Optimización Robusta de Distribuciones basada en Datos

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Octubre de 2017

Contenido

Optimización Robusta de Distribuciones basada en Datos

- En la aproximación RWPI al problema de inferencia y regularización estudiada, la distribución empírica solo se usa para centrar la bola en la métrica de Wasserstein: $U_\delta(P_n)$.
- El objetivo de este artículo es informar la *forma* de esta bola usando los datos observados.
- La definición de la distribuciones palusibles depende de la función de costo de transporte y la métrica de Wasserstein.

Optimización Robusta de Distribuciones basada en Datos

- En la aproximación RWPI al problema de inferencia y regularización estudiada, la distribución empírica solo se usa para centrar la bola en la métrica de Wasserstein: $U_\delta(P_n)$.
- El objetivo de este artículo es informar la *forma* de esta bola usando los datos observados.
- La definición de la distribuciones palusibles depende de la función de costo de transporte y la métrica de Wasserstein.

Optimización Robusta de Distribuciones basada en Datos

- En la aproximación RWPI al problema de inferencia y regularización estudiada, la distribución empírica solo se usa para centrar la bola en la métrica de Wasserstein: $U_\delta(P_n)$.
- El objetivo de este artículo es informar la *forma* de esta bola usando los datos observados.
- La definición de la distribuciones palusibles depende de la función de costo de transporte y la métrica de Wasserstein.

- Algunas otras métricas son útiles, por ejemplo Kullback - Leibler, pero fuerzan las distribuciones a tener el mismo soporte de la distribución empírica y pueden tener un problema de sobre ajuste (referencias).
- La métrica de Wasserstein captura características globales del espacio de distribuciones.

- Algunas otras métricas son útiles, por ejemplo Kullback - Leibler, pero fuerzan las distribuciones a tener el mismo soporte de la distribución empírica y pueden tener un problema de sobre ajuste (referencias).
- La métrica de Wasserstein captura características globales del espacio de distribuciones.

- Este artículo no estudia la pertinencia de la métrica de Wasserstein sino, formas informadas por los datos observados de construir la función de costos.
- En el problema estudiado hasta ahora:

$$c((x,y),(x',y')) = \left\|x - x'\right\|_q^2 + \infty \mathbb{I}(y \neq y')$$

- El énfasis es en la utilización de técnicas de aprendizaje de métricas.
- En los ejemplos el parámetro de regularización se elige usando validación cruzada.

Optimización Robusta de Distribuciones basada en Datos

- Este artículo no estudia la pertinencia de la métrica de Wasserstein sino, formas informadas por los datos observados de construir la función de costos.
- En el problema estudiado hasta ahora:

$$c((x,y),(x',y')) = \left\|x - x'\right\|_q^2 + \infty \mathbb{I}(y \neq y')$$

- El énfasis es en la utilización de técnicas de aprendizaje de métricas.
- En los ejemplos el parámetro de regularización se elige usando validación cruzada.

Optimización Robusta de Distribuciones basada en Datos

- Este artículo no estudia la pertinencia de la métrica de Wasserstein sino, formas informadas por los datos observados de construir la función de costos.
- En el problema estudiado hasta ahora:

$$c((x,y),(x',y')) = \left\|x - x'\right\|_q^2 + \infty \mathbb{I}(y \neq y')$$

- El énfasis es en la utilización de técnicas de aprendizaje de métricas.
- En los ejemplos el parámetro de regularización se elige usando validación cruzada.

- Este artículo no estudia la pertinencia de la métrica de Wasserstein sino, formas informadas por los datos observados de construir la función de costos.
- En el problema estudiado hasta ahora:

$$c((x,y),(x',y')) = \left\|x - x'\right\|_q^2 + \infty \mathbb{I}(y \neq y')$$

- El énfasis es en la utilización de técnicas de aprendizaje de métricas.
- En los ejemplos el parámetro de regularización se elige usando validación cruzada.

Optimización Robusta de Distribuciones basada en Datos

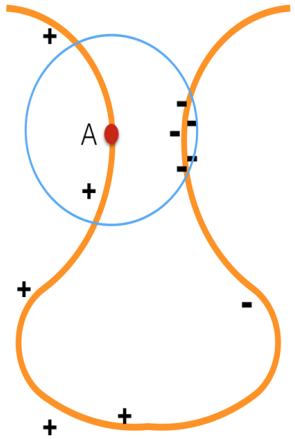
- La idea es elegir una función de costos que penalice el transporte entre dos puntos cuando esto implica impactar considerablemente el riesgo (esperado?).
- Como aplicación muestran como recuperar algunos estimadores en el contexto de regularización adaptativa.

Optimización Robusta de Distribuciones basada en Datos

- La idea es elegir una función de costos que penalice el transporte entre dos puntos cuando esto implica impactar considerablemente el riesgo (esperado?).
- Como aplicación muestran como recuperar algunos estimadores en el contexto de regularización adaptativa.

