Strategic Classification in the Dark

Vineet Nair (Speaker)

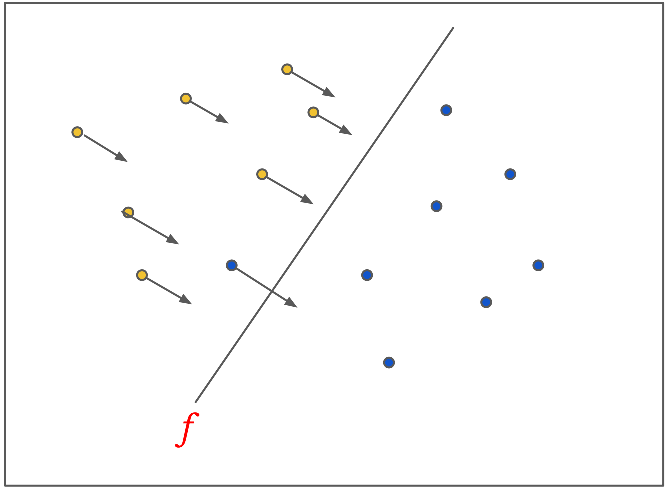
(supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 682203 -ERC-[Inf-Speed-Tradeoff])

TECHNION
Israel Institute of Technology



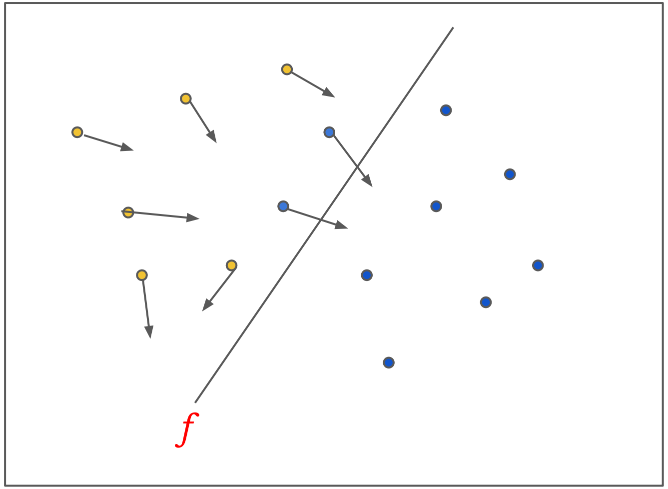

Ganesh Ghalme, Itay Eilat, Inbal Talgam-Cohen, Nir Rosenfeld

Main Idea



- Loan approved
- Loan denied

$f:$ Bank's classifier

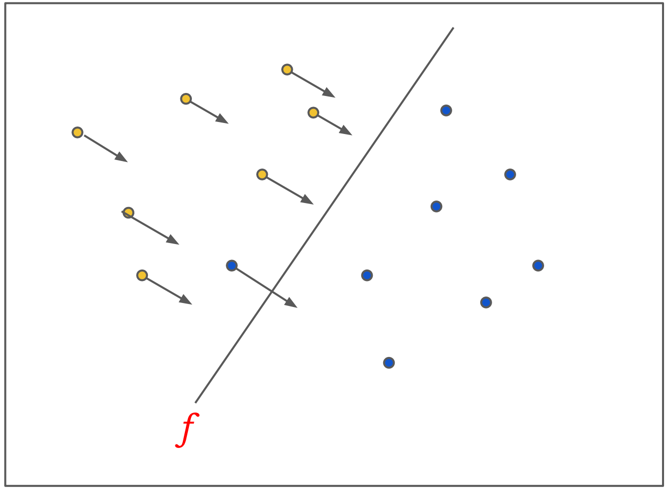


Strategic behaviour of users, dependent on its cost function, for a transparent strategic classifier

Strategic behaviour, dependent on its cost function, for an opaque strategic classifier (users in dark)

