So far we studied moment-based uncertainty sets. In some applications, we may be more interested in distributions that are "close" to a reference distribution that is induced by the data.

e.g.: Consider the scenario where we have iid realizations $\xi_1, \cdots, \xi_N$ of a random vector ξ. Define the empirical measure by

$$\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}$$

Intuitively, as $N \to \infty$, $\hat{\mathbb{P}}_N$ should tend to the true distribution $\mathbb{P}^*$ of ξ.

* For fixed measurable $A \subseteq \mathbb{R}^n$, $\hat{\mathbb{P}}_N(A) \to \mathbb{P}^*(A)$ (SLLN)
* Uniform convergence of $\hat{\mathbb{P}}_N$ to $\mathbb{P}^*$: Vapnik-Chervonenkis.

* How do we describe a "neighborhood" of the reference distribution?

1) Divergence-based measures

e.g.: f-divergence

Let $f: \mathbb{R}_+ \to \mathbb{R}$ be convex with $f(1)=0$. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability distributions on a probability space $(\Xi, \mathcal{B})$ with $\mathbb{P}$ being the reference distribution. Suppose that $\mathbb{Q}$ is absolutely continuous wrt $\mathbb{P}$ (intuitively, $\mathbb{P}, \mathbb{Q}$ have the same support).

The f-divergence of $\mathbb{Q}$ from $\mathbb{P}$ is defined as

$$D_f(\mathbb{Q} \,\|\, \mathbb{P}) = \int f\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right) d\mathbb{P}$$

e.g.: KL-divergence: $f(t) = t \log t$

TV: $f(t) = \frac{1}{2}|t-1|$

Then, we can define the "ball"

$$\mathbb{B}_{f,\varepsilon}(\mathbb{P}) = \{\mathbb{Q} : D_f(\mathbb{Q} \,\|\, \mathbb{P}) \le \varepsilon\}$$

See, e.g.,

Ben-Tal et al.: Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. Manag. Sci. 59(2): 341–357, 2013.

2) Probability metric-based measures

e.g.: Wasserstein distance

Let $\mathbb{P}$, $\mathbb{Q}$ be as above. Let $d(\cdot,\cdot)$ be a norm on $\mathbb{R}^n$. Define the d-Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ by

$$W(\mathbb{Q},\mathbb{P}) = \inf_{\Pi \in \mathcal{M}(\Xi\times\Xi)} \left\{ \int_{\Xi\times\Xi} \underbrace{d(\xi,\xi')}_{\text{transport cost}} \Pi(d\xi, d\xi') : \underbrace{\Pi(\Xi, d\xi') = \mathbb{P}(d\xi'),\ \Pi(d\xi, \Xi) = \mathbb{Q}(d\xi)}_{\text{marginal distributions}} \right\}$$

Then, we can define the ball

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

Wasserstein distance-based DRO has attracted much attention due to its connection to various problems in machine learning.

Application: Logistic Regression

Setup. * $x \in \mathbb{R}^n$: feature vector ; $y \in \{+1,-1\}$ binary label

* conditional distribution of y given x:

$$\Pr(y|x) = (1+\exp(-y\cdot\beta^T x))^{-1},$$

where β is the regression parameter.

* Suppose that N training samples $\{\hat{\xi}_i = (\hat{x}_i, \hat{y}_i)\}_{i=1}^N$ of the underlying random vector $\xi = (x,y)$ are observed.

Then, the MLE of β is given by

$$\hat{\beta}_{MLE} = \arg\min_\beta \frac{1}{N}\sum_{i=1}^N l_\beta(\hat{\xi}_i) = \arg\min_\beta \mathbb{E}_{\hat{\mathbb{P}}_N}[l_\beta(\xi)],$$

where $l_\beta(x,y) = \log(1+\exp(-y\beta^T x))$ is the log-loss.