Contenido

Aprendizaje de Métricas



(a)

- Usar una métrica a lo largo de la variedad en la que aproximadamente se encuentran las observaciones mejora los resultados del problema de optimización robusta. Fuerza al adversario a elegir distrbuciones plausibles que concentran su masa de tranporte a lo largo de la variedad donde localmente los ejemplos son de la misma clase.

Aprendizaje de Métricas: Mahalanobis

- Sea $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Supongamos que tenemos un grafo $(\{X_i\}_{i=1,\ldots,N}, R)$ y se definen dos conjuntos
  $\mathcal{M} = \{(X_i, X_j) : X_i R X_j\}, \mathcal{N} = \{(X_i, X_j) :\neq X_i R X_j\}$.
- Sea PSD el conjunto de matrices reales cuadradas simétricas semidefinidas positivas y $\Delta \in PSD$.
- La distancia de Mahalanobis se define como:

$$d_\Delta(x, x') = \sqrt{(x - x')\Delta(x - x')}$$

Aprendizaje de Métricas: Mahalanobis

- Sea $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Supongamos que tenemos un grafo $(\{X_i\}_{i=1,\ldots,N}, R)$ y se definen dos conjuntos
  $\mathcal{M} = \{(X_i, X_j) : X_i R X_j\}, \mathcal{N} = \{(X_i, X_j) :\neq X_i R X_j\}$.
- Sea PSD el conjunto de matrices reales cuadradas simétricas semidefinidas positivas y $\Delta \in PSD$.
- La distancia de Mahalanobis se define como:

$$d_\Delta(x, x') = \sqrt{(x - x')\Delta(x - x')}$$

Aprendizaje de Métricas: Mahalanobis

- Sea $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Supongamos que tenemos un grafo $(\{X_i\}_{i=1,\ldots,N}, R)$ y se definen dos conjuntos
  $\mathcal{M} = \{(X_i, X_j) : X_i R X_j\}, \mathcal{N} = \{(X_i, X_j) :\neq X_i R X_j\}$.
- Sea PSD el conjunto de matrices reales cuadradas simétricas semidefinidas positivas y $\Delta \in PSD$.
- La distancia de Mahalanobis se define como:

$$d_\Delta(x, x') = \sqrt{(x - x')\Delta(x - x')}$$

Aprendizaje de Métricas: Mahalanobis

- Sea $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Supongamos que tenemos un grafo $(\{X_i\}_{i=1,\ldots,N}, R)$ y se definen dos conjuntos
  $\mathcal{M} = \{(X_i, X_j) : X_i R X_j\}, \mathcal{N} = \{(X_i, X_j) :\neq X_i R X_j\}$.
- Sea PSD el conjunto de matrices reales cuadradas simétricas semidefinidas positivas y $\Delta \in PSD$.
- La distancia de Mahalanobis se define como:

$$d_\Delta(x, x') = \sqrt{(x - x')\Delta(x - x')}$$

Aprendizaje de Métricas: Mahalanobis

- Ahora considere el siguiente problema:

$$\min_{\Delta \in PSD} \sum_{(X_i, X_j) \in \mathcal{M}} d^2(X_i, X_j)$$

$$s.t. \sum_{(X_i, X_j) \in \mathcal{N}} d^2(X_i, X_j) \geq \lambda$$

- Se puede normalizar $\lambda = 1$.

Aprendizaje de Métricas: Mahalanobis

- Ahora considere el siguiente problema:

$$\min_{\Delta \in PSD} \sum_{(X_i, X_j) \in \mathcal{M}} d^2(X_i, X_j)$$
$$s.t. \sum_{(X_i, X_j) \in \mathcal{N}} d^2(X_i, X_j) \geq \lambda$$

- Se puede normalizar $\lambda = 1$.

