Aplicaciones de Inferencia Robusta de Wasserstein al Aprendizaje de Máquinas

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Octubre de 2017

Contenido

Introducción

- Blanchet, Kang, Murthy (2017). Robust Wasserstein Profile Inference and Applications to ML.
- Muestran dos cosas:
  1. Cómo hacer inferencia de los parámetros de un proceso generador de datos usando la función de Wasserstein.
  2. Cómo racionalizar la elección del parámetro de regularización y lo comparan con validación cruzada.
- Esto último requiere representacion distribucional robusta del problema de regularización en ML y sugiere una forma de inferencia robusta más allá de los problema de regularización.

Introducción

- Blanchet, Kang, Murthy (2017). Robust Wasserstein Profile Inference and Applications to ML.
- Muestran dos cosas:
  1. Cómo hacer inferencia de los parámetros de un proceso generador de datos usando la función de Wasserstein.
  2. Cómo racionalizar la elección del parámetro de regularización y lo comparan con validación cruzada.
- Esto último requiere representacion distribucional robusta del problema de regularización en ML y sugiere una forma de inferencia robusta más allá de los problema de regularización.

Introducción

- Blanchet, Kang, Murthy (2017). Robust Wasserstein Profile Inference and Applications to ML.
- Muestran dos cosas:
  1. Cómo hacer inferencia de los parámetros de un proceso generador de datos usando la función de Wasserstein.
  2. Cómo racionalizar la elección del parámetro de regularización y lo comparan con validación cruzada.
- Esto último requiere representacion distribucional robusta del problema de regularización en ML y sugiere una forma de inferencia robusta más allá de los problema de regularización.

Introducción

- Blanchet, Kang, Murthy (2017). Robust Wasserstein Profile Inference and Applications to ML.
- Muestran dos cosas:
  1. Cómo hacer inferencia de los parámetros de un proceso generador de datos usando la función de Wasserstein.
  2. Cómo racionalizar la elección del parámetro de regularización y lo comparan con validación cruzada.
- Esto último requiere representacion distribucional robusta del problema de regularización en ML y sugiere una forma de inferencia robusta más allá de los problema de regularización.

Introducción

- En esta formulación los datos se usan solo para centrar el conjunto de la distribuciones plausibles como perturbaciones de la distribución empírica.
- La función de costos es dada.
- Blanchet, Kang, Murthy (2017). Data Driven Optimal Transport Cost Selection for Distributional Robust Optimization extienden la teoría para hacer un mejor uso de los datos (elegir una función de costos más apropiada).

Introducción

- En esta formulación los datos se usan solo para centrar el conjunto de la distribuciones plausibles como perturbaciones de la distribución empírica.
- La función de costos es dada.
- Blanchet, Kang, Murthy (2017). Data Driven Optimal Transport Cost Selection for Distributional Robust Optimization extienden la teoría para hacer un mejor uso de los datos (elegir una función de costos más apropiada).

Introducción

- En esta formulación los datos se usan solo para centrar el conjunto de la distribuciones plausibles como perturbaciones de la distribución empírica.
- La función de costos es dada.
- Blanchet, Kang, Murthy (2017). Data Driven Optimal Transport Cost Selection for Distributional Robust Optimization extienden la teoría para hacer un mejor uso de los datos (elegir una función de costos más apropiada).

Introducción: Ejemplo modelo de regresión lineal

- Sea $\{(X_i, Y_i)\}_{i=1,\dots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Suponemos que $Y_i = \beta_*^T X_i + \epsilon_i$.
- Sea $l(x, y, \beta) = (y - \beta^T x)^2$ la función de pérdida cuadrática.
- Definimos la distribución empírica de la muestra como:

$$P_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(x \leq x_i, y \leq y_i)$$

Introducción: Ejemplo modelo de regresión lineal

- Sea $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Suponemos que $Y_i = \beta_*^T X_i + \epsilon_i$.
- Sea $l(x, y, \beta) = (y - \beta^T x)^2$ la función de pérdida cuadrática.
- Definimos la distribución empírica de la muestra como:

$$P_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(x \leq x_i, y \leq y_i)$$

Introducción: Ejemplo modelo de regresión lineal

- Sea $\{(X_i, Y_i)\}_{i=1,\dots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Suponemos que $Y_i = \beta_*^T X_i + \epsilon_i$.
- Sea $l(x, y, \beta) = (y - \beta^T x)^2$ la función de pérdida cuadrática.
- Definimos la distribución empírica de la muestra como:

$$P_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(x \leq x_i, y \leq y_i)$$

Introducción: Ejemplo modelo de regresión lineal

- Sea $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ una muestra i.i.d. $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$.
- Suponemos que $Y_i = \beta_*^T X_i + \epsilon_i$.
- Sea $l(x, y, \beta) = (y - \beta^T x)^2$ la función de pérdida cuadrática.
- Definimos la distribución empírica de la muestra como:

$$P_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(x \leq x_i, y \leq y_i)$$

Introducción: Inferencia

- Para hacer inferencia sobre β_* consideremos la función de Wasserstein.
- Sabemos que dado P, el β óptimo cuando la función de pérdida es cuadrática satisface:

$$E_P[(Y - \beta_*^T X)X] = 0$$

Introducción: Inferencia

- Para hacer inferencia sobre β_* consideremos la función de Wasserstein.
- Sabemos que dado P, el β óptimo cuando la función de pérdida es cuadrática satisface:

$$E_P[(Y - \beta_*^T X)X] = 0$$

Introducción: Inferencia

- Definimos la función de Wasserstein como:

$$R_n(\beta) = inf\{D_c(P, P_n) : E_P[(Y - \beta^T X)X] = 0\}$$

donde c es una función de costo y D_c es la distancia de Wasserstein.

- Sea $U_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$ es el conjunto de incertidumbre o de distribuciones de probabilidad (modelos) plausibles como perturbaciones de P_n.
- Dado P sea $\beta(P)$ tal que $E_P[(Y - \beta^T(P)X)X] = 0$.
- Sea $\Delta_n(\delta) = \{\beta(P) : P \in U_\delta(P_n)\}$, los estimadores plausibles de β.

Introducción: Inferencia

- Definimos la función de Wasserstein como:

$$R_n(\beta) = inf\{D_c(P, P_n) : E_P[(Y - \beta^T X)X] = 0\}$$

  donde c es una función de costo y D_c es la distancia de Wasserstein.
- Sea $U_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$ es el conjunto de incertidumbre o de distribuciones de probabilidad (modelos) plausibles como perturbaciones de P_n.
- Dado P sea $\beta(P)$ tal que $E_P[(Y - \beta^T(P)X)X] = 0$.
- Sea $\Delta_n(\delta) = \{\beta(P) : P \in U_\delta(P_n)\}$, los estimadores plausibles de β.

Introducción: Inferencia

- Definimos la función de Wasserstein como:

$$R_n(\beta) = inf\{D_c(P, P_n) : E_P[(Y - \beta^T X)X] = 0\}$$

  donde c es una función de costo y D_c es la distancia de Wasserstein.
- Sea $U_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$ es el conjunto de incertidumbre o de distribuciones de probabilidad (modelos) plausibles como perturbaciones de P_n.
- Dado P sea $\beta(P)$ tal que $E_P[(Y - \beta^T(P)X)X] = 0$.
- Sea $\Delta_n(\delta) = \{\beta(P) : P \in U_\delta(P_n)\}$, los estimadores plausibles de β.

