P(S = 1|Y = 1)$
- Propensity score $e(x) := P(S = 1|Y = 1, x)$

Fact 1

$P(X|S = 1) = \frac{e(x)}{c} P(X|Y = 1)$.

Proof. From definition of PU and Bayes Theorem we have:

$$P(X|S = 1) = P(X|S = 1, Y = 1) = \frac{P(S = 1|X, Y = 1)}{P(S = 1|Y = 1)} P(X|Y = 1).$$

For SCAR : $P(X|S = 1) = P(X|Y = 1)$.

PU learning- basics

Two important quantities:

- Label frequency $c := P(S = 1|Y = 1)$
- Propensity score $e(x) := P(S = 1|Y = 1, x)$

Fact 2 (Relationship between label frequency and class prior)

$$c = P(S = 1|Y = 1) = \frac{P(S = 1, Y = 1)}{P(Y = 1)} = \frac{P(S = 1)}{P(Y = 1)}.$$

$P(S = 1)$ is easily estimated from data as a fraction of labeled examples among all examples.

PU learning- basics

Two important quantities:

- Label frequency $c := P(S = 1|Y = 1)$
- Propensity score $e(x) := P(S = 1|Y = 1, x)$

Fact 3 (Relationship between posterior probabilities)

$$P(S = 1|X) = e(X)P(Y = 1|X).$$

Proof. From Law of Total Probability and definition of PU:

$$\begin{aligned} &P(S = 1|X) \\ &= P(S = 1|X, Y = 1)P(Y = 1|X) + P(S = 1|X, Y = 0)P(Y = 0|X) \\ &= P(S = 1|X, Y = 1)P(Y = 1|X). \end{aligned}$$

For SCAR

$$P(S = 1|X) = cP(Y = 1|X)$$

Prospective and i retrospective sampling

S-S and C-C scenarios are related to ..

- Prospective sampling: we sample n observations from P_{XY} $(Y \in \{0, 1\})$

$$n_1 = \#\{i : Y_i = 1\} \qquad n_0 = \#\{i : Y_i = 0\}.$$

  Ineffective when $\pi = P(Y = 1)$ - small

- retrospective sampling: we sample n_1 observations from $P_{X|Y=1}$ and n_0 observations from $P_{X|Y=0}$. We have control over n_1 and n_0 but not over π.

Identifiability of parameter β in retrospective sampling

Formalising retrospective sampling: W- variable indicating inclusion in the sample

$$P(W = 1|X, Y = 1) = p_1 \qquad P(W = 1|X, Y = 0) = p_0$$

Suppose that

$$\log\left(\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)}\right) = \beta' x$$

$$\begin{aligned} & \operatorname{logit} P(Y = 1|X, W = 1) = \log \frac{P(Y = 1|X, W = 1)}{P(Y = 0|X, W = 1)} \\ = \quad & \log\left(\frac{P(W = 1|X, Y = 1)}{P(W = 1|X, Y = 0)} \times \frac{P(Y = 1|X)}{P(Y = 0|X)}\right) \\ = \quad & \log\left(\frac{p_1}{p_0}\right) + \beta' x \end{aligned} \tag{1}$$

This holds for logistic regression model only!

Two algorithms: Expectation-Maximisation (EM) and Minorisation-Maximisation (MM)

- EM: popular when certain variables are not observed;
- MM: used when landscape of likelihood is complicated;
- EM concerns unobserved likelihood for $(X_i, Y_i), i = 1, \ldots, n$, MM algorithm concerns observed likelihood

Unobserved likelihood function for $(X_i, Y_i), i = 1, \ldots, n$

Let (X_i, Y_i, W_i) be a sample from $P_{X,Y,W}$ as above

$$p_1 = P(W = 1|X, Y = 1) = \frac{n_l + \pi n_u}{n\pi}$$

$$p_0 = P(W = 1|X, Y = 0) = \frac{n_u(1-\pi)}{n(1-\pi)} = \frac{n_u}{n}.$$

Then

$$P(Y_1, \ldots, Y_n|X_1, \ldots, X_n, W_1 = 1, \ldots, W_n = 1) =$$
$$= \prod_{i=1}^{n} \left(\frac{e^{\eta^*(X_i)}}{1 + e^{\eta^*(X_i)}}\right)^{Y_i} \left(\frac{1}{1 + e^{\eta^*(X_i)}}\right)^{1-Y_i},$$

$$\eta^*(X_i) = \beta' X_i + \log \frac{n_l + \pi n_u}{\pi n_u}$$

Algorithm EM, Ward et al 2009

Based on *unobservable* likelihood function

$$P(Y_1, \ldots, Y_n | X_1, \ldots, X_n, W_1 = 1, \ldots, W_n = 1) =$$
$$= \prod_{i=1}^{n} \left(\frac{e^{\eta^*(X_i)}}{1 + e^{\eta^*(X_i)}} \right)^{Y_i} \left(\frac{1}{1 + e^{\eta^*(X_i)}} \right)^{1-Y_i},$$

Assumption: π is known.

