Thomas Steinke
Lydia Zakynthinou

Reasoning About Generalization via Conditional Mutual Information

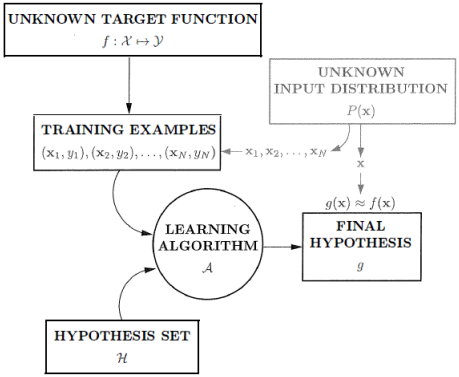
Edward Daniel Soto Mejia
Estudiante de Matemáticas
Facultad de Ciencias

Universidad Nacional de Colombia

August 13, 2021

Bases

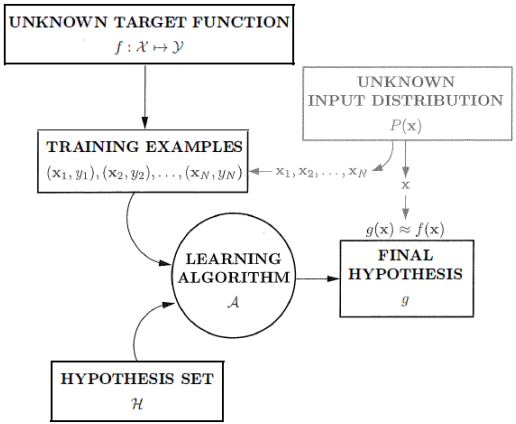
Configuración estándar del aprendizaje estadístico (Valiant [5]).



Elementos del aprendizaje.

Imagen tomada de [1].

Configuración estándar del aprendizaje estadístico



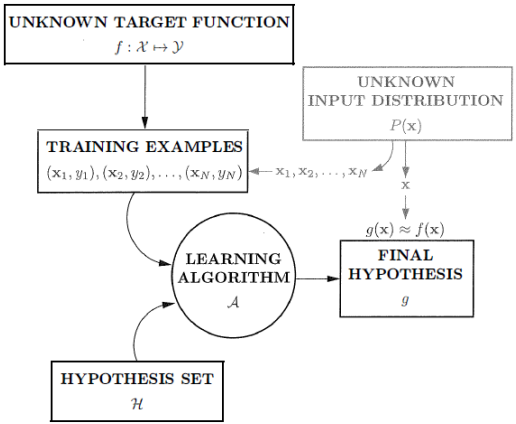
Elementos del aprendizaje.

Imagen tomada de [1].

Elementos

- Distribución D sobre X.
- Función f replicando D.
- Muestra Z de X.
- Conjunto de hipótesis H.
- Algoritmo de aprendizaje A.

Configuración estándar del aprendizaje estadístico



Elementos del aprendizaje.

Imagen tomada de [1].

Elementos

- Distribución D sobre X.
- Función f replicando D.
- Muestra Z de X.
- Conjunto de hipótesis H.
- Algoritmo de aprendizaje A.

Resultado

- Función g de H, $A(Z)$.

¿Como podemos asegurar que $g \approx f$?

Medidas de error

- $E_{in}(g, Z)$ error dentro de la muestra, *"riesgo empírico"*
- $E_{out}(g, D)$ error fuera de la muestra, *riesgo""*.

¿Como podemos asegurar que $g \approx f$?

Medidas de error

- $E_{in}(g, Z)$ error dentro de la muestra, *"riesgo empírico"* **Conocido**
- $E_{out}(g, D)$ error fuera de la muestra, *riesgo"*. **Desconocido**

¿Como podemos asegurar que $g \approx f$?

Medidas de error

- $E_{in}(g, Z)$ error dentro de la muestra, *"riesgo empírico"* **Conocido**
- $E_{out}(g, D)$ error fuera de la muestra, *riesgo"*. **Desconocido**

Generalización

$E_{out}(g, D) \approx E_{in}(g, Z)$

Generalización

Enfoques

- Convergencia uniforme (Vapnik and Chervonenkis [6]), dimensión *VC* y shattering.
- Esquemas de compresión (Littlestone and Warmuth [3]), tamaño del *kernel*.
- Privacidad diferencial (Dwork et al. [2]), privacidad de los datos.

Generalizacion

Problema

Los enfoques no se pueden relacionar o comparar.

Generalizacion

Problema

Los enfoques no se pueden relacionar o comparar.

Propuesta

Mediante teoría de la información agregar un nuevo concepto.

- Conditional mutual information.

Relacionarlo con los enfoques existentes.

Teoría de la información

Información mutua

Dada una muestra (input) Z, un algoritmo A y un modelo (output) $A(Z)$.

$$I(A(Z); Z))$$

Una medida de cuanta información el output $A(Z)$ contiene del input Z.