Introducción: Inferencia

- Definimos la función de Wasserstein como:

$$R_n(\beta) = inf\{D_c(P, P_n) : E_P[(Y - \beta^T X)X] = 0\}$$

  donde c es una función de costo y D_c es la distancia de Wasserstein.
- Sea $U_\delta(P_n) = \{P : D_c(P, P_n) \leq \delta\}$ es el conjunto de incertidumbre o de distribuciones de probabilidad (modelos) plausibles como perturbaciones de P_n.
- Dado P sea $\beta(P)$ tal que $E_P[(Y - \beta^T(P)X)X] = 0$.
- Sea $\Delta_n(\delta) = \{\beta(P) : P \in U_\delta(P_n)\}$, los estimadores plausibles de β.

Introducción: Inferencia

- Denotemos por $\chi_{1-\alpha}$ el $(1-\alpha)$ percentile de $R_n(\beta_*)$

$$\chi_{1-\alpha} = inf\{z : P(R_n(\beta_*) \leq z) \geq 1-\alpha\}$$

entonces: $\Delta_n(\chi_{1-\alpha})$ es un intervalo de confianza de tamaño $1-\alpha$ de β_*:

$$P(\beta_* \in \Delta_n(\chi_{1-\alpha})) = P(R_n(\beta_*) \leq \chi_{1-\alpha}) = 1-\alpha$$

- La estrategia va ser estudiar $R_n(\beta_*)$ asintóticamente.
- Obsérvese que $\chi_{1-\alpha}$ es el menor valor para el cual β_* es plausible con probabilidad $1-\alpha$.

Introducción: Inferencia

- Denotemos por $\chi_{1-\alpha}$ el $(1-\alpha)$ percentile de $R_n(\beta_*)$

$$\chi_{1-\alpha} = inf\{z : P(R_n(\beta_*) \leq z) \geq 1-\alpha\}$$

entonces: $\Delta_n(\chi_{1-\alpha})$ es un intervalo de confianza de tamaño $1-\alpha$ de β_*:

$$P(\beta_* \in \Delta_n(\chi_{1-\alpha})) = P(R_n(\beta_*) \leq \chi_{1-\alpha}) = 1-\alpha$$

- La estrategia va ser estudiar $R_n(\beta_*)$ asintóticamente.
- Obsérvese que $\chi_{1-\alpha}$ es el menor valor para el cual β_* es plausible con probabilidad $1-\alpha$.

Introducción: Inferencia

- Denotemos por $\chi_{1-\alpha}$ el $(1-\alpha)$ percentile de $R_n(\beta_*)$

$$\chi_{1-\alpha} = inf\{z : P(R_n(\beta_*) \leq z) \geq 1-\alpha\}$$

entonces: $\Delta_n(\chi_{1-\alpha})$ es un intervalo de confianza de tamaño $1-\alpha$ de β_*:

$$P(\beta_* \in \Delta_n(\chi_{1-\alpha})) = P(R_n(\beta_*) \leq \chi_{1-\alpha}) = 1-\alpha$$

- La estrategia va ser estudiar $R_n(\beta_*)$ asintóticamente.
- Obsérvese que $\chi_{1-\alpha}$ es el menor valor para el cual β_* es plausible con probabilidad $1-\alpha$.

Introducción: Inferencia

- Una interpretación de $R_n(\beta_*)$ se obtiene de observar lo siguiente.
- Sea $\mathbb{P}_{opt}(\beta_*) = \{P : E_P[(Y - \beta_*^T X)X] = 0\}$ entonces:

$$R_n(\beta_*) = inf\{D_c(P, P_n) : P \in \mathbb{P}_{opt}\}.$$

- Por lo tanto: $\{P : D_c(P, P_n) \leq R_n(\beta_*)\}$ es el conjunto más pequeño en términos de la métrica de Wasserstein para el cual existe una P consistente con β_*.

Introducción: Inferencia

- Una interpretación de $R_n(\beta_*)$ se obtiene de observar lo siguiente.
- Sea $\mathbb{P}_{opt}(\beta_*) = \{P : E_P[(Y - \beta_*^T X)X] = 0\}$ entonces:

$$R_n(\beta_*) = inf\{D_c(P, P_n) : P \in \mathbb{P}_{opt}\}.$$

- Por lo tanto: $\{P : D_c(P, P_n) \leq R_n(\beta_*)\}$ es el conjunto más pequeño en términos de la métrica de Wasserstein para el cual existe una P consistente con β_*.

Introducción: Inferencia

- Una interpretación de $R_n(\beta_*)$ se obtiene de observar lo siguiente.
- Sea $\mathbb{P}_{opt}(\beta_*) = \{P : E_P[(Y - \beta_*^T X)X] = 0\}$ entonces:

$$R_n(\beta_*) = inf\{D_c(P, P_n) : P \in \mathbb{P}_{opt}\}.$$

- Por lo tanto: $\{P : D_c(P, P_n) \leq R_n(\beta_*)\}$ es el conjunto más pequeño en términos de la métrica de Wasserstein para el cual existe una P consistente con β_*.

Introducción: Regularización

- El estimador LASSO (Least Absolut Shrinkage and Selection) del modelo de regresión lineal es:

$$= min_{\beta \in \mathbb{R}} \{ \sqrt{\frac{1}{n} \sum_{i=1}^{n} l(X_i, Y_i, \beta)} + \lambda \|\beta\|_1 \}$$

  donde λ es el parámetro de regularización.

- Alternativamente:

$$min_{\beta \in \mathbb{R}^d} \{ \sqrt{E_{P_n}[l(X, Y, \beta)]} + \lambda \|\beta\|_1 \}$$

- Por qué regularizar?

Introducción: Regularización

- El estimador LASSO (Least Absolut Shrinkage and Selection) del modelo de regresión lineal es:

$$= min_{\beta \in \mathbb{R}} \{ \sqrt{\frac{1}{n} \sum_{i=1}^{n} l(X_i, Y_i, \beta)} + \lambda \|\beta\|_1 \}$$

  donde λ es el parámetro de regularización.

- Alternativamente:

$$min_{\beta \in \mathbb{R}^d} \{ \sqrt{E_{P_n}[l(X, Y, \beta)]} + \lambda \|\beta\|_1 \}$$

- Por qué regularizar?

Introducción: Regularización

- El estimador LASSO (Least Absolut Shrinkage and Selection) del modelo de regresión lineal es:

$$= min_{\beta \in \mathbb{R}} \{ \sqrt{\frac{1}{n} \sum_{i=1}^{n} l(X_i, Y_i, \beta)} + \lambda \left\| \beta \right\|_1 \}$$

  donde λ es el parámetro de regularización.

- Alternativamente:

$$min_{\beta \in \mathbb{R}^d} \{ \sqrt{E_{P_n}[l(X, Y, \beta)]} + \lambda \left\| \beta \right\|_1 \}$$

- Por qué regularizar?

Regularización y Validación Cruzada

- Regularizamos para reducir la varianza (potencialmente a costa de sesgo) para obetener un menor error de generalización (riesgo).
- En ausencia de muchos datos, elegimos el parámetro de regularización usando validación cruzada.
- Validación cruzada es un estimador del riesgo esperado, no necesarimente del riesgo (condicional a la muestra).

