Aprendizaje Estadístico II: Consistencia, Minimización del Riesgo Empírico y Cotas

June de 2017

Aprendizaje Estadístico: Ejes de la teoría

- El problema de consistencia.
- Minimización del riesgo empírico.
- Capacidad y Cotas

Contenido

Consistencia

- Diferentes conceptos de consistencia:
  1. Decimos que una máquina de apredizaje M es consistente con respecto a P y $\mathbb{F}_0$ si:

$$P(R(f_n) - R(f_{\mathbb{F}_0}) > \epsilon) \to 0 \tag{1}$$

     cuando la muestra tiende a infinito.[1]
  2. Decimos que es Bayes consistente si M es consistente con respecto a P y $\mathbb{F}$.
  3. Es universal con respecto a $\mathbb{F}_0$ (Bayes universal) si es consistente con respecto a P y $\mathbb{F}_0$ ($\mathbb{F}$) para todo P.

[1] Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^{\infty}$. Esta convergencia es conocida como convergencia en probabilidad y sería más preciso hablar de consistencia débil. Cuando la convergencia es casi en todas partes, hablamos de convergencia fuerte.

Consistencia

- Diferentes conceptos de consistencia:
  1. Decimos que una máquina de apredizaje M es consistente con respecto a P y $\mathbb{F}_0$ si:

$$P(R(f_n) - R(f_{\mathbb{F}_0}) > \epsilon) \to 0 \tag{1}$$

     cuando la muestra tiende a infinito.[1]
  2. Decimos que es Bayes consistente si M es consistente con respecto a P y $\mathbb{F}$.
  3. Es universal con respecto a $\mathbb{F}_0$ (Bayes universal) si es consistente con respecto a P y $\mathbb{F}_0$ ($\mathbb{F}$) para todo P.

[1] Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^{\infty}$. Esta convergencia es conocida como convergencia en probabilidad y sería más preciso hablar de consistencia débil. Cuando la convergencia es casi en todas partes, hablamos de convergencia fuerte.

Consistencia

- Diferentes conceptos de consistencia:
  1. Decimos que una máquina de apredizaje M es consistente con respecto a P y $\mathbb{F}_0$ si:

$$P(R(f_n) - R(f_{\mathbb{F}_0}) > \epsilon) \to 0 \tag{1}$$

     cuando la muestra tiende a infinito.[1]
  2. Decimos que es Bayes consistente si M es consistente con respecto a P y $\mathbb{F}$.
  3. Es universal con respecto a $\mathbb{F}_0$ (Bayes universal) si es consistente con respecto a P y $\mathbb{F}_0$ ($\mathbb{F}$) para todo P.

[1] Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^\infty$. Esta convergencia es conocida como convergencia en probabilidad y sería más preciso hablar de consistencia débil. Cuando la convergencia es casi en todas partes, hablamos de convergencia fuerte.

Ejemplo Consistencia: 1-NN no es Bayes universal

- Sea X distribuida $U[0,1]$ y $P(Y=1\,|X=x)=0{,}9$ (i.e., en particular, Y, X son independientes.)
- Esto caracteriza a P:

$$P(Y=1|X\leq x)=\int_{X\leq x} P(Y=1\,|X=x)dF_X = 0{,}9x \quad (2)$$

Ejemplo Consistencia: 1-NN no es Bayes universal

- El clasificador de Bayes es $f_{Bayes}(x) = 1$ si $P(Y = 1 \,|X = x) \geq 0{,}5$ luego $f_{Bayes} = 1$.
- El error de prueba (riesgo) es:

$$E[L(X, Y, f_{Bayes}(X))] = \int E[L(X, Y, f_{Bayes}(X))|X = x]dF_X$$

$$= \int L(x, 1, f_{Bayes}(x))P(Y = 1|X = x)dF_X$$

$$+ \int L(x, 0, f_{Bayes}(x))P(Y = 0|X = x)dF_X$$

$$= \int (0 + 1(0{,}1))dF_X = 0{,}1$$

Ejemplo Consistencia: 1-NN no es Bayes universal

- El clasificador de Bayes es $f_{Bayes}(x) = 1$ si $P(Y = 1\,|X = x) \geq 0{,}5$ luego $f_{Bayes} = 1$.
- El error de prueba (riesgo) es:

$$E[L(X, Y, f_{Bayes}(X))] = \int E[L(X, Y, f_{Bayes}(X))|X = x]dF_X$$

$$= \int L(x, 1, f_{Bayes}(x))P(Y = 1|X = x)dF_X$$

$$+ \int L(x, 0, f_{Bayes}(x))P(Y = 0|X = x)dF_X$$

$$= \int (0 + 1(0{,}1))dF_X = 0{,}1$$

Ejemplo Consistencia: 1-NN no es Bayes universal

- Ahora seguimos los mismo pasos para calcular el error de prueba del algoritmo $1 - NN$.
- El error de prueba (riesgo) es:

$$E[L(X, Y, f_{1-NN}(X))] = \int E[L(X, Y, f_{1-NN}(X))|X = x]dF_X$$

$$= \int L(x, 1, f_{1-NN}(x))P(Y = 1|X = x)dF_X +$$

$$\int L(x, 0, f_{1-NN}(x))P(Y = 0|X = x)dF_X$$

$$= P(f_{1-NN}(X) \neq 1)0{,}9 + P(f_{1-NN}(X) \neq 0)0{,}1$$

$$\approx (0{,}1)(0{,}9) + (0{,}9)(0{,}1) = 0{,}18$$

- Obsérvese que la última igualdad es apenas aproximada (el tamaño de la muestra es fijo). Cuando la muestra es grande y $P(f_n(x) = 1) \approx 0{,}9$

Ejemplo Consistencia: 1-NN no es Bayes universal

- Ahora seguimos los mismo pasos para calcular el error de prueba del algoritmo $1 - NN$.
- El error de prueba (riesgo) es:

$$E[L(X, Y, f_{1-NN}(X))] = \int E[L(X, Y, f_{1-NN}(X))|X = x]dF_X$$

$$= \int L(x, 1, f_{1-NN}(x))P(Y = 1|X = x)dF_X +$$

$$\int L(x, 0, f_{1-NN}(x))P(Y = 0|X = x)dF_X$$

$$= P(f_{1-NN}(X) \neq 1)0{,}9 + P(f_{1-NN}(X) \neq 0)0{,}1$$

$$\approx (0{,}1)(0{,}9) + (0{,}9)(0{,}1) = 0{,}18$$

- Obsérvese que la última igualdad es apenas aproximada (el tamaño de la muestra es fijo). Cuando la muestra es grande y $P(f_n(x) = 1) \approx 0{,}9$

Ejemplo Consistencia: 1-NN no es Bayes universal

- Ahora seguimos los mismo pasos para calcular el error de prueba del algoritmo $1 - NN$.
- El error de prueba (riesgo) es:

$$E[L(X, Y, f_{1-NN}(X))] = \int E[L(X, Y, f_{1-NN}(X))|X = x]dF_X$$

$$= \int L(x, 1, f_{1-NN}(x))P(Y = 1|X = x)dF_X +$$

$$\int L(x, 0, f_{1-NN}(x))P(Y = 0|X = x)dF_X$$

$$= P(f_{1-NN}(X) \neq 1)0{,}9 + P(f_{1-NN}(X) \neq 0)0{,}1$$

$$\approx (0{,}1)(0{,}9) + (0{,}9)(0{,}1) = 0{,}18$$

- Obsérvese que la última igualdad es apenas aproximada (el tamaño de la muestra es fijo). Cuando la muestra es grande y $P(f_n(x) = 1) \approx 0{,}9$

Ejemplo Consistencia: 1-NN no es Bayes universal

- Por lo tanto no se puede cumplir para esta P que:

$$P(R(f_{1-NN}) - R(f_{Bayes}) > \epsilon) \to 0 \tag{3}$$

Consistencia

- El teorema de Stone (1977) afirma que en el algoritmo del vecino más cercano, si $\frac{k}{n}$ tiende a cero y k, n ambos tienden a infinito entonces $k - NN$ es Bayes universal.
- La máquina o algoritmo de aprendizaje $k - NN$ es Bayes universal.

Consistencia

- El teorema de Stone (1977) afirma que en el algoritmo del vecino más cercano, si $\frac{k}{n}$ tiende a cero y k, n ambos tienden a infinito entonces $k - NN$ es Bayes universal.
- La máquina o algoritmo de aprendizaje $k - NN$ es Bayes universal.

