Distributionally Robust Logistic Regression

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn

École Polytechnique Fédérale de Lausanne (EPFL)

Abstract

Motivation: Classification problems face the following challenges

- Data-generating distribution $\mathbb{P}$ is unknown
- Overfitting when # of training samples N is small
- Ad hoc regularization techniques lack theoretical justification

Our solution:

- Use **Distributionally Robust Optimization (DRO)**
- Use the **Wasserstein distance** to construct ambiguity sets

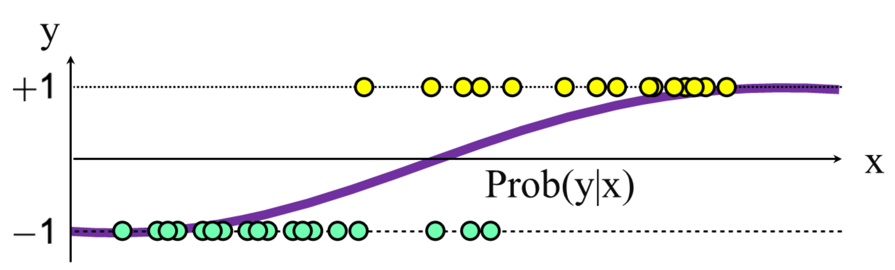
Logistic Regression (LR)

Assumption:

$\text{Prob}(y|x) = \left[1+\exp(-y\langle\beta,x\rangle)\right]^{-1}$

Maximum likelihood estimation:

LR: $\min_\beta \frac{1}{N}\sum_{i=1}^N l_\beta(\hat{x}_i,\hat{y}_i)$



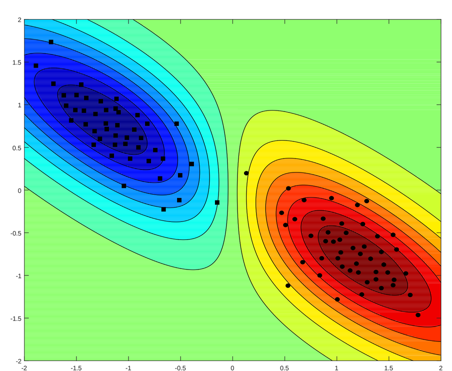
Logloss function:

$l_\beta(x,y) = \log(1+\exp(-y\langle\beta,x\rangle))$

Statistical Learning

The ideal classifier would be a solution of

$$\min_\beta \mathbb{E}^{\mathbb{P}}[l_\beta(x,y)]$$



Challenges:

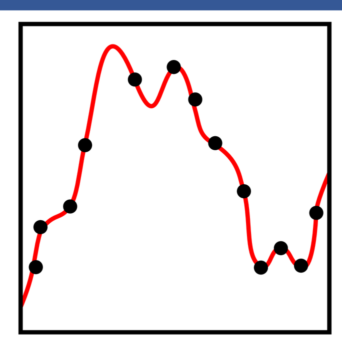
- $\mathbb{P}$ is unknown
- Integration is hard even if $\mathbb{P}$ is uniform on a box

Remedy: replace $\mathbb{P}$ with the empirical distribution $\widehat{\mathbb{P}}_N$

LR $\Leftrightarrow$ $\min_\beta \mathbb{E}^{\widehat{\mathbb{P}}_N}[l_\beta(x,y)]$

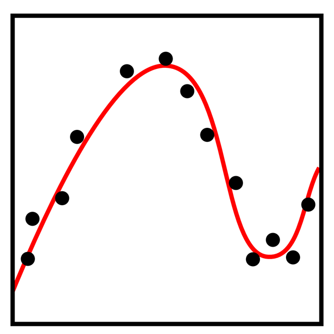
Regularized LR (RLR)