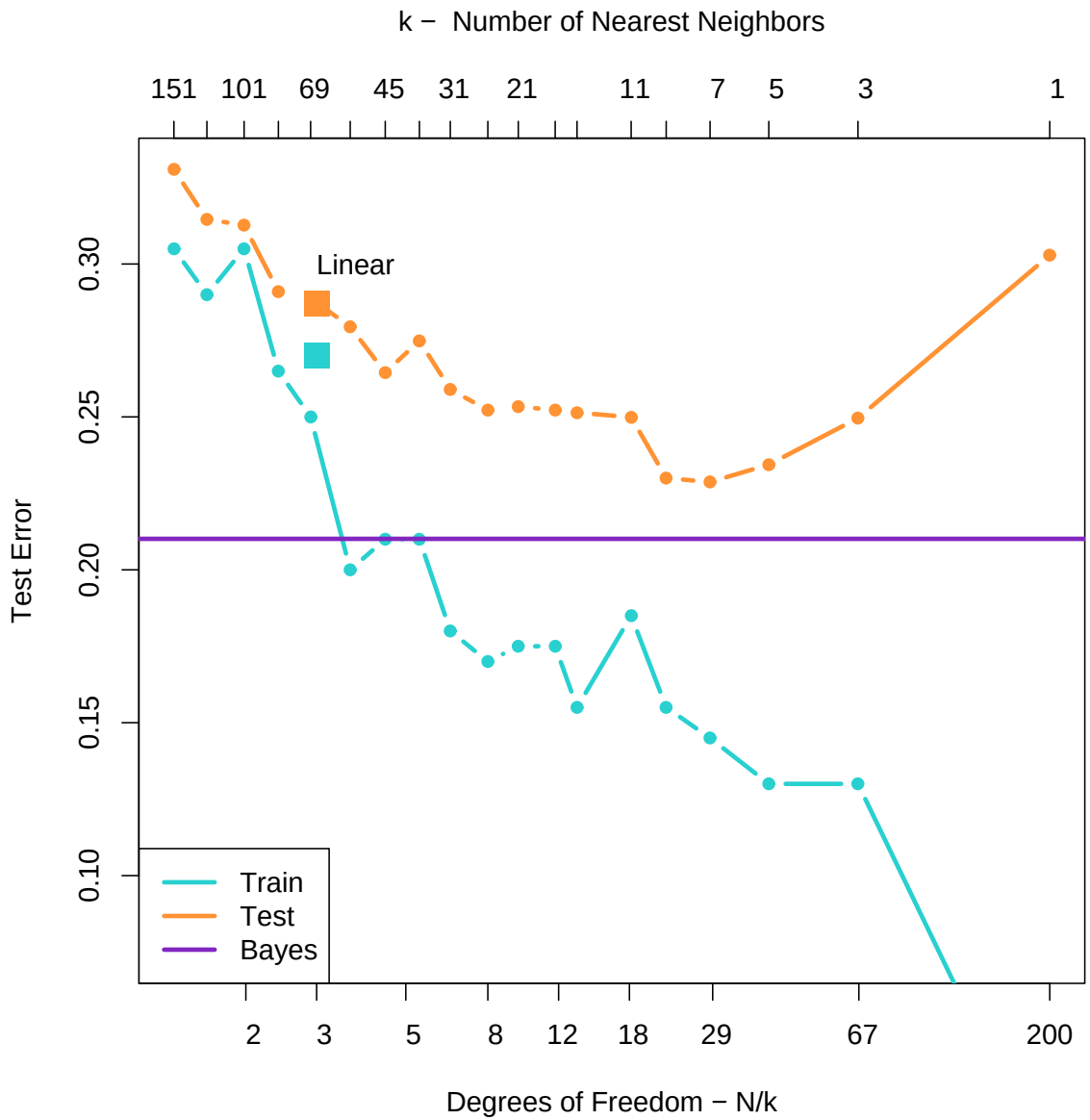
Regularización y Validación Cruzada

- Regularizamos para reducir la varianza (potencialmente a costa de sesgo) para obetener un menor error de generalización (riesgo).
- En ausencia de muchos datos, elegimos el parámetro de regularización usando validación cruzada.
- Validación cruzada es un estimador del riesgo esperado, no necesarimente del riesgo (condicional a la muestra).

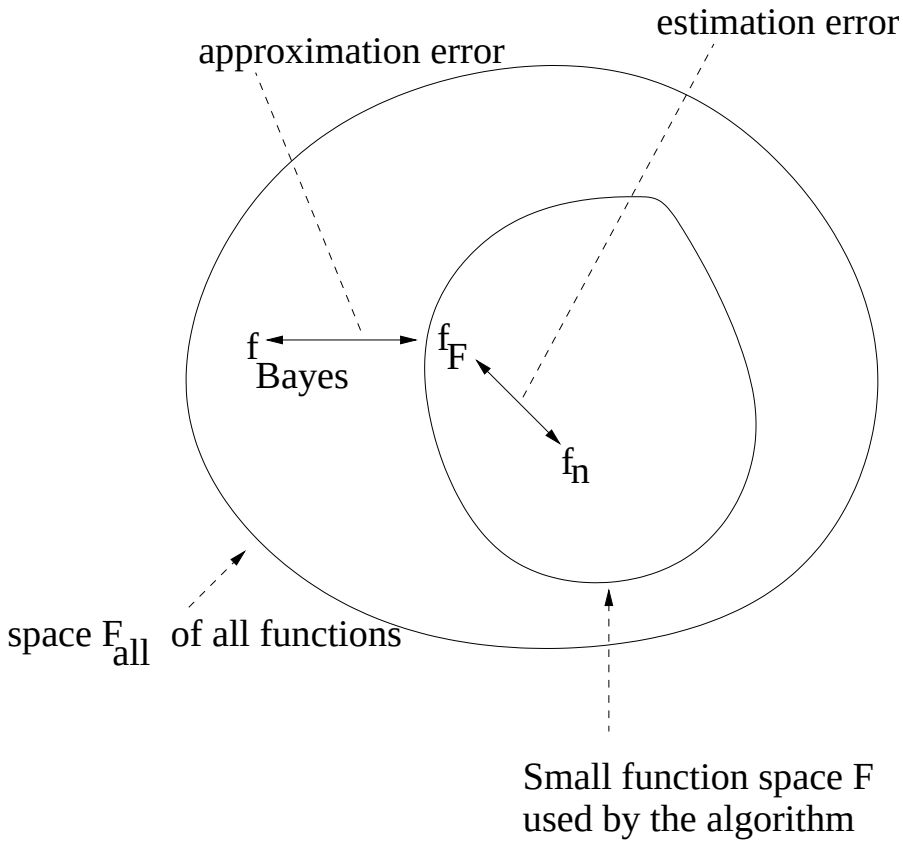
Regularización y Validación Cruzada

- Regularizamos para reducir la varianza (potencialmente a costa de sesgo) para obetener un menor error de generalización (riesgo).
- En ausencia de muchos datos, elegimos el parámetro de regularización usando validación cruzada.
- Validación cruzada es un estimador del riesgo esperado, no necesarimente del riesgo (condicional a la muestra).

Sesgo vrs. Varianza (riesgo)



Sesgo vrs. Varianza (espacio funciones)



Por qué regularizar?

- Regresión lineal regularizada

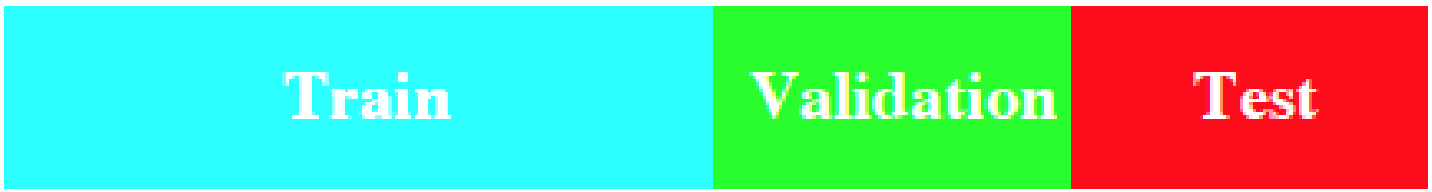
$$min\{\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\|(\beta)\|^2\} \quad (1)$$

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Cómo elegir el parámetro de penalización?

- En ambientes ricos en datos podemos seleccionar modelos (metaparámetros) y estimar el riesgo.