Contenido

Minimización del riesgo empírico

- El problema de clasificación consiste en la minimización del riesgo de tal forma que sea lo más cercano posible al riesgo de Bayes.
- Este problema es imposible de resolver porque no conocemos la distribucion que genera los datos (lo único que suponemos es que los datos son una muestra i.i.d).
- Definimos el problema de minimización del riesgo empírico como:

$$f^{emp}_{\mathbb{F}_0,n} = argmin_{f \in \mathbb{F}_0} R_{emp}[f] \tag{4}$$

obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones y el problema solo hace sentido dada una muestra.
- La esperanza en la capacidad de generalización de esta función de aprendizaje se llama el principio de inducción de minimización del riesgo.
- Vamos a presentar dos formas de acotar la distacia entre el

Minimización del riesgo empírico

- El problema de clasificación consiste en la minimización del riesgo de tal forma que sea lo más cercano posible al riesgo de Bayes.
- Este problema es imposible de resolver porque no conocemos la distribucion que genera los datos (lo único que suponemos es que los datos son una muestra i.i.d).
- Definimos el problema de minimización del riesgo empírico como:

$$f^{emp}_{\mathbb{F}_0,n} = argmin_{f \in \mathbb{F}_0} R_{emp}[f] \tag{4}$$

obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones y el problema solo hace sentido dada una muestra.
- La esperanza en la capacidad de generalización de esta función de aprendizaje se llama el principio de inducción de minimización del riesgo.
- Vamos a presentar dos formas de acotar la distacia entre el

Minimización del riesgo empírico

- El problema de clasificación consiste en la minimización del riesgo de tal forma que sea lo más cercano posible al riesgo de Bayes.
- Este problema es imposible de resolver porque no conocemos la distribucion que genera los datos (lo único que suponemos es que los datos son una muestra i.i.d).
- Definimos el problema de minimización del riesgo empírico como:

$$f^{emp}_{\mathbb{F}_0,n} = argmin_{f \in \mathbb{F}_0} R_{emp}[f] \tag{4}$$

obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones y el problema solo hace sentido dada una muestra.

- La esperanza en la capacidad de generalización de esta función de aprendizaje se llama el principio de inducción de minimización del riesgo.
- Vamos a presentar dos formas de acotar la distacia entre el

Minimización del riesgo empírico

- El problema de clasificación consiste en la minimización del riesgo de tal forma que sea lo más cercano posible al riesgo de Bayes.
- Este problema es imposible de resolver porque no conocemos la distribucion que genera los datos (lo único que suponemos es que los datos son una muestra i.i.d).
- Definimos el problema de minimización del riesgo empírico como:

$$f^{emp}_{\mathbb{F}_0,n} = argmin_{f \in \mathbb{F}_0} R_{emp}[f] \tag{4}$$

obsérvese que $\mathbb{F}_0$ puede ser el conjunto de todas las funciones y el problema solo hace sentido dada una muestra.
- La esperanza en la capacidad de generalización de esta función de aprendizaje se llama el principio de inducción de minimización del riesgo.
- Vamos a presentar dos formas de acotar la distacia entre el

Minimización del riesgo empírico I

- Ahora, la ley de los grandes números implica que si fijamos una función de aprendizaje f entonces:

$$R_{emp}(f) \to R(f) \tag{5}$$

cuando la muestra τ_n se hace cada vez más grande.

- Por la desiguladad de Chernoff:

$$P(\mid R_{emp}(f) - R(f) \mid \geq \epsilon) \leq 2\exp(-2n\epsilon^2) \tag{6}$$

donde n es el tamaño de la muestra.[2].

- Obsérvese que en las dos ecuaciones anteriores f está fijo mientras que en el problema de consistencia la función de aprendizaje f_n depende de la muestra.
- Es precisamente esto lo que puede hacer que el riesgo empírico sea inconsistente.

[2]Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^n$. Por simplicidad la denotamos también por P

Minimización del riesgo empírico I

- Ahora, la ley de los grandes números implica que si fijamos una función de aprendizaje f entonces:

$$R_{emp}(f) \to R(f) \tag{5}$$

  cuando la muestra τ_n se hace cada vez más grande.
- Por la desiguladad de Chernoff:

$$P(\mid R_{emp}(f) - R(f) \mid \geq \epsilon) \leq 2\exp(-2n\epsilon^2) \tag{6}$$

  donde n es el tamaño de la muestra.[2].
- Obsérvese que en las dos ecuaciones anteriores f está fijo mientras que en el problema de consistencia la función de aprendizaje f_n depende de la muestra.
- Es precisamente esto lo que puede hacer que el riesgo empírico sea inconsistente.

