Classification with Strategically Withheld Data

Anilesh K. Krishnaswamy*, Haoming Li+, David Rein*, Hanrui Zhang*, Vincent Conitzer*
*Duke University, +University of Southern California

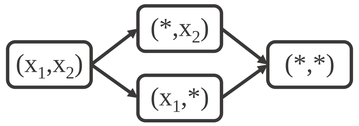
Problem

Strategic Withholding of Feature Values:

- College admission
- Credit approval
- Online dating

In General:

- Agent's type is defined by feature values.
- Agent wants to be accepted, but each type can only (mis)report as certain types.



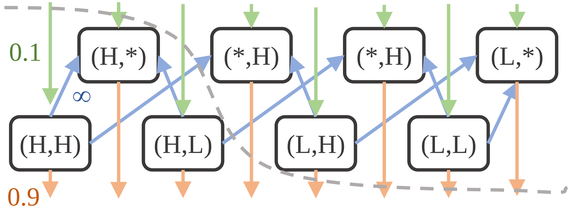
Problem Statement:

- Assume the principal knows the prior distribution over agents' types.
- How to learn a truthful classifier, knowing that agents can withhold feature values?

Solutions

The Min-Cut Classifier:

X	(H,H)	(H,L)	(L,H)	(L,L)	(H,*)	(*,H)	(L,*)	(*,L)
Pr(X)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
Pr(Y=1|X)	0.9	0.7	0.3	0.1	0.6	0.6	0.2	0.2
Pr(Y=0|X)	0.1	0.3	0.7	0.9	0.4	0.4	0.8	0.8



The Hill-Climbing Classifier:

- How to train a strategy-proof ensemble?
- Idea: MAX-ensemble: accept the agent if any of the applicable subclassifiers accepts.
- One subclassifier for each combination of features, each iteratively trained on the data rejected by all other subclassifiers.

The Incentive-Compatible Log Reg Classifier:

- What imputed value gives no better signal?
- Idea: Nonnegative feature values + nonnegative coefficients
- After each gradient step, project the coefficients to feasible nonnegative region.

Evaluation

Vs. mean/mode imputation & reduced-feature; Tru.: truthful report; Str.: strategic (mis)report; disc.: discretizing the feature values.

Classifier	Australia		Germany		Poland		Taiwan	
	Tru.	Str.	Tru.	Str.	Tru.	Str.	Tru.	Str.
HC(LR)	.792	**.792**	.639	.639	.659	.659	.648	.648
MINCUT	.770	.770	.580	.580	.501	.501	.652	**.652**
IC-LR	.788	.788	.654	**.654**	.639	.639	.499	.499
IMP(LR)	.796	.791	**.663**	.580	**.714**	**.660**	**.670**	.618
R-F(LR)	**.808**	.545	.631	.508	.670	.511	.665	.590

Classifier	Australia		Germany		Poland		Taiwan	
	Tru.	Str.	Tru.	Str.	Tru.	Str.	Tru.	Str.
HC(LR) w/ disc.	.794	.794	.641	.641	.692	.692	.650	**.650**
MINCUT w/ disc.	.789	.789	.629	.629	.692	.692	.649	.649
IC-LR w/ disc.	**.800**	**.800**	.651	**.651**	.698	**.698**	.646	.646
IMP(LR) w/ disc.	.799	.762	**.652**	.577	**.719**	.631	**.686**	.541
R-F(LR) w/ disc.	.796	.542	.633	.516	.708	.522	.684	.587