Challenge: LR suffers from overfitting



Remedy: *ad hoc* regularization techniques

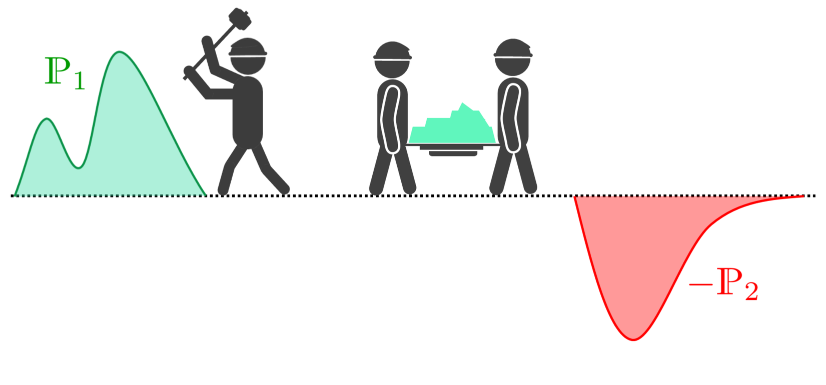
RLR: $\min_\beta \frac{1}{N}\sum_{i=1}^N l_\beta(\hat{x}_i,\hat{y}_i) + c\,R(\beta)$



Questions:

- Probabilistic interpretation for regularization?
- How to choose c and $R(\beta)$?

Wasserstein distance



$W(\mathbb{P}_1,\mathbb{P}_2)$ = least cost of moving $\mathbb{P}_1$ to $\mathbb{P}_2$

$$W(\mathbb{P}_1,\mathbb{P}_2) := \inf_\Pi \mathbb{E}^\Pi\left[d(\xi_1,\xi_2)\right],$$

where $\mathbb{P}_1$ and $\mathbb{P}_2$ are the marginals of ξ_1 and ξ_2 under Π

Metric on feature-label-space:

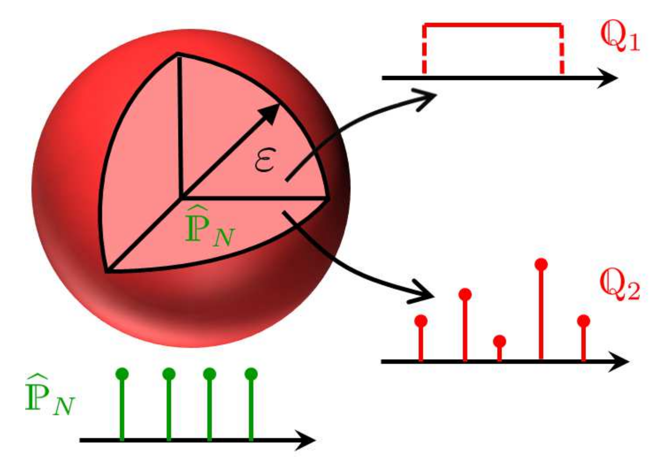
$$d\big((x_1,y_1),(x_2,y_2)\big) = \|x_1-x_2\| + \frac{\kappa}{2}|y_1-y_2|$$

- κ : trust in labels
- $\kappa=\infty \implies$ exact labels

Distributionally Robust LR (DRLR)

Minimize the **worst-case** expected logloss

DRLR: $\min_\beta \sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[l_\beta(x,y)]$



The worst-case is taken over all distributions $\mathbb{Q}$ in the **Wasserstein ball**

$$\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N) := \{\mathbb{Q} : W(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\varepsilon\}$$

Tractability

Theorem 1: DRLR is equivalent to the finite convex program

$$\min_\beta \sup_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\Big[l_\beta(x,y)\Big] = \left\{\begin{array}{lll} \min\limits_{\beta,\lambda,s_i} & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N s_i & \\ \text{s.t.} & l_\beta(\hat{x}_i,\hat{y}_i)\le s_i & \forall i\le N \\ & l_\beta(\hat{x}_i,-\hat{y}_i)-\lambda\kappa\le s_i & \forall i\le N \\ & \|\beta\|_*\le\lambda. & \end{array}\right.$$