[2]Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^n$. Por simplicidad la denotamos también por P

Minimización del riesgo empírico I

- Ahora, la ley de los grandes números implica que si fijamos una función de aprendizaje f entonces:

$$R_{emp}(f) \to R(f) \tag{5}$$

  cuando la muestra τ_n se hace cada vez más grande.
- Por la desiguladad de Chernoff:

$$P(\mid R_{emp}(f) - R(f) \mid \geq \epsilon) \leq 2\exp(-2n\epsilon^2) \tag{6}$$

  donde n es el tamaño de la muestra.[2].
- Obsérvese que en las dos ecuaciones anteriores f está fijo mientras que en el problema de consistencia la función de aprendizaje f_n depende de la muestra.
- Es precisamente esto lo que puede hacer que el riesgo empírico sea inconsistente.

[2]Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^n$. Por simplicidad la denotamos también por P

Minimización del riesgo empírico I

- Ahora, la ley de los grandes números implica que si fijamos una función de aprendizaje f entonces:

$$R_{emp}(f) \to R(f) \tag{5}$$

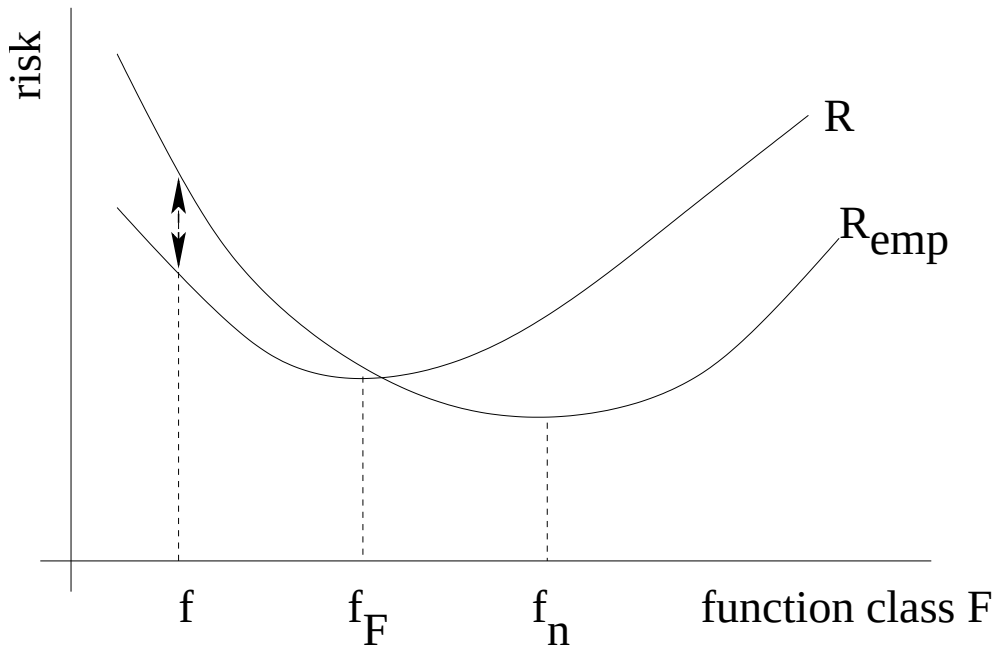
  cuando la muestra τ_n se hace cada vez más grande.
- Por la desiguladad de Chernoff:

$$P(\mid R_{emp}(f) - R(f) \mid \geq \epsilon) \leq 2\exp(-2n\epsilon^2) \tag{6}$$

  donde n es el tamaño de la muestra.[2].
- Obsérvese que en las dos ecuaciones anteriores f está fijo mientras que en el problema de consistencia la función de aprendizaje f_n depende de la muestra.
- Es precisamente esto lo que puede hacer que el riesgo empírico sea inconsistente.

[2]Para los más sofisticados matemáticamente: P define una distribución de probabilidad sobre el espacio $(\Xi \times \Upsilon)^n$. Por simplicidad la denotamos también por P

Minimización del riesgo empírico I



Ejemplo: Inconsistencia de la minimización del riesgo

- Sea $\chi = [0, 1]$ y supongamos que los datos x son generados con la distribución uniforme.
- La variable objetivo se define (de forma determinística) como $f(x)$ es -1 si $x < 0{,}5$ y 1 caso contrario.
- Considere el siguiente clasificador: $f_n(x) = y_i$ si $x = x_i$ para algún i en la muestra τ_n y uno caso contrario.
- Esta función de aprendizaje clasifica perfectamente dentro de muestra. Su riesgo empírico es cero. Puesto que la minimización del riesgo empírico no puede ser negativa entonces este es el clasificador de mínimo riesgo cuando el espacio de funciones admisible es el espacio de todas las funciones.