Validación Cruzada

- Cuando no se tiene tanta información se puede hacer validación cruzada.
- Esto permite estimar el riesgo esperado y seleccionar modelos.
- Validación cruzada de K muestras:
  1. Dividir en K muestras aleatorias la muestra originaL de tamañ N. Dada la muestra k se entrena el modelo sin los datos de esta muestra y se estima el error en esa muestra. El promedio de los errores es la estimación del riesgo esperado.
  2. Cuando $K = N$ se conoce como *leave out one cross validation*. En este caso el modelo estimado puede tener una varianza alta pero el sesgo en la estimación del error esperado es bajo.
- Validación cruzada no es un estimador del riesgo sino del riesgo esperado.

Validación Cruzada

- Cuando no se tiene tanta información se puede hacer validación cruzada.
- Esto permite estimar el riesgo esperado y seleccionar modelos.
- Validación cruzada de K muestras:
  1. Dividir en K muestras aleatorias la muestra originaL de tamañ N. Dada la muestra k se entrena el modelo sin los datos de esta muestra y se estima el error en esa muestra. El promedio de los errores es la estimación del riesgo esperado.
  2. Cuando $K = N$ se conoce como *leave out one cross validation*. En este caso el modelo estimado puede tener una varianza alta pero el sesgo en la estimación del error esperado es bajo.
- Validación cruzada no es un estimador del riesgo sino del riesgo esperado.

Sesgo, Varianza y Riesgo Esperado

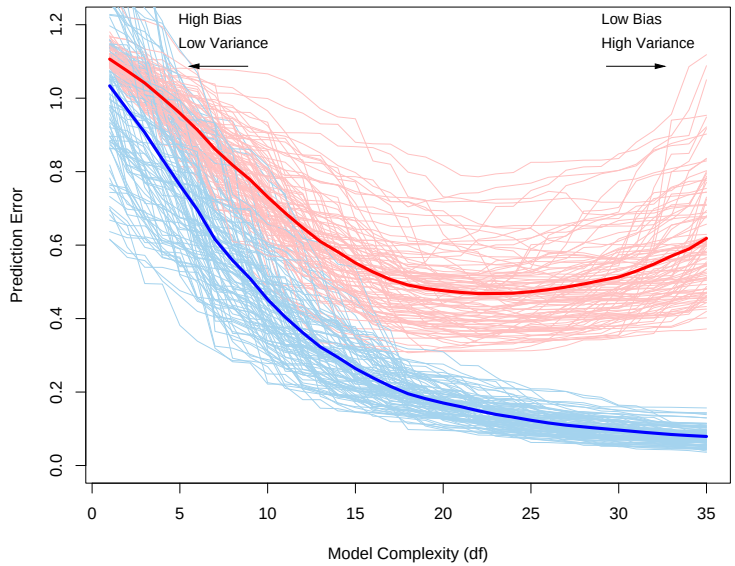


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

Representación Robusta del Problema LASSO

- El problema del estimador LASSO se puede escribir como:

$$min_{\beta \in \mathbb{R}^d} \{ \sqrt{E_{P_n}[l(X, Y, \beta)]} + \lambda \|\beta\|_1 \}^2$$
$$= \min_{\beta \in R^d} \max_{P \in U_\delta(P_n)} E_P[l(X, Y, \beta)]$$

  donde $\delta = \lambda^{\frac{1}{2}}$.

- Intuitivamente: minimizar la pédida en el peor caso (i.e. estrategia minmax).

Representación Robusta del Problema LASSO

- El problema del estimador LASSO se puede escribir como:

$$min_{\beta \in \mathbb{R}^d} \{ \sqrt{E_{P_n}[l(X, Y, \beta)]} + \lambda \|\beta\|_1 \}^2$$
$$= \min_{\beta \in R^d} \max_{P \in U_\delta(P_n)} E_P[l(X, Y, \beta)]$$

  donde $\delta = \lambda^{\frac{1}{2}}$.
- Intuitivamente: minimizar la pédida en el peor caso (i.e. estrategia minmax).

Representación Robusta del Problema LASSO

- El problema minmax introducido anteriormente es equivalente a LASSO y el estimador es plausible (esto se hace con la ayuda del teorema minmax).
- Decimos que β_* es plausible con $(1-\alpha)$ de confianza si δ es lo suficientemente grande para que $\beta_* \in \Delta_n(\delta)$ con probabilidad $1-\alpha$:

$$P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)$$

- El criterio de selección de δ es: escoger δ lo más pequeño posible de tal form que β_* sea plausible con $(1-\alpha)$ de confianza:

$$inf\{\delta : P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)\}$$

- De la discucón anterior sobre inferencia se sigue que $\delta = \chi_{1-\alpha}$.

Representación Robusta del Problema LASSO

- El problema minmax introducido anteriormente es equivalente a LASSO y el estimador es plausible (esto se hace con la ayuda del teorema minmax).
- Decimos que β_* es plausible con $(1-\alpha)$ de confianza si δ es lo suficientemente grande para que $\beta_* \in \Delta_n(\delta)$ con probabilidad $1-\alpha$:

$$P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)$$

- El criterio de selección de δ es: escoger δ lo más pequeño posible de tal form que β_* sea plausible con $(1-\alpha)$ de confianza:

$$inf\{\delta : P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)\}$$

- De la discucón anterior sobre inferencia se sigue que $\delta = \chi_{1-\alpha}$.

Representación Robusta del Problema LASSO

- El problema minmax introducido anteriormente es equivalente a LASSO y el estimador es plausible (esto se hace con la ayuda del teorema minmax).
- Decimos que β_* es plausible con $(1-\alpha)$ de confianza si δ es lo suficientemente grande para que $\beta_* \in \Delta_n(\delta)$ con probabilidad $1-\alpha$:

$$P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)$$

- El criterio de selección de δ es: escoger δ lo más pequeño posible de tal form que β_* sea plausible con $(1-\alpha)$ de confianza:

$$inf\{\delta : P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)\}$$

- De la discucón anterior sobre inferencia se sigue que $\delta = \chi_{1-\alpha}$.

Representación Robusta del Problema LASSO

- El problema minmax introducido anteriormente es equivalente a LASSO y el estimador es plausible (esto se hace con la ayuda del teorema minmax).
- Decimos que β_* es plausible con $(1-\alpha)$ de confianza si δ es lo suficientemente grande para que $\beta_* \in \Delta_n(\delta)$ con probabilidad $1-\alpha$:

$$P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)$$

- El criterio de selección de δ es: escoger δ lo más pequeño posible de tal form que β_* sea plausible con $(1-\alpha)$ de confianza:

$$inf\{\delta : P(\{\beta_* \in \Delta_n(\delta)\}) \geq (1-\alpha)\}$$

- De la discucón anterior sobre inferencia se sigue que $\delta = \chi_{1-\alpha}$.

Introducción: Observaciones

- δ no depende de los datos observados (se usan todas las muestras).
- La única parte donde se utilizan los datos observados es en la estimación del β donde sriven para centrar la bola $U_\delta(P_n)$.
- La representación robusta sugiere una forma general de inferencia.

Introducción: Observaciones

- δ no depende de los datos observados (se usan todas las muestras).
- La única parte donde se utilizan los datos observados es en la estimación del β donde sriven para centrar la bola $U_\delta(P_n)$.
- La representación robusta sugiere una forma general de inferencia.

Introducción: Observaciones

- δ no depende de los datos observados (se usan todas las muestras).
- La única parte donde se utilizan los datos observados es en la estimación del β donde sriven para centrar la bola $U_\delta(P_n)$.
- La representación robusta sugiere una forma general de inferencia.