Special Case: DRLR reduces to RLR when $\kappa=\infty$

$$\inf_\beta \frac{1}{N}\sum_{i=1}^N \ell_\beta(\hat{x}_i,\hat{y}_i) + \varepsilon\|\beta\|_*$$

Out-of-Sample Performance

Theorem 2: If the tail of $\mathbb{P}$ decays as $\exp(-\|2x\|^a)$ and the radius ε of the Wasserstein ball is set to

$$\varepsilon_N(\eta) = \left(\frac{\log(c_1\eta^{-1})}{c_2 N}\right)^{\frac{1}{a}} \mathbb{1}_{\{N<\frac{\log(c_1\eta^{-1})}{c_2c_3}\}} + \left(\frac{\log(c_1\eta^{-1})}{c_2 N}\right)^{\frac{1}{n}} \mathbb{1}_{\{N\ge\frac{\log(c_1\eta^{-1})}{c_2c_3}\}},$$

then we have

$$\mathbb{P}^N\{\mathbb{P}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)\}\ge 1-\eta \implies \mathbb{P}^N\{\mathbb{E}^{\mathbb{P}}[l_{\hat\beta}(x,y)]\le\hat{J}\}\ge 1-\eta \qquad (*)$$

Note: $1-\eta$ = confidence that $\mathbb{P}$ is in the Wasserstein ball.

Risk Estimation

Theorem 3: The worst/best-case risk $\mathfrak{R}_{\max/\min} := \sup/\inf_{\mathbb{Q}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\mathbb{1}_{\{y\langle\hat\beta,x\rangle\le 0\}}]$ is given by

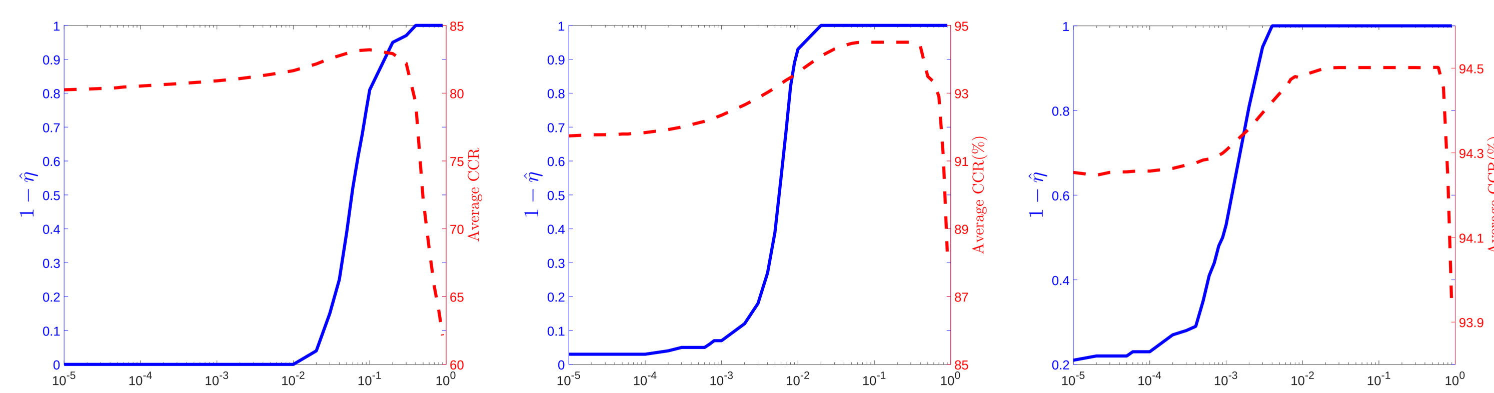
$$\mathfrak{R}_{\max/\min} = 1 - \left\{\begin{array}{lll} \min\limits_{\lambda,s_i,r_i,t_i} & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N s_i & \\ \text{s.t.} & 1\mp r_i\hat{y}_i\langle\hat\beta,\hat{x}_i\rangle\le s_i & \forall i\le N \\ & 1\pm t_i\hat{y}_i\langle\hat\beta,\hat{x}_i\rangle-\lambda\kappa\le s_i & \forall i\le N \\ & r_i\|\hat\beta\|_*\le\lambda,\quad t_i\|\hat\beta\|_*\le\lambda & \forall i\le N \\ & r_i,t_i,s_i\ge 0 & \forall i\le N \end{array}\right.$$