Ejemplo: Inconsistencia de la minimización del riesgo

- Sea $\chi = [0, 1]$ y supongamos que los datos x son generados con la distribución uniforme.
- La variable objetivo se define (de forma determinística) como $f(x)$ es -1 si $x < 0{,}5$ y 1 caso contrario.
- Considere el siguiente clasificador: $f_n(x) = y_i$ si $x = x_i$ para algún i en la muestra τ_n y uno caso contrario.
- Esta función de aprendizaje clasifica perfectamente dentro de muestra. Su riesgo empírico es cero. Puesto que la minimización del riesgo empírico no puede ser negativa entonces este es el clasificador de mínimo riesgo cuando el espacio de funciones admisible es el espacio de todas las funciones.

Ejemplo: Inconsistencia de la minimización del riesgo

- Sea $\chi = [0, 1]$ y supongamos que los datos x son generados con la distribución uniforme.
- La variable objetivo se define (de forma determinística) como $f(x)$ es -1 si $x < 0{,}5$ y 1 caso contrario.
- Considere el siguiente clasificador: $f_n(x) = y_i$ si $x = x_i$ para algún i en la muestra τ_n y uno caso contrario.
- Esta función de aprendizaje clasifica perfectamente dentro de muestra. Su riesgo empírico es cero. Puesto que la minimización del riesgo empírico no puede ser negativa entonces este es el clasificador de mínimo riesgo cuando el espacio de funciones admisible es el espacio de todas las funciones.

Ejemplo: Inconsistencia de la minimización del riesgo

- Sea $\chi = [0, 1]$ y supongamos que los datos x son generados con la distribución uniforme.
- La variable objetivo se define (de forma determinística) como $f(x)$ es -1 si $x < 0{,}5$ y 1 caso contrario.
- Considere el siguiente clasificador: $f_n(x) = y_i$ si $x = x_i$ para algún i en la muestra τ_n y uno caso contrario.
- Esta función de aprendizaje clasifica perfectamente dentro de muestra. Su riesgo empírico es cero. Puesto que la minimización del riesgo empírico no puede ser negativa entonces este es el clasificador de mínimo riesgo cuando el espacio de funciones admisible es el espacio de todas las funciones.

Ejemplo: Inconsistencia de la minimización del riesgo

- Ahora calculemos el riesgo de este clasificador. Dada una realización de X si $x \geq 0{,}5$ el clasificador acierta. Si $x < 0{,}5$ se equivoca excepto por un conjunto de probabilidad cero (si $x = x_i$ para algún i).
- Luego, el clasificador se equivoca la mitad de las veces y su riesgo es 0,5.
- Se sigue que la máquina de aprendizaje que construye estas funciones de aprendizaje no es consistente.

Ejemplo: Inconsistencia de la minimización del riesgo

- Ahora calculemos el riesgo de este clasificador. Dada una realización de X si $x \geq 0{,}5$ el clasificador acierta. Si $x < 0{,}5$ se equivoca excepto por un conjunto de probabilidad cero (si $x = x_i$ para algún i).
- Luego, el clasificador se equivoca la mitad de las veces y su riesgo es 0,5.
- Se sigue que la máquina de aprendizaje que construye estas funciones de aprendizaje no es consistente.

Ejemplo: Inconsistencia de la minimización del riesgo

- Ahora calculemos el riesgo de este clasificador. Dada una realización de X si $x \geq 0{,}5$ el clasificador acierta. Si $x < 0{,}5$ se equivoca excepto por un conjunto de probabilidad cero (si $x = x_i$ para algún i).
- Luego, el clasificador se equivoca la mitad de las veces y su riesgo es 0,5.
- Se sigue que la máquina de aprendizaje que construye estas funciones de aprendizaje no es consistente.