RWPI

- Estimación e inferencia
  1. Estimación: plantear problema como: $E(h(x, y, \theta) = 0)$
  2. Construir función de Wasserstein.
  3. Estudiar (asíntoticamente) función de Wasserstein.
  4. Inferencia: Intervalos de confianza.
- Optimalidad
  1. Representación robusta de problemas de regularización en ML.
  2. Optimalidad: Parámetro de regularización.

LASSO Generalizado

Example (LASSO Generalizado)

Se postula como modelo: $Y_i = \beta_*^T X_i + e_i$ para algún $\beta_* \in \mathbb{R}^d$ y errores $\{e_1, ..., e_n\}$. Si $l(x, y; \beta) = \left(y - \beta^T x\right)^2$ entonces el estimador LASSO generalizado se obtiene como solución:

$$\min_{\beta \in \mathbb{R}^d} \left\{ \sqrt{E_{P_n}\left[l\left(X, Y; \beta\right)\right]} + \lambda \left\|\beta\right\|_p \right\}, \tag{2}$$

para cualquier $p \in [1, \infty)$.

Example (Regresión Logística Generalizada)

Supongamos que tenemos un problema de clasificación binaria $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, con $X_i \in \mathbb{R}^d$, $Y_i \in \{-1, 1\}$ y suponemos que:

$$\log\left(\frac{P(Y_i = 1|X_i = x)}{1 - P(Y_i = 1|X_i = x)}\right) = \beta_*^T x$$

para algún $\beta_* \in \mathbb{R}^d$. En este caso:

$$l(x, y; \beta) = \log\left(1 + \exp(-y \cdot \beta^T x)\right),$$

y el problema de estimación es:

$$\min_{\beta \in \mathbb{R}^d} \left\{ E_{P_n}[l(X, Y; \beta)] + \lambda \|\beta\|_p \right\}, \tag{3}$$

para $p \in [1, \infty)$.

Contenido

Costo de Transporte Óptimo y Distancia de Wasserstein

- Sea $c : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$ tal que $c(u, w) = 0$ si y solo si $u = w$.
- Dadas dos distribuciones P, Q en R^m la discrepancia (o transporte óptimo entre P y Q):

$$D_c(P, Q) = inf\{E_\pi[c(U, W)] : \pi \in P(\mathbb{R}^m \times \mathbb{R}^m), \pi_U = P, \pi_W = Q\}$$

- Esta formulación del costo óptimo de transporte corresponde a la formulación de Kantorovich (usando planes de transporte).
- La formulación de Monge es más natural (usando mapas de transporte) pero es un caso particular de esta. ver

Costo de Transporte Óptimo y Distancia de Wasserstein

- Sea $c : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$ tal que $c(u, w) = 0$ si y solo si $u = w$.
- Dadas dos distribuciones P, Q en R^m la discrepancia (o transporte óptimo entre P y Q):

$$D_c(P, Q) = inf\{E_\pi[c(U, W)] : \pi \in P(\mathbb{R}^m \times \mathbb{R}^m), \pi_U = P, \pi_W = Q\}$$

- Esta formulación del costo óptimo de transporte corresponde a la formulación de Kantorovich (usando planes de transporte).
- La formulación de Monge es más natural (usando mapas de transporte) pero es un caso particular de esta. ver

Costo de Transporte Óptimo y Distancia de Wasserstein

- Sea $c : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$ tal que $c(u, w) = 0$ si y solo si $u = w$.
- Dadas dos distribuciones P, Q en R^m la discrepancia (o transporte óptimo entre P y Q):

$$D_c(P, Q) = inf\{E_\pi[c(U, W)] : \pi \in P(\mathbb{R}^m \times \mathbb{R}^m), \pi_U = P, \pi_W = Q\}$$

- Esta formulación del costo óptimo de transporte corresponde a la formulación de Kantorovich (usando planes de transporte).
- La formulación de Monge es más natural (usando mapas de transporte) pero es un caso particular de esta. ver

Costo de Transporte Óptimo y Distancia de Wasserstein

- Sea $c : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty]$ tal que $c(u, w) = 0$ si y solo si $u = w$.
- Dadas dos distribuciones P, Q en R^m la discrepancia (o transporte óptimo entre P y Q):

$$D_c(P, Q) = inf\{E_\pi[c(U, W)] : \pi \in P(\mathbb{R}^m \times \mathbb{R}^m), \pi_U = P, \pi_W = Q\}$$

- Esta formulación del costo óptimo de transporte corresponde a la formulación de Kantorovich (usando planes de transporte).
- La formulación de Monge es más natural (usando mapas de transporte) pero es un caso particular de esta. ver

Costo de Transporte Óptimo y Distancia de Wasserstein

- Si c es simétrica y existe $\rho \geq 1$ tal que $c^{\frac{1}{\rho}}$ satisface la desigualdad del triangulo entonces $D_c^{\frac{1}{\rho}}$ es una métrica.

Costo de Transporte Óptimo y Distancia de Wasserstein

Example

Supongamos que $c(u, w) = \|u - w\|_2^2$ donde $\|u - w\|_2$ es la métrica Euclideana. En ese caso ($\rho = 2$):

$$D_c^{\frac{1}{2}}(P, Q) = inf\{\sqrt{E_\pi[\|U - W\|_2^2]}, \pi \in P(\mathbb{R}^m \times \mathbb{R}^m), \pi_U = P, \pi_W = Q\}$$

La distancia de Wasserstein de orden 2.

Costo de Transporte Óptimo y Distancia de Wasserstein

- Bajo ciertas condiciones las métricas de Wasserstein caracterizan convergencia débil en probabilidad.
- *Earth movers distance* es un caso particular de la distancia de W. Se utiliza en procesamiento de imágenes.

Contenido

Estimación y Función de Wasserstein para Inferencia

- Dada una variable aleatoria W en $\mathbb{R}^m$ y $h : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}^r$ el problema es estimar θ_* tal que:

$$E[h(W, \theta_*)] = 0$$

donde el valor esperado se calcula usando la distribución de W en $\mathbb{R}^m$

- Supongamos que tenemos una muestra i.i.d de W, $\{W_1, ..., W_n\}$. La función de Wasserstein se define como:

$$R_n(\theta) = inf\{D_c(P, P_n) : E_P[h(W, \theta)] = 0\}$$

- Usualmente vamos a suponer que $c(u, w) = \|x\|_q^p$, $p, q \geq 1$.

Estimación y Función de Wasserstein para Inferencia

- Dada una variable aleatoria W en $\mathbb{R}^m$ y $h : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}^r$ el problema es estimar θ_* tal que:

$$E[h(W, \theta_*)] = 0$$

donde el valor esperado se calcula usando la distribución de W en $\mathbb{R}^m$

- Supongamos que tenemos una muestra i.i.d de W, $\{W_1, ..., W_n\}$. La función de Wasserstein se define como:

$$R_n(\theta) = inf\{D_c(P, P_n) : E_P[h(W, \theta)] = 0\}$$

- Usualmente vamos a suponer que $c(u, w) = \|x\|_q^p$, $p, q \geq 1$.

Estimación y Función de Wasserstein para Inferencia

- Dada una variable aleatoria W en $\mathbb{R}^m$ y $h : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}^r$ el problema es estimar θ_* tal que:

$$E[h(W, \theta_*)] = 0$$

donde el valor esperado se calcula usando la distribución de W en $\mathbb{R}^m$

- Supongamos que tenemos una muestra i.i.d de W, $\{W_1, ..., W_n\}$. La función de Wasserstein se define como:

$$R_n(\theta) = inf\{D_c(P, P_n) : E_P[h(W, \theta)] = 0\}$$

- Usualmente vamos a suponer que $c(u, w) = \|x\|_q^p$, $p, q \geq 1$.