To avoid overfitting, a typical approach is regularization:

To avoid overfitting, a typical approach is regularization:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell_\beta(\xi)] + \varepsilon R(\beta)$$

e.g.: $R(\beta) = \|\beta\|_1$, $R(\beta) = \|\beta\|_2^2$, ...

sparsity-inducing ridge

One way of understanding the overfitting phenomenon is that $\hat{\beta}_{MLE}$ above does not account for the unseen data. This motivates the following formulation:

$$\min_{\beta} \sup_{\mathbb{B}_\varepsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{P}}[\ell_\beta(\xi)]$$

The hope is that if ε is chosen appropriately, then the true distribution $\mathbb{P}^* \in \mathbb{B}_\varepsilon(\hat{\mathbb{P}}_N)$, so that the solution β^* takes into account the effect of $\mathbb{P}^*$.

Here, for $\xi = (x, y)$, $\xi' = (x', y')$, we take

$$d(\xi, \xi') = \|x - x'\| + \frac{\kappa}{2}|y - y'|$$

as the transport cost, where $\|\cdot\|$ is an arbitrary norm and $\kappa > 0$ is a parameter that specifies the relative emphasis between feature mismatch and label uncertainty.

* In particular, if $\kappa = +\infty$, then the label needs to be a deterministic function of the feature, and label measurements are exact.

Recall our problem of interest:

$$\inf_{\beta \in \mathbb{R}^n} \sup_{\mathbb{Q} \in B_\varepsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\ell_\beta(x,y) \right],$$

where

$$\ell_\beta(x,y) = \log(1+\exp(-y\beta^T x)),$$

$$B_\varepsilon(\hat{\mathbb{P}}_N) = \left\{ \mathbb{Q} \in \mathcal{M}(\Xi) : W(\mathbb{Q}, \hat{\mathbb{P}}_N) \le \varepsilon \right\}, \quad \hat{\mathbb{P}}_N(\xi) = \frac{1}{N}\sum_{i=1}^N \delta_{(\hat{x}_i,\hat{y}_i)}(\xi),$$

$$W(\mathbb{Q},\mathbb{P}) = \inf_{\Pi \in \mathcal{M}(\Xi\times\Xi)} \left\{ \int_{\Xi\times\Xi} d(\xi,\xi')\Pi(d\xi,d\xi') : \Pi(\Xi,d\xi') = \mathbb{P}(d\xi'),\ \Pi(d\xi,\Xi) = \mathbb{Q}(d\xi) \right\},$$

$$\Xi = \mathbb{R}^n \times \{-1,+1\}, \quad \xi = (x,y), \quad d(\xi,\xi') = \|x-x'\| + \frac{\kappa}{2}|y-y'|$$

As before, consider the inner sup problem:

$$\sup_{\mathbb{Q} \in B_\varepsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\ell_\beta(x,y) \right] = \begin{array}{rl} \sup\limits_{\Pi \in \mathcal{M}(\Xi\times\Xi)} & \int_\Xi \ell_\beta(\xi) \underbrace{\Pi(d\xi,\Xi)}_{\mathbb{Q}(d\xi)} \\ \text{s.t.} & \int_{\Xi\times\Xi} d(\xi,\xi')\Pi(d\xi,d\xi') \le \varepsilon \\ & \Pi(\Xi,d\xi') = \hat{\mathbb{P}}_N(d\xi') \end{array}$$

$*$ Since $\hat{\mathbb{P}}_N$ is discrete, we have

$$\begin{aligned} \mathbb{Q}(d\xi) = \Pi(d\xi,\Xi) &= \int_{\xi'\in\Xi} \Pi(d\xi,d\xi') \\ &= \sum_{i=1}^N \Pi\left(d\xi \,\middle|\, \xi' = (\hat{x}_i,\hat{y}_i)\right) \cdot \hat{\mathbb{P}}_N(\hat{x}_i,\hat{y}_i) \\ &= \frac{1}{N}\sum_{i=1}^N \mathbb{Q}^i(d\xi) \end{aligned}$$

conditional distribution of ξ given $\xi' = (\hat{x}_i,\hat{y}_i)$

Similarly,

$$\Pi(d\xi,d\xi') = \Pi(d\xi \mid \xi')\cdot \hat{\mathbb{P}}_N(\xi') = \frac{1}{N}\sum_{i=1}^N \delta_{(\hat{x}_i,\hat{y}_i)}(\xi')\mathbb{Q}^i(d\xi)$$