Teoría de la información

Información mutua

Dada una muestra (input) Z, un algoritmo A y un modelo (output) $A(Z)$.

$$I(A(Z); Z))$$

Una medida de cuanta información el output $A(Z)$ contiene del input Z.

Generalización

Información mutua acotada $\Rightarrow$ generalización.

$$|\mathbb{E}[E_{in} - E_{out}]| \leq \sqrt{\frac{2}{n} \cdot I(A(Z); Z)}$$

Teoría de la información

Propiedades

- La privacidad diferencial es una cota para la información mutua.

Problemas

- Depende del tamaño del espacio de entrada X.
- Puede haber generalización sin que la información mutua este acotada.

Nuevo acercamiento por teoría de la información

Conditional mutual information (CMI)

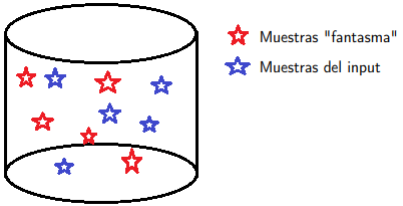
La CMI mide que tan bien podemos "reconocer" un input Z dado un output $A(Z)$.

Nuevo acercamiento por teoría de la información

Conditional mutual information (CMI)

La CMI mide que tan bien podemos "reconocer" un input Z dado un output $A(Z)$.

Idea: Se toma una *"súper muestra"* de tamaño $2n$ de la distribución D, Z'.



Nuevo acercamiento por teoría de la información

Conditional mutual information (CMI)

La *"súper muestra"* es particionada aleatoriamente mediante un conjunto de índices $S = \{0,1\}^n$.

$Z'_{1,0}$	$Z'_{1,1}$
$Z'_{2,0}$	$Z'_{2,1}$
$\vdots$	$\vdots$
$Z'_{n,0}$	$Z'_{n,1}$

$$S = \{0,1,\ldots,1\} \rightarrow Z'_S = \{Z'_{1,0}, Z'_{2,1}, \ldots, Z'_{n,1}\}$$

Nuevo acercamiento por teoría de la información

Conditional mutual information (CMI)

La *"súper muestra"* es particionada aleatoriamente mediante un conjunto de índices $S = \{0,1\}^n$.

$Z'_{1,0}$	$Z'_{1,1}$
$Z'_{2,0}$	$Z'_{2,1}$
$\vdots$	$\vdots$
$Z'_{n,0}$	$Z'_{n,1}$

$$S = \{0,1,\ldots,1\} \rightarrow Z'_S = \{Z'_{1,0}, Z'_{2,1}, \ldots, Z'_{n,1}\}$$

Mediante información mutua queremos saber que tan capaz es el output $A(Z'_S)$ en reconocer información de la partición S.

Nuevo acercamiento por teoría de la información

Conditional mutual information (CMI)

Sea A un algoritmo determinista o aleatorio. Sea D una distribución de probabilidad sobre X y sea $Z' \in X^{nx2}$ una muestra independiente de tamaño $2n$ de D. Sea $S \in \{0,1\}^n$ uniformemente aleatorio, independiente de Z' y A. Definimos $Z'_S \in X^n$ como $(Z'_S)_i = Z'_{i,S_i}$ para todo i, es decir, Z'_S es un subconjunto de Z' indexado por S.

La *conditional mutual information* (Steinke and Zakynthinou [4]) de A respecto a D es:

$$CMI_D(A) := I(A(Z'_S); S|Z')$$

Conditional mutual information (CMI)

Propiedades

- $0 \leq CMI_D(A) \leq n \ln 2$ para todo A y todo D.
- CMI es finita.
- Entropía de Shannon; $CMI_D(A) \leq H(A(Z)) \leq \ln |H|$.
- Si A y B son algoritmos tal que $B(A(\cdot))$ existe, entonces $CMI_D(B(A(\cdot)) = CMI_D(A)$ para todas las distribuciones D.

Generalización y CMI

Teorema 1

Sea A un algoritmo, $E(H, X) \to [0, 1]$ y D una distribución de probabilidad sobre X. Defina $E_{out}(h, D) = \mathbb{E}[E_{in}(h, Z)]$ y $E_{in}(h, Z) = \frac{1}{n} \sum_{i=1}^{n} E(h, z_i)$ para todo $h \in H$ y $Z \in X^n$. Entonces:

Generalización y CMI

Teorema 1

Sea A un algoritmo, $E(H, X) \to [0, 1]$ y D una distribución de probabilidad sobre X. Defina $E_{out}(h, D) = \mathbb{E}[E_{in}(h, Z)]$ y $E_{in}(h, Z) = \frac{1}{n} \sum_{i=1}^{n} E(h, z_i)$ para todo $h \in H$ y $Z \in X^n$. Entonces:

$$|\mathbb{E}[E_{in}(g, Z) - E_{out}(g, D)]| \leq \sqrt{\frac{2}{n} CMI_D(A)} \tag{1}$$