Ejemplos Estimación

Example (Valor Esperado)

Sea $h(w, \theta) = w - \theta$

Example (Regresión Lineal)

Sea $W = (X, Y)$ y $h(x, y, \theta) = (y - \theta^T x)x$.

Example (Regresión Logística)

$$h(x, y; \theta) = \frac{-yx}{(1 + \exp(y \cdot \theta^T x))},$$

Función de Wasserstein para Inferencia

- El objetivo es estudiar la convergencia en distribución de $n^{\frac{\rho}{2}} R_n(\theta_*; \rho)$ a un $\bar{R}(\rho)$, donde ρ es el parámetro de la función de costo.
- Esta distribución asintótica es la base para hacer inferencia.

Función de Wasserstein para Inferencia

- Sea η_α el $(1-\alpha)$ percentil de $\bar{R}(\rho)$
- Entonces:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) = P(n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \leq \eta_\alpha)$$

- Si $n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \to \bar{R(\rho)}$ entonces podemos construir intervalos de confianza basados en:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) \approx P(\bar{R(\rho)} \leq \eta_\alpha) = 1 - \alpha$$

- Para estudiar el comportamiento asintótico de la función de Wasserstein se usa el problema dual.

Función de Wasserstein para Inferencia

- Sea η_α el $(1-\alpha)$ percentil de $\bar{R}(\rho)$
- Entonces:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) = P(n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \leq \eta_\alpha)$$

- Si $n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \to \bar{R(\rho)}$ entonces podemos construir intervalos de confianza basados en:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) \approx P(\bar{R(\rho)} \leq \eta_\alpha) = 1-\alpha$$

- Para estudiar el comportamiento asintótico de la función de Wasserstein se usa el problema dual.

Función de Wasserstein para Inferencia

- Sea η_α el $(1-\alpha)$ percentil de $\bar{R}(\rho)$
- Entonces:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) = P(n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \leq \eta_\alpha)$$

- Si $n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \to \bar{R(\rho)}$ entonces podemos construir intervalos de confianza basados en:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) \approx P(\bar{R(\rho)} \leq \eta_\alpha) = 1 - \alpha$$

- Para estudiar el comportamiento asintótico de la función de Wasserstein se usa el problema dual.

Función de Wasserstein para Inferencia

- Sea η_α el $(1-\alpha)$ percentil de $\bar{R}(\rho)$
- Entonces:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) = P(n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \leq \eta_\alpha)$$

- Si $n^{\frac{\rho}{2}} R_n(\theta_*; \rho) \to \bar{R(\rho)}$ entonces podemos construir intervalos de confianza basados en:

$$P(\theta_* \in \Delta(\frac{\eta_\alpha}{n^{\frac{\rho}{2}}})) \approx P(\bar{R(\rho)} \leq \eta_\alpha) = 1 - \alpha$$

- Para estudiar el comportamiento asintótico de la función de Wasserstein se usa el problema dual.

Formulación Dual

- Reescribamos $R_n(\theta)$ como:

$$R_n(\theta) = inf_{\pi \in P(R^m \times R^n)}\{E_\pi[c(U,W)] : E_\pi[h(U,\theta)] = 0, \pi_W = P_n\}$$

$$= inf_{\pi \in P(R^m \times R^n)}\{E_\pi c(U,W) : E_\pi[h(U,\theta)] = 0, \pi(\mathbf{1}(W = W_i)) = \frac{1}{n},$$
$$i = 1, ..., n\}.$$

que es un problema de momentos.

- Este problema admite una formulación dual como un programa lineal semi-infinito.

Formulación Dual

- Reescribamos $R_n(\theta)$ como:

$$R_n(\theta) = inf_{\pi \in P(R^m \times R^n)}\{E_\pi[c(U,W)] : E_\pi[h(U,\theta)] = 0, \pi_W = P_n\}$$

$$= inf_{\pi \in P(R^m \times R^n)}\{E_\pi c(U,W) : E_\pi[h(U,\theta)] = 0, \pi(\mathbf{1}(W = W_i)) = \frac{1}{n},$$
$$i = 1, ..., n\}.$$

  que es un problema de momentos.
- Este problema admite una formulación dual como un programa lineal semi-infinito.

Formulación Dual

Proposición

$$R_n(\theta) = sup_{\lambda \in R^r}\{-\frac{1}{n}\sum_{i=1}^{n} \sup_{u \in R^m}\{\lambda^T h(u,\theta) - c(u, W_i)\}\}$$

Ejemplo

Example (Valor esperado)

Supongamos $c(u, w) =\mid u - w \mid^{\rho}, \rho > 1$. Si θ está en el interior del soporte de W podemos escribir:

$$R_n(\theta, \rho) = sup_{\lambda \in \mathbb{R}}\{\frac{-\lambda}{n}\sum_{i=1}^{n}(W_i - \theta) -$$

$$\frac{1}{n}\sum_{i=1}^{n} sup_{\mu \in \mathbb{R}}\{\lambda(u - W_i) - \mid W_i - u \mid^{p}\}$$

Reemplazando el argumento que maximiza el término de la forma: $\lambda\Theta - \mid \Theta \mid^{\rho}$ se obtiene:

$$R_n(\theta, \rho) =\mid \frac{1}{n}\sum_{i=1}^{n}(W_i - \theta) \mid^{\rho}$$

Ejemplo continuación

Example (Valor esperado - continuación)

Si la varianza de W, σ_W^2 es finita se obtiene:

$$\bar{R}(\rho) \sim \sigma_W^{\rho} \mid N(0,1) \mid^{\rho}$$

El resultado también es válido cuando $\rho = 1$.

- Para el modelo lineal y la regresión logístca no se obtienen representaciones tan sencillas de la función de Wesserstein (aún usando el teorema de dualidad), solo resultados asintóticos.

Contenido

Teoría Asintótica

- Con hipótesis adicionales se pueden obtener una cota superior estocástica a la convergencia en distribución de la función de Wasserstein.
- Estas hipótesis adicionales son importantes para las dos aplicaciones.
- En particular: función de costos igual a la métrica de Wasserstein, existe una solución θ_* que resuleve el problema de estimación, etc.

Teoría Asintótica

- Con hipótesis adicionales se pueden obtener una cota superior estocástica a la convergencia en distribución de la función de Wasserstein.
- Estas hipótesis adicionales son importantes para las dos aplicaciones.
- En particular: función de costos igual a la métrica de Wasserstein, existe una solución θ_* que resuleve el problema de estimación, etc.

Teoría Asintótica

- Con hipótesis adicionales se pueden obtener una cota superior estocástica a la convergencia en distribución de la función de Wasserstein.
- Estas hipótesis adicionales son importantes para las dos aplicaciones.
- En particular: función de costos igual a la métrica de Wasserstein, existe una solución θ_* que resuleve el problema de estimación, etc.

Teoría Asintótica: Aplicaciones

- Las aplicaciones se desarrollan de la siguiente forma:
  1. Regresión Lineal (pérdida $h = (x, y, \beta) = (y - \beta^T x)x$).
  2. LASSO Generalizado.
  3. Regresión Logística (pérdida log-exponencial: $h = (x, y, \beta) = \frac{-yx}{1+exp(y\beta^T x)}$).
  4. Regresión Logístisa Regularizada.
- El estudio del comportamiento asintótico de la función de Wasserstein en el caso de la regresión lineal y logística es lo esencial.