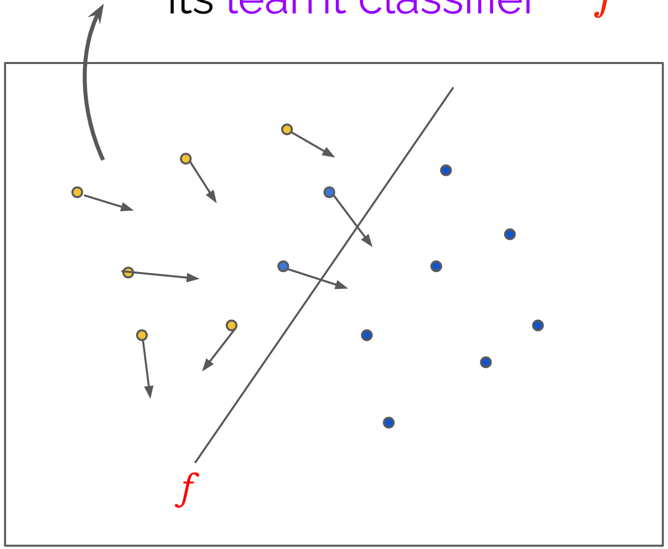
Main Idea

Users moves strategically as per its learnt classifier $\widehat{f}$



- Loan approved
- Loan denied

$f:$ Bank's classifier

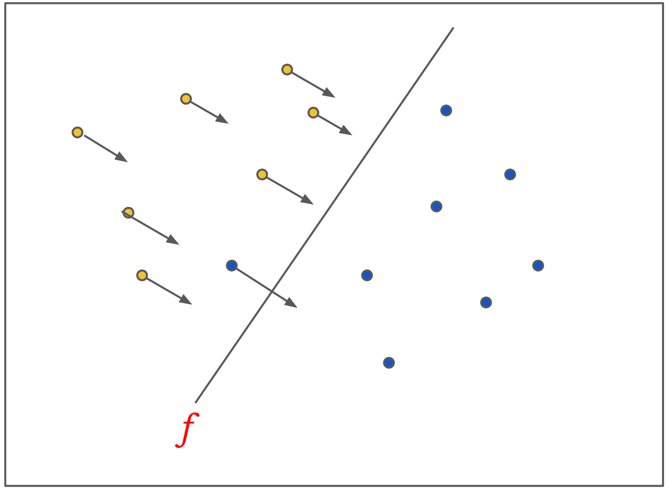


Strategic behaviour of users, dependent on its cost function, for a transparent strategic classifier

Strategic behaviour, dependent on its cost function, for an opaque strategic classifier (users in dark)

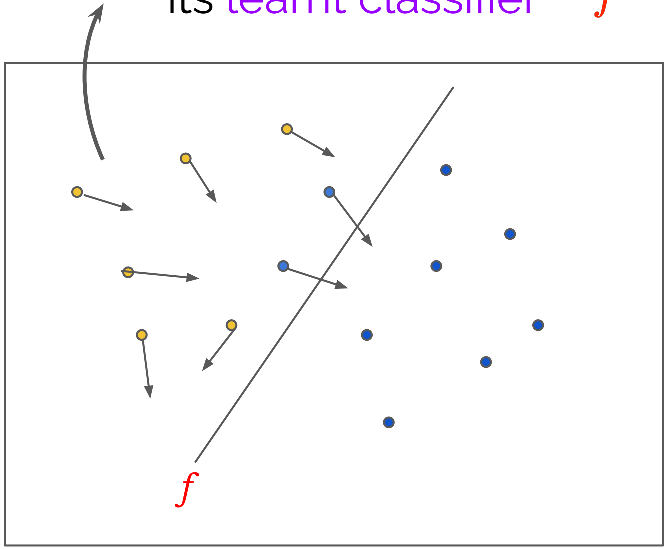
Main Idea

Users moves strategically as per its learnt classifier $\hat{f}$



- Loan approved
- Loan denied

f : Bank's classifier



Strategic behaviour of users, dependent on its cost function, for a transparent strategic classifier

Strategic behaviour, dependent on its cost function, for an opaque strategic classifier (users in dark)

Objective: Compare the classification errors of transparent and opaque strategic classifiers

Main Contribution

- Price of OPacity (POP): Difference between the errors of opaque and transparent strategic classifiers.
    - POP > 0 implies transparency prevails.
- A sufficient condition for POP > 0 which we show is also necessary in some cases.
    - The sufficiency condition depends on the probability mass of the enlargement set (defined next).

Main Contribution

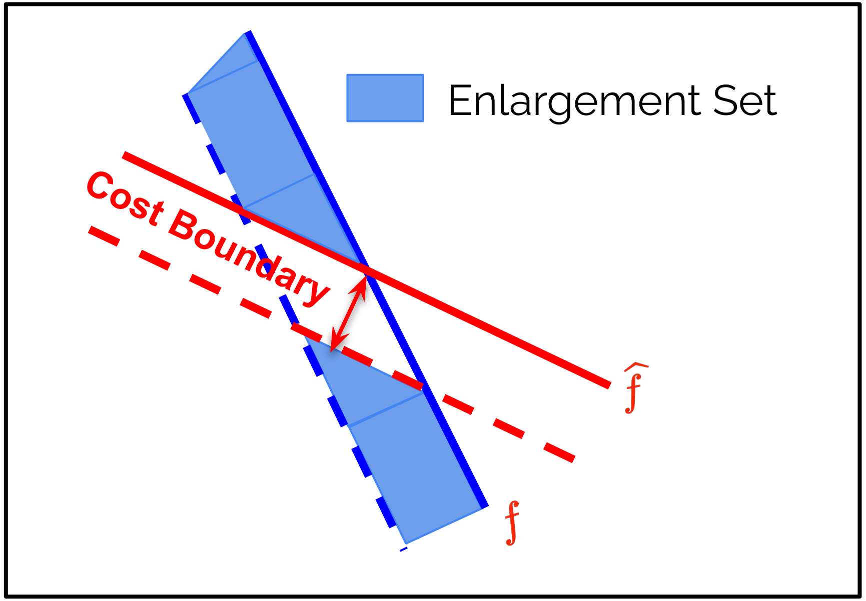
- Price of OPacity (POP): Difference between the errors of opaque and transparent strategic classifiers.
  - POP > 0 implies transparency prevails.
- A sufficient condition for POP > 0 which we show is also necessary in some cases.
  - The sufficiency condition depends on the probability mass of the enlargement set (defined next).
  - We demonstrate the utility of these results by analyzing a normally distributed population classified linearly and show that POP can become arbitrarily large.