Generalización y CMI

Teorema 1

Sea A un algoritmo, $E(H, X) \to [0, 1]$ y D una distribución de probabilidad sobre X. Defina $E_{out}(h, D) = \mathbb{E}[E_{in}(h, Z)]$ y $E_{in}(h, Z) = \frac{1}{n} \sum_{i=1}^{n} E(h, z_i)$ para todo $h \in H$ y $Z \in X^n$. Entonces:

$$|\mathbb{E}[E_{in}(g, Z) - E_{out}(g, D)]| \leq \sqrt{\frac{2}{n} CMI_D(A)} \tag{1}$$

$$\mathbb{E}[(E_{in}(g, Z) - E_{out}(g, D))^2] \leq \frac{3 \cdot CMI_D(A) + 2}{n} \tag{2}$$

Generalización y CMI

Teorema 1

Sea A un algoritmo, $E(H, X) \to [0, 1]$ y D una distribución de probabilidad sobre X. Defina $E_{out}(h, D) = \mathbb{E}[E_{in}(h, Z)]$ y $E_{in}(h, Z) = \frac{1}{n} \sum_{i=1}^{n} E(h, z_i)$ para todo $h \in H$ y $Z \in X^n$. Entonces:

$$|\mathbb{E}[E_{in}(g, Z) - E_{out}(g, D)]| \leq \sqrt{\frac{2}{n} CMI_D(A)} \tag{2}$$

$$\mathbb{E}[(E_{in}(g, Z) - E_{out}(g, D))^2] \leq \frac{3 \cdot CMI_D(A) + 2}{n} \tag{2}$$

$$\mathbb{E}[E_{out}(g, D)] \leq 2 \cdot \mathbb{E}[E_{in}(g, Z)] + \frac{3}{n} \cdot CMI_D(A) \tag{3}$$

Relación con otros resultados

Convergencia uniforme - Dimensión *VC*

- Dimension *VC* $\longleftrightarrow$ Espacio de hipotesis.
- CMI $\longleftrightarrow$ Algoritmo

Relación con otros resultados

Convergencia uniforme - Dimensión VC

- Dimension $VC \longleftrightarrow$ Espacio de hipotesis.
- CMI $\longleftrightarrow$ Algoritmo

Teorema 2

Sea Xx$\{0,1\}$ el espacio de entrada y $H = \{h, X \to \{0,1\}\}$ un espacio de hipótesis con dimensión VC d. Entonces, existe un algoritmo $A : X^n \to H$ que minimiza el riesgo empírico para $E(H, X) \to [0,1]$ tal que para toda distribucion D.

$$CMI_D(A) \leq O(d \ln n)$$

Relación con otros resultados

Esquemas de compresión

- Esquemas de compresión $\longleftrightarrow$ Algoritmo.
- CMI $\longleftrightarrow$ Algoritmo

Relación con otros resultados

Esquemas de compresión

- Esquemas de compresión $\longleftrightarrow$ Algoritmo.
- CMI $\longleftrightarrow$ Algoritmo

Teorema 3

Sea $A : X^n \to H$ un algoritmo con un esquema de compresión de tamaño k.
$A(Z) \to$ reconstruir un subconjunto de tamaño k de la muestra.
Entonces, para toda distribución D:

$$CMI_D(A) \leq O(k \ln n)$$

Relación con otros resultados

Privacidad diferencial

- Privacidad diferencial $\longleftrightarrow$ Algoritmo y los datos.
- CMI $\longleftrightarrow$ Algoritmo

Relación con otros resultados

Privacidad diferencial

- Privacidad diferencial $\longleftrightarrow$ Algoritmo y los datos.
- CMI $\longleftrightarrow$ Algoritmo

Teorema 4

Sea $A : X^n \to H$ un algoritmo. Si A es $\sqrt{2\varepsilon}$-privacidad diferencial. $\sqrt{2\varepsilon}$ una medida de la perdida de privacidad cuando se manipulan los datos.
Entonces, para toda distribución D:

$$CMI_D(A) \leq \varepsilon n$$

¿Qué sigue?

Problemas

- No garantiza una alta probabilidad, δ.
- Aprendizaje adaptivo.
- Estabilidad uniforme.

¿Qué sigue?

Problemas

- No garantiza una alta probabilidad, δ.
- Aprendizaje adaptivo.
- Estabilidad uniforme.

Trabajo

- Mejorar las cotas y encontrar otras.

Referencias I

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*. Vol. 4. AMLBook New York, NY, USA: 2012.

[2] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284. ISBN: 978-3-540-32732-5.

[3] Nick Littlestone and Manfred Warmuth. "Relating data compression and learnability". In: (1986).

[4] Thomas Steinke and Lydia Zakynthinou. "Reasoning about generalization via conditional mutual information". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3437–3452.

Referencias II

[5] Leslie G Valiant. "A theory of the learnable". In: *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[6] Vladimir N Vapnik and A Ya Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In: *Measures of complexity*. Springer, 2015, pp. 11–30.