Teoría Asintótica: Aplicaciones

- Las aplicaciones se desarrollan de la siguiente forma:
  1. Regresión Lineal (pérdida $h = (x, y, \beta) = (y - \beta^T x)x)$.
  2. LASSO Generalizado.
  3. Regresión Logística (pérdida log-exponencial: $h = (x, y, \beta) = \frac{-yx}{1+exp(y\beta^T x)}$).
  4. Regresión Logístisa Regularizada.
- El estudio del comportamiento asintótico de la función de Wasserstein en el caso de la regresión lineal y logística es lo esencial.

Teoría Asintótica: Aplicaciones

- Las aplicaciones se desarrollan de la siguiente forma:
  1. Regresión Lineal (pérdida $h = (x, y, \beta) = (y - \beta^T x)x)$.
  2. LASSO Generalizado.
  3. Regresión Logística (pérdida log-exponencial: $h = (x, y, \beta) = \frac{-yx}{1+exp(y\beta^T x)}$).
  4. Regresión Logístisa Regularizada.
- El estudio del comportamiento asintótico de la función de Wasserstein en el caso de la regresión lineal y logística es lo esencial.

Teoría Asintótica

- $\bar{R}(\rho)$ es una cota superior estocástica si para toda función no decreciente f:

$$limsup_{n\to\infty} E(f(n^{\frac{\rho}{2}} R_n(\theta,\rho)) \leq_D E[f(\bar{R}(\rho))]$$

- Una definición análoga aplica para cotas inferiores. Cuando ambas cotas son iguales se obtiene convergencia en distribución.

Teoría Asintótica

- $\bar{R}(\rho)$ es una cota superior estocástica si para toda función no decreciente f:

$$limsup_{n\to\infty} E(f(n^{\frac{\rho}{2}} R_n(\theta,\rho)) \leq_D E[f(\bar{R}(\rho))]$$

- Una definición análoga aplica para cotas inferiores. Cuando ambas cotas son iguales se obtiene convergencia en distribución.

Teorema

Cuando $n \to \infty$,

$$n^{\rho/2} R_n(\theta_*; \rho) \lesssim_D \bar{R}(\rho),$$

para $\rho > 1$,

$$\bar{R}(\rho) := \max_{\zeta \in \mathbb{R}^r} \left\{ \rho \zeta^T H - (\rho - 1) E \left\| \zeta^T D_w h(W, \theta_*) \right\|_p^{\rho/(\rho-1)} \right\},$$

y si $\rho = 1$,

$$\bar{R}(1) := \max_{\zeta : P\left(\left\| \zeta^T D_w h(W, \theta_*) \right\|_p > 1 \right) = 0} \{ \zeta^T H \}.$$

En ambos casos $H \sim \mathcal{N}(\mathbf{0}, Cov[h(W, \theta_*)])$, *y* $Cov[h(W, \theta_*)] = E\left[h(W, \theta_*) h(W, \theta_*)^T \right]$.

Aplicación Regresión Lineal

Sea H_0 la hipótesis nula que: $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ son i.i.d y vienen del modelo $Y = \beta_*^T X + e$, donde el término de error e tien media cero, varianza σ^2, y es independiente de X. Sea $\Sigma = \text{Cov}[X]$.

Teorema

Usando una función de costo adecuada $\mathcal{D}_c(\cdot)$, para $\beta \in \mathbb{R}^d$, sea:

$$R_n(\beta) = \text{ínf}\left\{D_c(P, P_n) : \ E_P\left[(Y - \beta^T X)X\right] = \mathbf{0}\right\}.$$

Entonces bajo la hipótesis nula H_0,

$$nR_n(\beta_*) \Rightarrow L_1 := \ \underset{\xi \in \mathbb{R}^d}{\text{máx}} \left\{2\sigma\xi^T Z - E\left\|e\xi - (\xi^T X)\beta_*\right\|_p^2\right\},$$

cuando $n \to \infty$ y donde $Z \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

Teoría Asintótica: Aplicación Regresión Logística

- Una estrategia similar se sigue para el caso de la regresión logística.

Contenido

Representación Robusta de Problemas de Aprendizaje de Máquinas

- Queremos mostrar que los ejemplos logísticos y LASSO pueden escribirse de la siguiente forma:

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: \mathcal{D}_c\left(P, P_n\right) \leq \delta} E_P\left[l\left(X, Y; \beta\right)\right] \tag{4}$$

- Si bien estos ejemplos son casos particulares, en principio esta representación puede utilizarse como una forma general de aproximarse al problema de inferencia.

Representación Robusta de Problemas de Aprendizaje de Máquinas

- Queremos mostrar que los ejemplos logísticos y LASSO pueden escribirse de la siguiente forma:

$$\min_{\beta \in \mathbb{R}^d} \sup_{P: \mathcal{D}_c\left(P, P_n\right) \leq \delta} E_P\left[l\left(X, Y; \beta\right)\right] \tag{4}$$

- Si bien estos ejemplos son casos particulares, en principio esta representación puede utilizarse como una forma general de aproximarse al problema de inferencia.

Representación Robusta: Regresión Lineal

Teorema (Representación Robusta LASSO)

$$\min_{\beta\in\mathbb{R}^d} \sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[l(X,Y;\beta)\right] = \min_{\beta\in\mathbb{R}^d} \left(\sqrt{MSE_n(\beta)} + \sqrt{\delta}\ \|\beta\|_p\right)^2,$$

donde $MSE_n(\beta) = E_{P_n}[(Y-\beta^T X)^2] = n^{-1}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ *y* p *es tal que* $1/p + 1/q = 1$.

Representación Robusta: Regresión Lineal

- Una representación similar se obtiene para el modelo de regresión logística

Representación Robusta: Dualidad

Proposición

Sea $c(\cdot)$ nonegativa, semicontinua por debajo. Para $\gamma \geq 0$ y $I(x, y; \beta)$ semicontinuas por encima en (x, y) para cada β, sea

$$\phi_\gamma(X_i, Y_i; \beta) = \sup_{u \in \mathbb{R}^d,\ v \in \mathbb{R}} \left\{ I(u, v; \beta) - \gamma c\big((u, v), (X_i, Y_i)\big) \right\}. \quad (5)$$

Luego:

$$\sup_{P:\ \mathcal{D}_c(P, P_n) \leq \delta} E_P\big[I(X, Y; \beta)\big] = \min_{\gamma \geq 0} \left\{ \gamma\delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(X_i, Y_i; \beta) \right\}.$$

Contenido

Regularización Óptima de Problemas de Aprendizaje de Máquinas

- Sea $\mathcal{U}_\delta(P_n)$ el conjunto de incertidumbre: $\mathcal{U}_\delta(P_n) = \{P : \mathcal{D}_c(P, P_n) \leq \delta\}$, y β_* el verdadero parámetro (en cualquiera de los modelos ML)
- La convexidad de $l(x, y; \beta)$, como función de β, implica que:

$$\mathcal{P}_{opt}(\beta) := \left\{ P \in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}\right) : E_P\left[D_\beta l(X, Y; \beta_*)\right] = \mathbf{0} \right\}$$

Regularización Óptima de Problemas de Aprendizaje de Máquinas

- β_* es plausible para un δ dado si,

$$\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing.$$

- β_* es plausible con confianza: least $1-\alpha$ si,

$$P\left(\mathcal{P}_{opt}(\beta_*) \cap \mathcal{U}_\delta(P_n) \neq \varnothing\right) \geq 1-\alpha.$$

Regularización: Solución Representación Robusta es Plausible

Lemma

Bajo ciertas condiciones $E\|X\|_2^2 < \infty$, *tenemos que:*

$$\inf_{\beta \in \mathbb{R}^d} \sup_{P \in \mathcal{U}_\delta(P_n)} E_P\left[l(X, Y; \beta)\right] = \sup_{P \in \mathcal{U}_\delta(P_n)} \inf_{\beta \in \mathbb{R}^d} E_P\left[l(X, Y; \beta)\right]. \tag{6}$$

- Este lemma permite demostrar que el estimador de la representación robusta es plausible.