Main Contribution

- Price of OPacity (POP): Difference between the errors of opaque and transparent strategic classifiers.
    - POP > 0 implies transparency prevails.
- A sufficient condition for POP > 0 which we show is also necessary in some cases.
    - The sufficiency condition depends on the probability mass of the enlargement set (defined next).
    - We demonstrate the utility of these results by analyzing a normally distributed population classified linearly and show that POP can become arbitrarily large.
- Experiments on synthetic as well as a large dataset on loan requests show that POP can be quite large in practice.

Enlargement Set

- The set of users that were classified differently by f because of opacity.
- We show that even small differences between f and $\hat{f}$ could result in a large probability mass on the enlargement set.



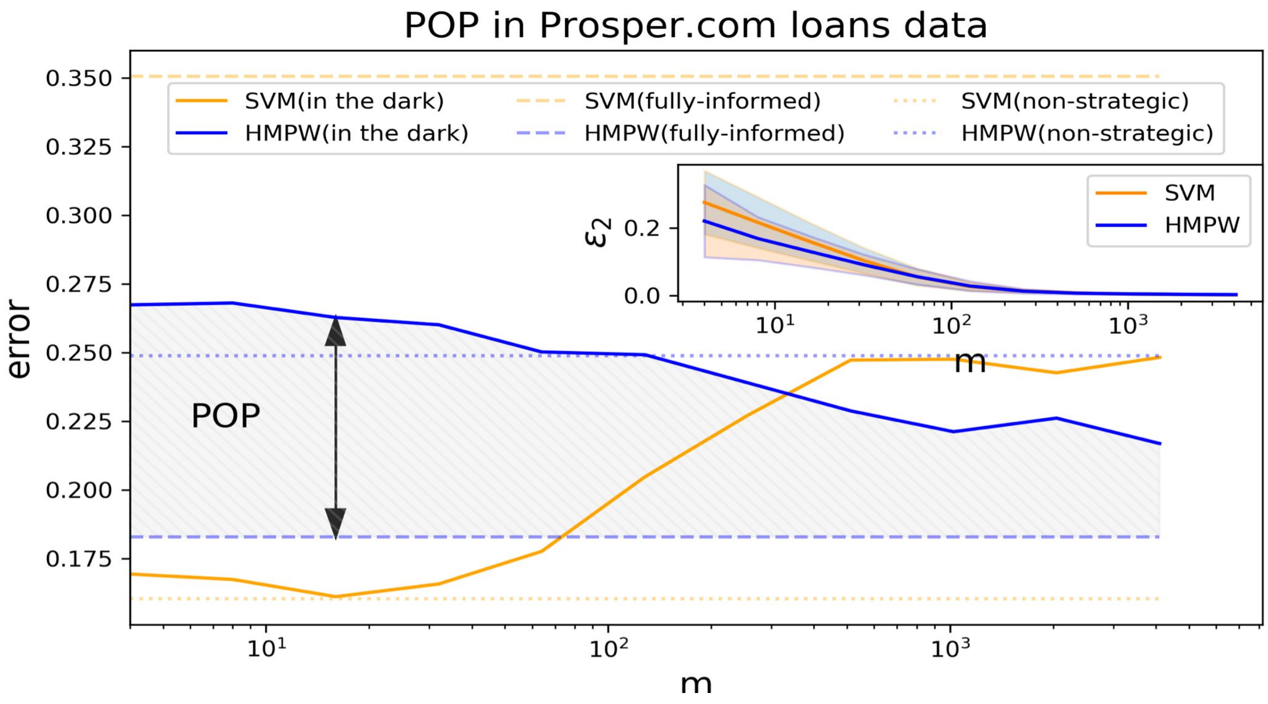
Example of Enlargement Set when f and $\hat{f}$ are linear classifiers

Enlargement Set (Contd.)

- We show a sufficient condition on the probability mass of the enlargement set for POP > 0.
  - The sufficient condition depends on the errors of the optimal classifier for the system and the system's classifier f.

- From the user's perspective, the enlargement set is undesirable.
  - Under opacity these users are classified negatively, whereas under transparency they would have been classified positively.

Experiment showing positive POP



m is the number of samples for learning $\hat{f}$

[HMPW]: Strategic Classification, M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters, Innovations in Theoretical Computer Science (ITCS) 2016.

Key Takeaways

- The System cannot guarantee higher payoff by keeping the users in the dark.

- Even small errors in estimating f by users in dark could result in a big enlargement set implying POP > 0.

- Under an opaque policy Users with access to more samples have greater likelihood of being classified accurately than those with access to fewer samples.

Thank You!