Robust regression

Christophe Giraud

Université Paris-Sud

M2 DS

Don't trust too much theoreticians' fables

The theoretical / practical gap

Typical assumptions in statistical learning theory

- observations $(X_i, Y_i)_{i=1\ldots,n}$ i.i.d.
- large sample size asymptotic $(n \to \infty)$ or sub-Gaussian "errors" $\mathbb{P}(|\varepsilon| > x) \leq e^{-x^2/2}$

Yet, when data are collected in a poorly controlled way, it is not likely to be true...

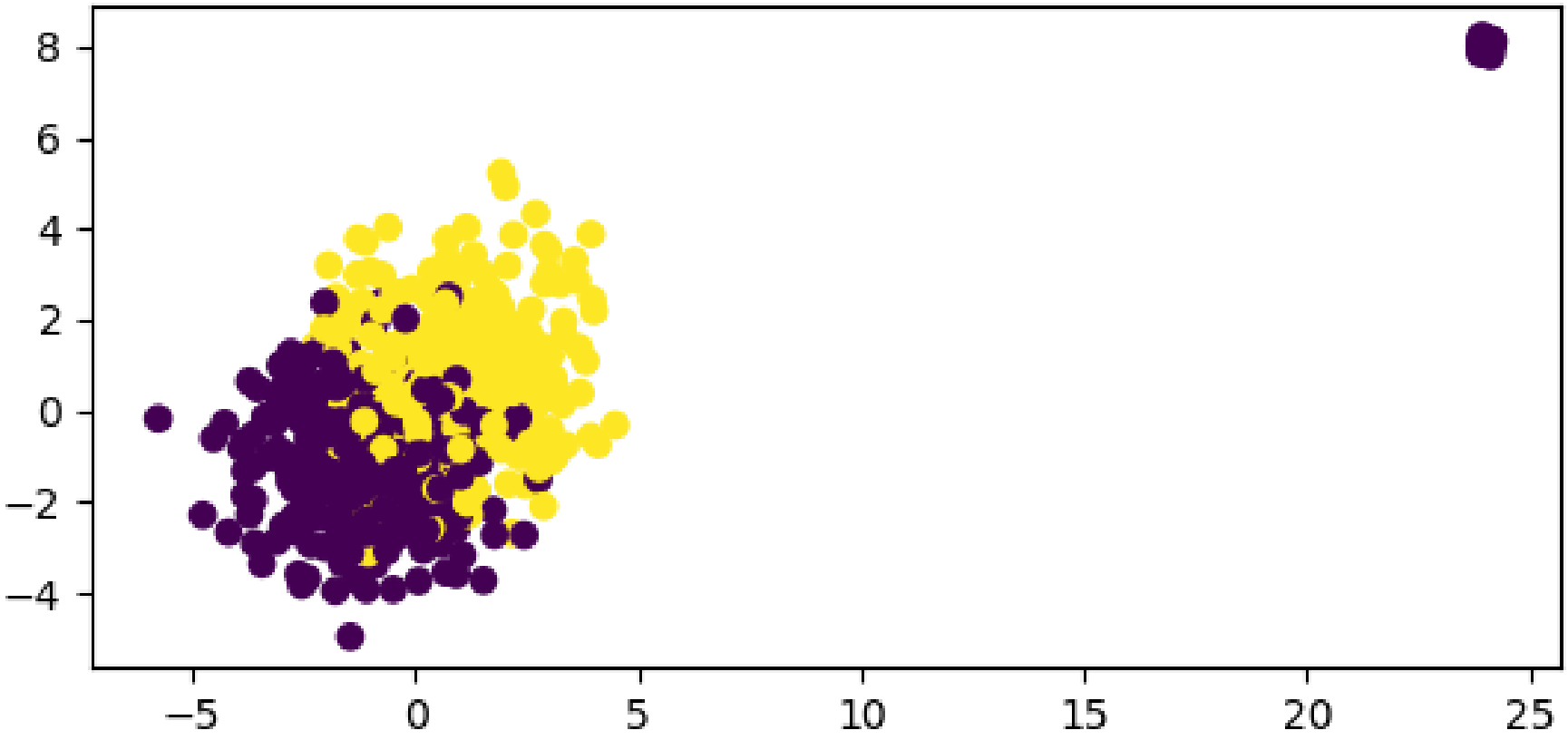
Contaminated data

What happens when a fraction of the data is spurious?

Typically, what if you have some "good" data $(X_i, Y_i)_{i=1\ldots,n_G}$ mixed with some "junk" data $(X_i, Y_i)_{i=1,\ldots,n_J}$?