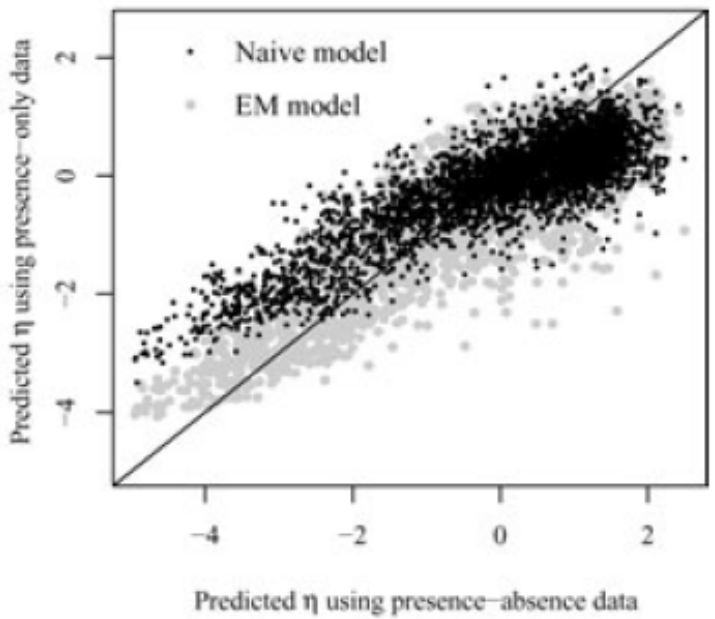
$$\eta^*(X_i) = \beta' X_i + \log \frac{n_l + \pi n_u}{\pi n_u}$$

Algorihtm EM cont'd

Assume that π is known.

- $\hat{y}_i^{(0)} = \pi$ dla $s_i = 0$
- Step M:
  Calculate $\hat{\eta}_i^{*(k)}$ fitting $\hat{y}_i^{(k-1)} \sim x_i$ (logistic model);
- Correction of an intercept: $\hat{\eta}_i^{(k)} := \hat{\eta}_i^{*(k)} - \log \frac{n_l + \pi n_u}{\pi n_u}$
  (modification related to cc)
- Step E:
  $\hat{y}_i^{(k)} := \frac{e^{\eta_i^{(k)}}}{1+e^{\eta_i^{(k)}}}$ for $s_i = 0$ i $\hat{y}_i^{(k)} = 1$ for $s_i = 1$.

Occurrence of an eel *Anguilla diefenbachii* in New Zeland.
$\pi = 0.513$. PU-cc data: sampling form data base of occurrences and data base of all habitats.
(less shrinkage for EM than for naive estimator (based on logistic model fitted to (X_i, S_i)).



PU-cc: Algorithm MM for *observed* likelihood function

$$\tilde{\beta}_0 = \beta_0 + \log \frac{n_p + \pi n_u}{\pi n_u} \quad \tilde{\beta} = (\tilde{\beta}_0, \beta_1, \ldots, \beta_p)$$

$$c = \frac{n_l}{n_l + \pi n_u}$$

$$L_c(\tilde{\beta}) = \prod_{i=1}^{n} \left(\frac{c e^{\tilde{\beta}' x_i}}{1 + e^{\tilde{\beta}' x_i}} \right)^{S_i} \left(1 - \frac{c e^{\tilde{\beta}' x_i}}{1 + e^{\tilde{\beta}' x_i}} \right)^{1 - S_i}$$

Is **not** a concave function of $\tilde{\beta}$.
Concave majorisation $L_c(\tilde{\beta})$ and MM algorithm
Final modification of $\tilde{\beta}_0$. Comparison with EM ??.