If $\varepsilon\ge\varepsilon_N(\eta)$, then $\mathfrak{R}_{\min}(\hat\beta)\le\mathfrak{R}(\hat\beta)\le\mathfrak{R}_{\max}(\hat\beta)$ with probability $1-2\eta$.

Results

Synthetic Experiments (out-of-sample performance):

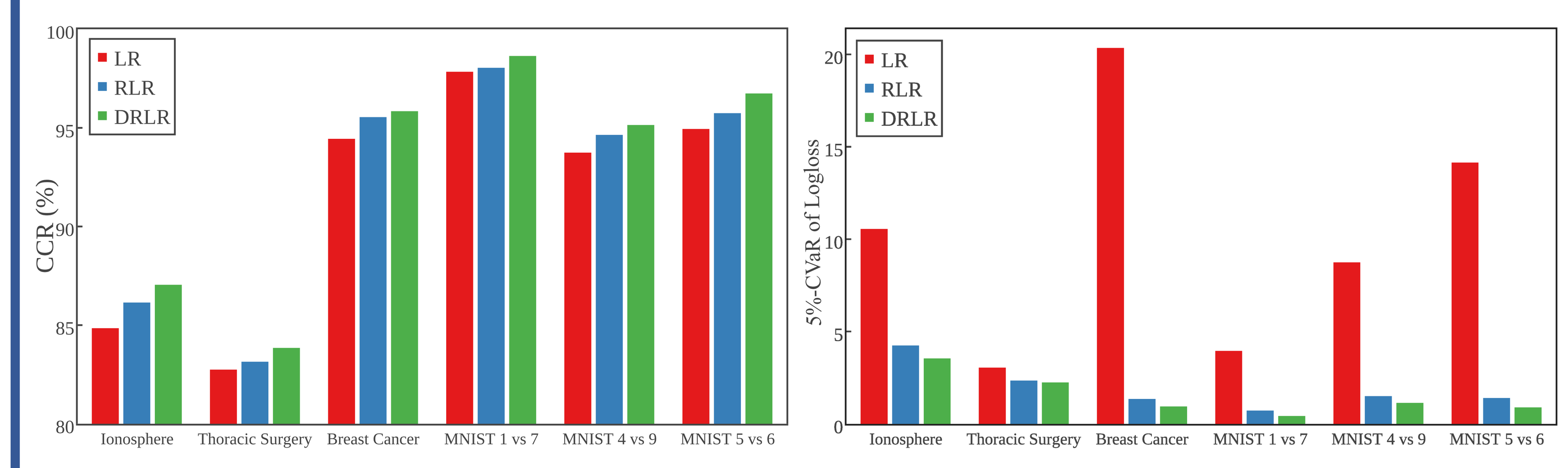
- $\dim(x)=10$
- $x\sim\mathcal{N}(0,I)$
- $\text{Prob}(y|x) = \left[1+\exp(-10yx_1)\right]^{-1}$



Observations:

- Empirical confidence that $\mathbb{P}\in\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)(\widehat{\mathbb{P}}_N)$ saturates when the out-of-sample CCR is maximal
- The saturation point scales with N^{-1} consistent with $(*)$

Empirical Experiments: (MNIST & UCI datasets)



- Improvement of DRLR over RLR ≈ improvement of RLR over LR

Ionosphere dataset: (UCI datasets)