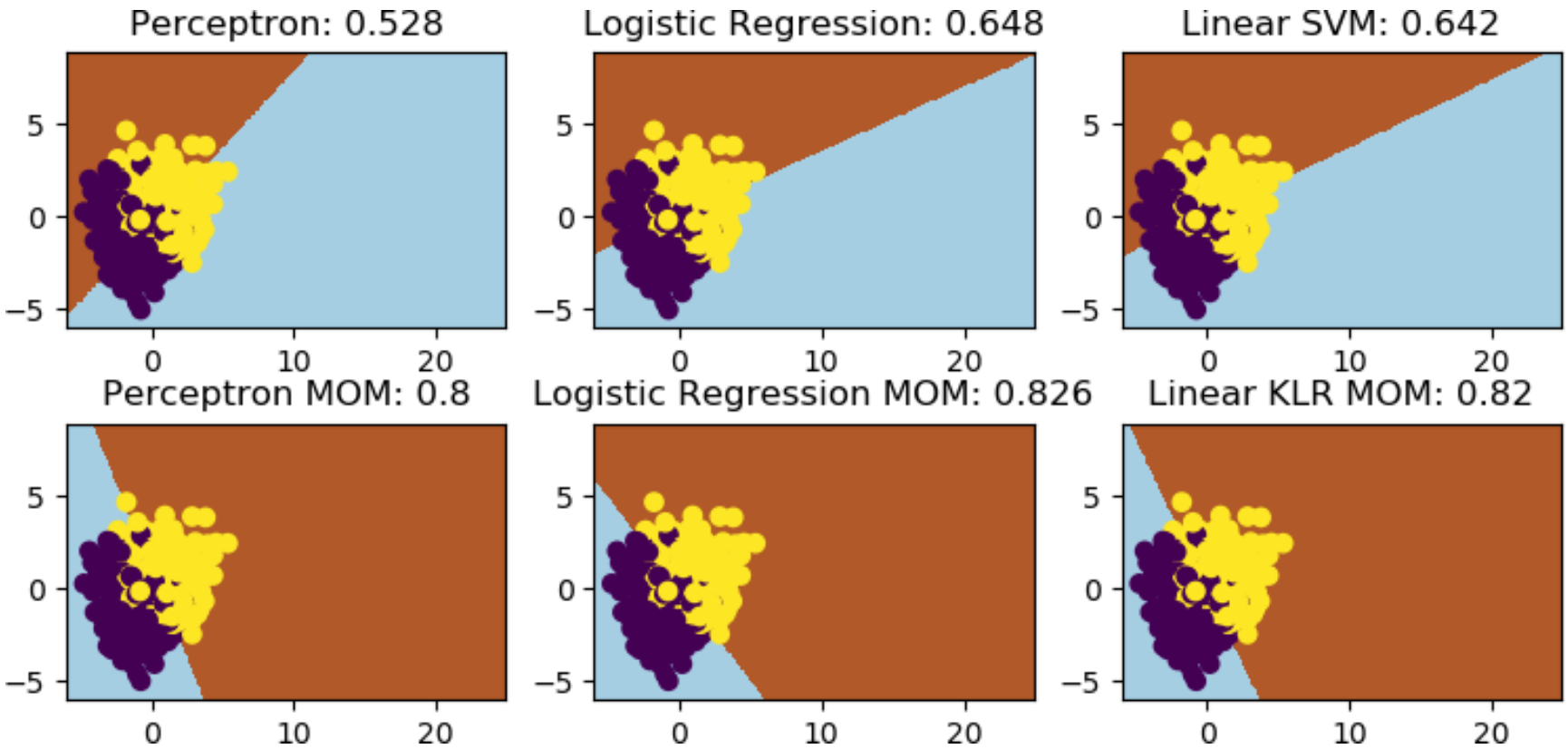
- are the estimators stable?
- at which fraction $n_J/(n_J + n_G)$ of spurious data does an algorithm breakdown?

Example of (badly) contaminated data



Classification problem with a small fraction of spurious data

Classical algorithms versus robustified version



Error with heavy tails

What happens when the errors have heavy tails?

Typically, what if we only have $\mathbb{P}(|\varepsilon| > x) \leq c/x^{1+\delta}$?
Or if we only have $\mathbb{E}\left[|\varepsilon|^{1+\delta}\right] < +\infty$?

- are the classical estimators stable?
- what is the best that we can do?

Topics of the day

Plan

- starter: estimation of mean in presence of heavy tails
- robust regression with Huber loss
- Discussion of the paper: Adaptive Huber Regression: Optimality and Phase Transition