Hence,

$$\sup_{\Pi\in\mathcal{M}(\Xi\times\Xi)} \int_\Xi \ell_\beta(\xi)\underbrace{\Pi(d\xi,\Xi)} \qquad \sup_{\mathbb{Q}^i \ge 0} \frac{1}{N}\sum_{i=1}^N \int_\Xi \ell_\beta(\xi)\mathbb{Q}^i(d\xi)$$

$$\sup_{\Pi \in M(\Xi\times\Xi)} \int_{\Xi} \ell_\beta(\xi)\underbrace{\Pi(d\xi,\Xi)}_{Q(d\xi)} \qquad = \qquad \sup_{Q^i \geq 0} \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi} \ell_\beta(\xi)\, Q^i(d\xi)$$

$$\text{s.t.} \int_{\Xi\times\Xi} d(\xi,\xi')\,\Pi(d\xi,d\xi') \leq \varepsilon \qquad\qquad \text{s.t.}\ \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi} d(\xi,(\hat{x}_i,\hat{y}_i))\,Q^i(d\xi) \leq \varepsilon,$$

$$\Pi(\Xi, d\xi') = \hat{\mathbb{P}}_N(d\xi') \qquad\qquad \int_{\Xi} Q^i(d\xi) = 1.$$

Now, recall $\xi \in \mathbb{R}^n \times \{-1,+1\}$. We decompose

$$Q^i(d\xi) = Q^i_{+1}(dx) + Q^i_{-1}(dx) \quad \text{where } Q^i_{+1}(dx) = Q(dx, y=1)$$

Then, we rewrite

$$\sup_{Q^i \geq 0} \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi} \ell_\beta(\xi)\, Q^i(d\xi)$$

$$\text{s.t.}\ \frac{1}{N}\sum_{i=1}^{N}\int_{\Xi} d(\xi,(\hat{x}_i,\hat{y}_i))\,Q^i(d\xi) \leq \varepsilon,$$

$$\int_{\Xi} Q^i(d\xi) = 1.$$

$$=$$

$$\sup_{Q^i \geq 0} \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}^n} \ell_\beta(x,+1)\,Q^i_{+1}(dx) + \ell_\beta(x,-1)\,Q^i_{-1}(dx)$$

$$\text{s.t.}\ \frac{1}{N}\sum_{i=1}^{N}\Big[\int_{\mathbb{R}^n} d((x,+1),(\hat{x}_i,\hat{y}_i))\,Q^i_{+1}(dx) + \int_{\mathbb{R}^n} d((x,-1),(\hat{x}_i,\hat{y}_i))\,Q^i_{-1}(dx)\Big] \leq \varepsilon,$$

$$\int_{\Xi} Q^i_{+1}(dx) + Q^i_{-1}(dx) = 1$$

$$= \sup_{Q^i \geq 0} \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}^n} \ell_\beta(x,+1)\,Q^i_{+1}(dx) + \ell_\beta(x,-1)\,Q^i_{-1}(dx)$$

$$\text{s.t.}\ \frac{1}{N}\Big[\sum_{i:\hat{y}_i=+1}\int_{\mathbb{R}^n} \|x-\hat{x}_i\|\,Q^i_{+1}(dx) + (\|x-\hat{x}_i\|+\kappa)\,Q^i_{-1}(dx)$$

$$+ \sum_{i:\hat{y}_i=-1}\int_{\mathbb{R}^n} (\|x-\hat{x}_i\|+\kappa)\,Q^i_{+1}(dx) + \|x-\hat{x}_i\|\,Q^i_{-1}(dx)\Big] \leq \varepsilon,$$