Optimización Robusta de Distribuciones Basada en Datos

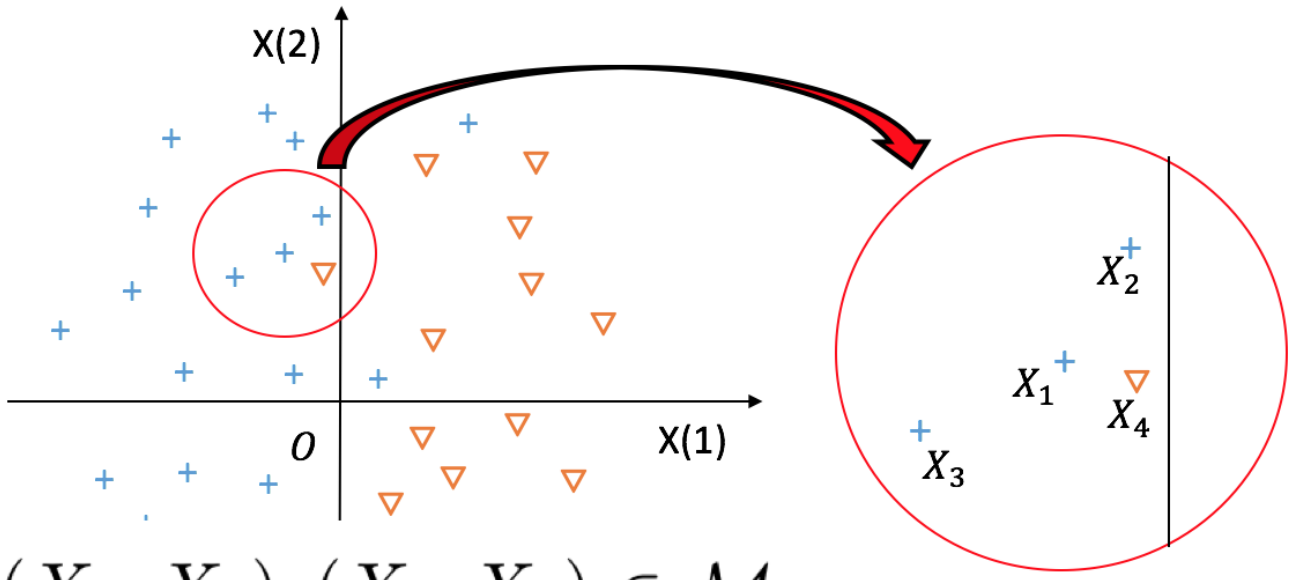
- Dada Δ, considere la siguiente función de costos:

$$c_\Delta((x,y),(x',y')) = d^2_\Delta(x,x')\mathbb{I}(y=y') + \infty\mathbb{I}(y\neq y')$$

y el problema de optimización robusta de distribuciones basado en datos que se propone es:

$$\min_{\beta} \sup_{U_\delta(P_n;c_\Delta)} E[l(X,Y,\beta)]$$

- Veamos el siguiente ejemplo.



$$(X_1, X_2), (X_1, X_3) \in \mathcal{M}$$

$$(X_1, X_4) \in \mathcal{N}$$

$$\Lambda = \begin{bmatrix} 1.16 & 0 \\ 0 & 0.04 \end{bmatrix}$$

- Obsérvese que a lo largo de la variable X_1 se encuentra más observaciones en distintas clases (no relacionadas). Luego si queremos minimizar la distancia entre ellas debemos reducir el peso en esta dimensión relativo a X_2.

Contenido

Experimentos Numéricos

- Introducen técnicas de optimización para resolver el problema de forma eficiente.
- Experimentos numéricos con 5 bases reales.

Experimentos Numéricos

		BC	BN	QSAR	Magic	MB	SB
LR	Train	0 ± 0	$.008 \pm .003$	$.026 \pm .008$	$.213 \pm .153$	0 ± 0	0 ± 0
	Test	8.75 ± 4.75	2.80 ± 1.44	35.5 ± 12.8	17.8 ± 6.77	18.2 ± 10.0	14.5 ± 9.04
	Accur	$.762 \pm .061$	$.926 \pm .048$	$.701 \pm .040$	$.668 \pm .042$	$.678 \pm .059$	$.789 \pm .035$
LRL1	Train	$.185 \pm .123$	$.080 \pm .030$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
	Test	$.428 \pm .338$	$.340 \pm .228$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .140$
	Accur	$.929 \pm .023$	$.930 \pm .042$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
DRO-L	Train	$.022 \pm .019$	$.197 \pm .112$	$.402 \pm .039$	$.469 \pm .064$	$.294 \pm .046$	$.166 \pm .031$
	Test	$.126 \pm .034$	$.275 \pm .093$	$.557 \pm .023$	$.571 \pm .043$	$.613 \pm .053$	$.333 \pm .018$
	Accur	$.954 \pm .015$	$.919 \pm .050$	$.733 \pm .0.026$	$.727 \pm .039$	$.714 \pm .032$	$.887 \pm .011$
DRO-NL	Train	$.032 \pm .015$	$.113 \pm .035$	$.339 \pm .044$	$.381 \pm .084$	$.287 \pm .049$	$.195 \pm .034$
	Test	$.119 \pm .044$	$.194 \pm .067$	$.554 \pm .032$	$.576 \pm ..049$	$.607 \pm .060$	$.332 \pm .015$
	Accur	$.955 \pm .016$	$.931 \pm .036$	$.736 \pm .027$	$.730 \pm .043$	$.716 \pm .054$	$.889 \pm .009$
Num Predictors		30	4	30	10	20	56
Train Size		40	20	80	30	30	150
Test Size		329	752	475	9990	125034	2951

TABLE 1. Numerical results for real data sets.