Minimización del riesgo empírico

- El problema que queremos resolver es cuando el clasificador de mínimo riesgo es consistente.
- La esperanza en que el clasificador de mínimo riesgo sea consistente radica en restringir el espacio donde se minimiza el riesgo.
- La siguiente cota uniforme será la clave:

$$|R(f) - R_{emp}(f)| \leq \sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)|) \tag{7}$$

por lo tanto:

$$P(|R(f_n) - R_{emp}(f_n)| \geq \epsilon) \leq P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \tag{8}$$

Minimización del riesgo empírico

- El problema que queremos resolver es cuando el clasificador de mínimo riesgo es consistente.
- La esperanza en que el clasificador de mínimo riesgo sea consistente radica en restringir el espacio donde se minimiza el riesgo.
- La siguiente cota uniforme será la clave:

$$|R(f) - R_{emp}(f)| \leq \sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)|) \tag{7}$$

por lo tanto:

$$P(|R(f_n) - R_{emp}(f_n)| \geq \epsilon) \leq P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \tag{8}$$

Minimización del riesgo empírico

- El problema que queremos resolver es cuando el clasificador de mínimo riesgo es consistente.
- La esperanza en que el clasificador de mínimo riesgo sea consistente radica en restringir el espacio donde se minimiza el riesgo.
- La siguiente cota uniforme será la clave:

$$|R(f) - R_{emp}(f)| \leq \sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)|) \tag{7}$$

por lo tanto:

$$P(|R(f_n) - R_{emp}(f_n)| \geq \epsilon) \leq P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \tag{8}$$

Minimización del riesgo empírico

- Obsérvese que (por definición de $f_{\mathbb{F}_0}$):

$$\left| R(f^{emp}_{\mathbb{F}_0,n}) - R(f_{\mathbb{F}_0}) \right| = R(f^{emp}_{\mathbb{F}_0,n}) - R(f_{\mathbb{F}_0}) \tag{9}$$

$$= R(f^{emp}_{\mathbb{F}_0,n}) - R_{emp}(f^{emp}_{\mathbb{F}_0,n}) + R_{emp}(f^{emp}_{\mathbb{F}_0,n}) - R_{emp}(f_{\mathbb{F}_0}) \tag{10}$$

$$+ R_{emp}(f_{\mathbb{F}_0}) - R(f_{\mathbb{F}_0}) \tag{11}$$

Minimización del riesgo empírico

- El segundo término es no positivo por lo tanto:

$$\leq R(f^{emp}_{\mathbb{F}_0,n}) - R_{emp}(f^{emp}_{\mathbb{F}_0,n}) + R_{emp}(f_{\mathbb{F}_0}) - R(f_{\mathbb{F}_0}) \tag{12}$$

$$\leq 2 \sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \tag{13}$$

Minimización del riesgo empírico

- Se sigue que:

$$\left| R(f^{emp}_{\mathbb{F}_0,n}) - R(f_{\mathbb{F}_0}) \right| \leq 2 \sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \tag{14}$$

luego

$$P(\left| R(f^{emp}_{\mathbb{F}_0,n}) - R(f_{\mathbb{F}_0}) \right| \geq \epsilon) \leq P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \frac{\epsilon}{2}) \tag{15}$$

Teorema de Vapnik y Chervonenkis

- La máquina de aprendizaje de minimización del riesgo empírico en $\mathbb{F}_0$ es consistente con respecto a P en $\mathbb{F}_0$ si y sólo si:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \to 0 \tag{16}$$

  cuando $n \to \infty$.
- La convergencia anterior se denomina convergencia uniforme.
- Obsérvese que este teorema permite controlar el error de estimación cuando la muestra es grande. Pero nada dice sobre el sesgo.

Contenido

Capacidad y cotas

- En la sección anterior hemos identificado la convergencia uniforme como la caracterización fundamental para que el el riesgo empírico sea consistente.
- Para estudiar cuando la convergencia uniforme hace sentido usamos dos trucos: la idea de cotas uniformes y la simetrización usando muestras fantasma.
- La idea de cotas uniformes es una extensión elemental de la desigualdad de Chernoff al caso de un conjunto finito de funciones. Si el espacio de funciones de interes $\mathbb{F}_0$ tiene únicamente m funciones es fácil probar que:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2) \qquad (17)$$

Capacidad y cotas

- En la sección anterior hemos identificado la convergencia uniforme como la caracterización fundamental para que el el riesgo empírico sea consistente.
- Para estudiar cuando la convergencia uniforme hace sentido usamos dos trucos: la idea de cotas uniformes y la simetrización usando muestras fantasma.
- La idea de cotas uniformes es una extensión elemental de la desigualdad de Chernoff al caso de un conjunto finito de funciones. Si el espacio de funciones de interes $\mathbb{F}_0$ tiene únicamente m funciones es fácil probar que:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2) \qquad (17)$$