$$\int_{\Xi} Q^i_{+1}(dx) + Q^i_{-1}(dx) = 1$$

$$= \sup_{Q^i \geq 0} \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{R}^n} \ell_\beta(x,+1)\,Q^i_{+1}(dx) + \ell_\beta(x,-1)\,Q^i_{-1}(dx)$$

$$\text{s.t.}\ \frac{1}{N}\int_{\mathbb{R}^n} \kappa\sum_{i:\hat{y}_i=+1} Q^i_{-1}(dx) + \kappa\sum_{i:\hat{y}_i=-1} Q^i_{+1}(dx)$$

$$+ \sum_{i=1}^{N}\|x-\hat{x}_i\|\big(Q^i_{+1}(dx) + Q^i_{-1}(dx)\big) \leq \varepsilon, \quad (\lambda) \qquad —— (\Delta)$$

$$\int_{\Xi} \mathbb{Q}^i_{+1}(dx) + \mathbb{Q}^i_{-1}(dx) = 1. \quad (s_i)$$

Now, take the dual of (Δ) to get (Exercise)

$$\inf_{\lambda, s_i} \quad \lambda \varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i$$

$$\text{s.t.} \quad \sup_{x \in \mathbb{R}^n} \; l_\beta(x,+1) - \lambda\|x-\hat{x}_i\| - \frac{1}{2}\lambda k(1-\hat{y}_i) \le s_i, \qquad \text{— (D)}$$

$$\sup_{x \in \mathbb{R}^n} \; l_\beta(x,-1) - \lambda\|x-\hat{x}_i\| - \frac{1}{2}\lambda k(1+\hat{y}_i) \le s_i,$$

$$\lambda \ge 0.$$

Exercise: Verify that strong duality holds for any $\varepsilon > 0$

Claim: For every $\lambda > 0$,

$$\sup_{x \in \mathbb{R}^n} l_\beta(x,\pm 1) - \lambda\|x-x'\| = \begin{cases} l_\beta(x',\pm 1) & \text{if } \|\beta\|_* \le \lambda, \\ -\infty & \text{o/w} \end{cases}$$

(Exercise, use conjugate functions)

It then follows that (D) is equivalent to

$$\begin{aligned} \inf_{\lambda, s_i} \quad & \lambda \varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{s.t.} \quad & l_\beta(\hat{x}_i,+1) - \frac{1}{2}\lambda k(1-\hat{y}_i) \le s_i, \\ & l_\beta(\hat{x}_i,-1) - \frac{1}{2}\lambda k(1+\hat{y}_i) \le s_i, \\ & \|\beta\|_* \le \lambda. \end{aligned} \quad = \quad \begin{aligned} \inf_{\lambda, s_i} \quad & \lambda \varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{s.t.} \quad & l_\beta(\hat{x}_i,\hat{y}_i) \le s_i, \\ & l_\beta(\hat{x}_i,-\hat{y}_i) - \lambda k \le s_i, \\ & \|\beta\|_* \le \lambda. \end{aligned}$$

* Putting back the original $\inf_{\beta \in \mathbb{R}^n}$, we obtain a convex optimization problem in the variables (β, λ, s_i).

* Suppose that we take $k = +\infty$ (all labels are deterministic). Then, we get

$$\inf_{\beta \in \mathbb{R}^n} \quad \frac{1}{N}\sum_{i=1}^{N} l_\beta(\hat{x}_i, \hat{y}_i) + \varepsilon\|\beta\|_*$$

This provides a DRO interpretation of regularization

References:

1) Shafieezadeh-Abadeh, Mohajerin Esfahani, Kuhn. Distributionally Robust Logistic Regression. NIPS 2015

2) Shafieezadeh-Abadeh, Kuhn, Mohajerin Esfahani: Regularization via Mass Transportation. JMLR 2019.

Further Reading

1) Kuhn et al.: Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. INFORMS TutORials in Operations Research, 2019

2) Rahimian, Mehrotra: Distributionally Robust Optimization: A Review. arXiv 2019.