MM algorihtm

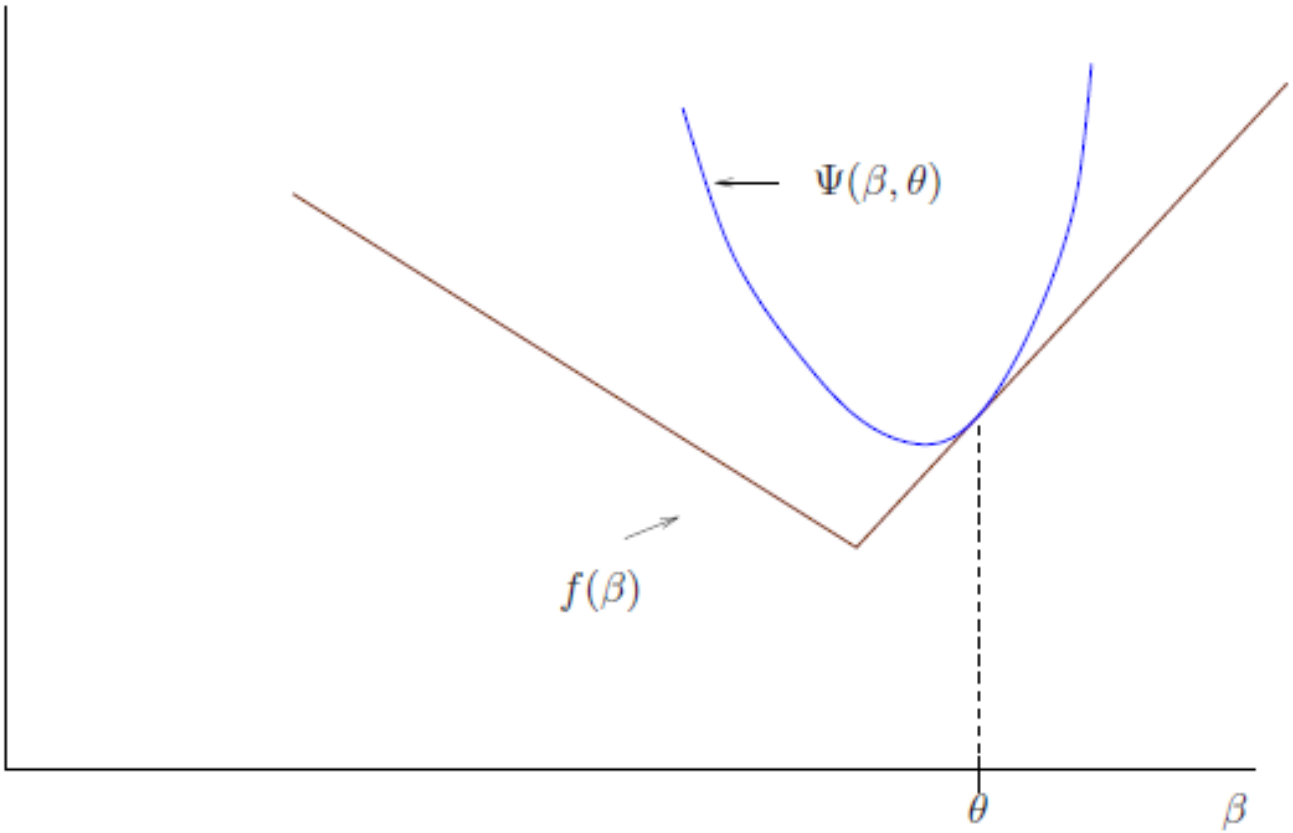
Problem: given $f : R^p \longrightarrow R$. Find

$$\operatorname{argmin}_{\beta \in R^p} f(\beta)$$

$f(\beta)$ usually non-convex, p- large - hard problem
MM algorithm is based on function $\Psi(\beta, \theta) : R^p \times R^p \longrightarrow R$ such that

$$(i) \quad f(\beta) \leqslant \Psi(\beta, \theta) \ , \theta \in R^p$$

$$(ii) \quad f(\beta) = \Psi(\beta, \beta)$$


$\Psi(\beta, \theta)$
$f(\beta)$
θ
β

β^0 -some starting point. β^t -given from t^{th} iteration

$$\beta^{t+1} = \text{argmin}_{\beta \in R^p} \Psi(\beta, \beta^t)$$

Main property of MM algorithm

$$f(\beta^t) =_{(ii)} \Psi(\beta^t, \beta^t) \geqslant \Psi(\beta^{t+1}, \beta^t) \geqslant_{(i)} f(\beta^{t+1})$$

f-convex - procedure yields global maximum

MM algorithm in logistic regression

Usually Ψ obtained by modifying term of order 2 in Taylor expansion of $\log L$

$$\frac{1}{2}(\theta - \beta)^T H(\tilde{\beta})(\theta - \beta)$$

In logistic regression

$$H = diag(\pi(\tilde{\beta}^T x_i)(1 - \pi(\tilde{\beta}^T x_i)) \rightarrow H^* = diag(1/4, \ldots 1/4)$$

$$H \leqslant H^*$$

References

1. P. Teisseyre, J. Mielniczuk, M. Łazęcka, *Different strategies of fitting logistic regression for positive unlabeled data*, Proceedings of ICCS'20, 2020.
2. M. Kubkowski, J. Mielniczuk, *Active sets of predictors for misspecified logistic regression*, Statistics, 2017.
3. C. Elkan, K. Noto, *Learning classifiers from only positive and unlabelled data*, Proceedings of ACM SIGKDD'08, 2008.
4. J. Bekker, J. Davis, *Learning from positive and unlabeled data: a survey*, Machine Learning, 2020.