Capacidad y cotas

- En la sección anterior hemos identificado la convergencia uniforme como la caracterización fundamental para que el el riesgo empírico sea consistente.
- Para estudiar cuando la convergencia uniforme hace sentido usamos dos trucos: la idea de cotas uniformes y la simetrización usando muestras fantasma.
- La idea de cotas uniformes es una extensión elemental de la desigualdad de Chernoff al caso de un conjunto finito de funciones. Si el espacio de funciones de interes $\mathbb{F}_0$ tiene únicamente m funciones es fácil probar que:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2) \qquad (17)$$

Capacidad y cotas

- Luego el problema de consistencia del riesgo empírico queda resuelto cuando el conjunto $\mathbb{F}_0$ es finito.
- Ahora el truco es reducir el problema general a un problema con un número finito de funciones.
- Esto se basa en dos ideas: la idea de simetrización y la idea de aplastar.

Capacidad y cotas

- Luego el problema de consistencia del riesgo empírico queda resuelto cuando el conjunto $\mathbb{F}_0$ es finito.
- Ahora el truco es reducir el problema general a un problema con un número finito de funciones.
- Esto se basa en dos ideas: la idea de simetrización y la idea de aplastar.

Simetrización

- Considere una muestra ficticia o fantasma $\widehat{\tau} = (x_i, y_i)_{i=1,\ldots,N}$ que sea i.i.d de la misma distribución P. Este es solo un artifico matemático y en la práctica no es necesario tener una segunda muestra.
- Vapnik y Chervonenkis demostraron el siguiente teorema:

Theorem (Vapnik y Chervonenkis)

Sea $m\epsilon^2 \geq 2$ *entonces:*

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2P(\sup_{f \in \mathbb{F}_0} \left|R_{emp}(f) - \widehat{R}_{emp}(f)\right| \geq \frac{\epsilon}{2})$$

- La segunda probabilidad se toma con respecto a la distribucion de las dos muestras.
- Intuitivamente, si el riesgo empírico en dos muestras independientes está cerca, entonces es porque están cerca del verdadero riesgo.

Simetrización

- Considere una muestra ficticia o fantasma $\widehat{\tau} = (x_i, y_i)_{i=1,\ldots,N}$ que sea i.i.d de la misma distribución P. Este es solo un artifico matemático y en la práctica no es necesario tener una segunda muestra.
- Vapnik y Chervonenkis demostraron el siguiente teorema:

Theorem (Vapnik y Chervonenkis)

Sea $m\epsilon^2 \geq 2$ *entonces:*

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2P(\sup_{f \in \mathbb{F}_0} \left|R_{emp}(f) - \widehat{R}_{emp}(f)\right| \geq \frac{\epsilon}{2})$$

- La segunda probabilidad se toma con respecto a la distribucion de las dos muestras.
- Intuitivamente, si el riesgo empírico en dos muestras independientes está cerca, entonces es porque están cerca del verdadero riesgo.

Simetrización

- Considere una muestra ficticia o fantasma $\widehat{\tau} = (x_i, y_i)_{i=1,\ldots,N}$ que sea i.i.d de la misma distribución P. Este es solo un artifico matemático y en la práctica no es necesario tener una segunda muestra.
- Vapnik y Chervonenkis demostraron el siguiente teorema:

Theorem (Vapnik y Chervonenkis)

Sea $m\epsilon^2 \geq 2$ *entonces:*

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2P(\sup_{f \in \mathbb{F}_0} \left|R_{emp}(f) - \widehat{R}_{emp}(f)\right| \geq \frac{\epsilon}{2})$$

- La segunda probabilidad se toma con respecto a la distribucion de las dos muestras.
- Intuitivamente, si el riesgo empírico en dos muestras independientes está cerca, entonces es porque están cerca del verdadero riesgo.

Simetrización

- Ahora obsérvese que si bien $\mathbb{F}_0$ puede ser infinito, para calcular el riesgo empírico en cualquier muestra finita solamente importa el valor de las funciones en un número finito de puntos (en este caso solo importa el valor en $2n$ puntos).