Regularización: Aplicación Regresión Lineal

- Los siguiente pasos resumen la estrategia para elegir el parámetro de regularización para el caso de regresión lineal.
  1) Muestrear Z de $\mathcal{N}(\mathbf{0}, \Sigma)$ para estimar el $1-\alpha$ percentil de la variable aleatoria L_1 sea $\hat{\eta}_{1-\alpha}$ el percentil estimado. Para obtener relaizaciones de L_1 hay que resolver el problema de optimización para cada realiazación de Z. Si $\Sigma = \text{Cov}[X]$ no se se usa el estimador muestral $\text{Cov}[X]$ de Σ.
  2) Elegir λ como:

$$\lambda = \sqrt{\delta} = \sqrt{\hat{\eta}_{1-\alpha}/n}.$$

Contenido

Ejemplos Numéricos

Example (Simulación)

Sea $Y = 3X_1 + 2X_2 + 1{,}5X_4 + e$, $X = (X_1, \ldots, X_d)$ se distribuye normal multivariada $\mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma_{k,j} = 0{,}5^{|k-j|}$, e se distribuye normal con media 0 y $\sigma = 10$. El ejemplo ilustra el efecto de usar diferentes valores de d y n. Sean $d = 300, 600$, y $n = 350, 700, 3500, 10000, 1 - \alpha = 0{,}95$. Para los datos de prueba se usa una simulación de $N = 10{,}000$. En cada instancia se estima LASSO usando la metodología RWPI.Se repite 100 veces y se reporta el promedio en base de prueba y entrenamiento MSE en la tabla comparando con OLS y CV.

Ejemplos Numéricos

Training data size, n	Method	Training Error	Test Error	ℓ_1 loss $\|\beta - \beta_*\|_1$	ℓ_2 loss $\|\beta - \beta_*\|_2$
350	RWPI	101,16($\pm$8,11)	122,59($\pm$6,64)	4,08($\pm$0,69)	5,23($\pm$0,76)
	G-LASSO CV	92,23($\pm$7,91)	117,25($\pm$6,07)	3,91($\pm$0,42)	5,02($\pm$1,28)
	OLS	13,95($\pm$2,63)	702,73($\pm$188,05)	31,59($\pm$3,64)	436,19($\pm$50,55)
700	RWPI	101,81($\pm$3,01)	117,96($\pm$4,80)	3,31($\pm$0,40)	4,38($\pm$0,48)
	G-LASSO CV	99,66($\pm$4,64)	115,46($\pm$4,36)	2,96($\pm$0,37)	3,98($\pm$0,66)
	OLS	56,82($\pm$3,94)	178,44($\pm$21,74)	10,99($\pm$0,57)	152,04($\pm$8,25)
3500	RWPI	102,55($\pm$2,39)	108,44($\pm$2,54)	2,18($\pm$0,16)	3,28($\pm$1,66)
	G-LASSO CV	100,74($\pm$2,35)	113,83($\pm$2,33)	2,66($\pm$0,14)	3,91($\pm$2,18)
	OLS	90,37($\pm$2,17)	114,78($\pm$5,50)	3,96($\pm$0,20)	54,67($\pm$3,09)
10000	RWPI	102,12($\pm$8,11)	105,97($\pm$0,88)	1,13($\pm$0,08)	1,63($\pm$0,11)
	G-LASSO CV	100,69($\pm$7,91)	112,82($\pm$0,71)	1,15($\pm$0,07)	1,94($\pm$0,12)
	OLS	95,91($\pm$1,11)	107,74($\pm$2,96)	2,23($\pm$0,10)	30,91($\pm$1,43)

Cuadro: Sparse linear regression for $d = 300$ predictor variables. The training and test mean square errors of RWPI based generalized LASSO regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based generalized LASSO estimator (written as G-LASSO CV)

Ejemplos Numéricos

Example (Diabetes)

Tenemos 64 predictores (incluyendo interaciones) y una variable de respuesta. En cada iteración (se hace 100 veces) se divide la muestra aleatoriamente en 442 observaciones de $n = 142$ observaciones de entreno y $N = 300$ de prueba. Véase la tabla.

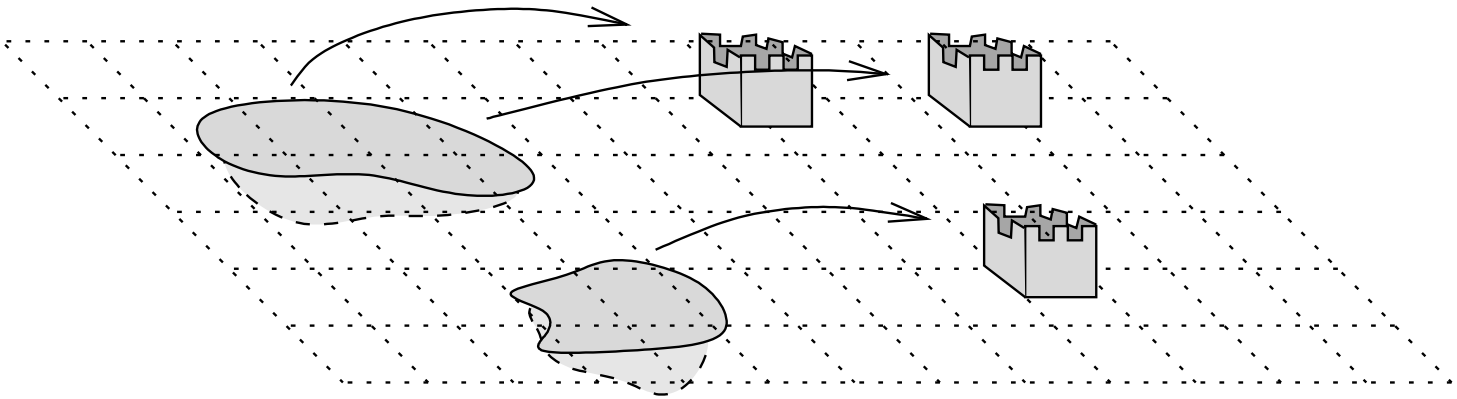
Ejemplos Numéricos

	Training Error	Testing Error
RWPI	0,58($\pm$0,05)	0,60($\pm$0,04)
G-LASSO CV	0,44($\pm$0,06)	0,57($\pm$0,03)
OLS	0,26($\pm$0,05)	1,38($\pm$0,68)

Cuadro: Linear Regression for Diabetes data in Example with 142 training samples and 300 test samples. The training and test mean square errors of RWPI based generalized LASSO regularization parameter selection is compared with ordinary least squares estimator (written as OLS) and cross-validation based generalized LASSO estimator (written as G-LASSO CV).

Contenido

Problema de Transporte Óptimo de Monge



Problema de Transporte Óptimo de Monge

- Sea μ una distribución de probabilidad sobre los Borelianos de $\mathbb{R}^n$.
- Sea $T : \mathbb{R}^n \to \mathbb{R}^n$. Denotamos por T el operador inducido entre distribuciones de porbabilidad en $\mathbb{R}^n$.
- T es un mapa de transporte entre dos distribuciones μ, ν si $T(\mu) = \nu$.
- Sea $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ una función de costo.

Problema de Transporte Óptimo de Monge

- El problema de transporte óptimo de Monge entre μ, ν es:

$$min_{\{T:T(\mu)=\nu\}} \int c(x, T(x)) d\mu(x)$$

- Este problema no esta siempre bien puesto (i.e., puede no existir un mapa de transporte entre las dos distribuciones).
- Puede no existir solución aún cuando existen mapas.

Regresar