Simetrización

- Luego en efecto hemos reducido el problema a uno con, efectivamente, un número finito de funciones (i.e, 2^{2n}).
- Obsérvese que el truco de la muestra fantasma es fundamental pues de lo contrario el lado izquierdo:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon)$$

no dependería únicamente de los valores de las hipótesis en las muestras.

- Esto casi resuelve nuestro problema, sin embargo la cota depende del número de elementos de la muestra n y cómo veremos más adelantes con esta cota tan laxa no se puede garantizar convergencia uniforme.
- Quizás se puede hacer algo mejor.

Simetrización

- Luego en efecto hemos reducido el problema a uno con, efectivamente, un número finito de funciones (i.e, 2^{2n}).
- Obsérvese que el truco de la muestra fantasma es fundamental pues de lo contrario el lado izquierdo:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon)$$

no dependería únicamente de los valores de las hipótesis en las muestras.

- Esto casi resuelve nuestro problema, sin embargo la cota depende del número de elementos de la muestra n y cómo veremos más adelantes con esta cota tan laxa no se puede garantizar convergencia uniforme.
- Quizás se puede hacer algo mejor.

Función de crecimiento y dimensión VC

- Ahora el objetivo es usar la cota de uniones y obtener una desigualdad similar a la de Chernoff pero donde m se reemplaza por un coeficiente, denominado coeficiente de aplastamiento, que es simplemente una forma de contar bien, en $\mathbb{F}_0$ el número de funciones efectivamente diferentes.

Función de crecimiento

- Sea $\tau_m = \{x_1, ..., x_m\}$ y definamos $\Pi_H(\tau_m) = \{(h(x_1), ..., h(x_m)); h \in H\}$.
- $\Pi_H(\tau_m)$ es el conjunto de todas las marcas posibles que se pueden lograr con las hipótesis en H.
- La función de crecimiento para una muestra de tamaño m se define como: $\Pi_H(m) = \text{máx}_{\tau_m \subset \Xi} \mid \Pi_H(\tau_m) \mid$
- $\Pi_H(m)$ es el mayor número de marcas posibles que se pueden lograr con las hipotésis de H para alguna muestra de tamaño m.
- $\Pi_H(m)$ también se conoce como la función de aplastamiento para la clase de hipótesis H con respecto a las muestras de tamaño m.

Cotas: Función de crecimiento

Ahora podemos ser más precisos en la cota:

$$P(\sup_{f \in \mathbb{F}_0} |R(f) - R_{emp}(f)| \geq \epsilon)$$

$$\leq 2P(\sup_{f \in \mathbb{F}_0} \left|R_{emp}(f) - \widehat{R}_{emp}(f)\right| \geq \frac{\epsilon}{2})$$

$$\leq 2P(\sup_{f \in \Pi_H(\tau_{2m})} \left|R_{emp}(f) - \widehat{R}_{emp}(f)\right| \geq \frac{\epsilon}{2})$$

y usando el hecho de que el número de funciones en $\Pi_H(\tau_{2m})$ es a lo sumo $\Pi_H(2m)$ entonces:

$$\leq 2\Pi_H(2m) \exp(\frac{-m\epsilon^2}{4})$$

Cotas: Función de crecimiento

- Utilizando esta expresión podemos llegar a algunas condiciones suficientes para consistencia.
- Supongamos que $\Pi_H(2m) \leq (2m)^k$ para algún k (crecimiento exponencial). Al sustituir en el lado derecho se obtiene:

$$2\exp(k\log(2m) - m\frac{\epsilon^2}{4}) \to 0$$

luego, si el número de funciones efectivas crece polinomialmente, entonces la minimización del riesgo es consistente.

Cotas: Función de crecimiento

- Ahora supongamos que $\Pi_H(2m) = 2^{2m}$ como es en el caso de $\mathbb{F}_0 = \mathbb{F}$.
- Entonces el lado derecho ahora es:

$$2\exp(m(2\log(2) - \frac{\epsilon^2}{4})) \to 0$$

que no tiene a cero con el tamaño de la muestra. Entonce por lo menos usando esta cota, no es posible concluir que en el espacio de todas las hipótesis la minimización del riesgo empírico es consistente.

- De hecho un teorema de Mendelson (2003) muestra que una condición necesaria y suficiente es:

$$\frac{\log \Pi_H(m)}{m} \to 0.$$

- Ejercicio: Verificar las dos afirmaciones anteriores con este teorema. En particular, en la minimización del riesgo no es consistente sobre el espacio de todas las hipótesis.