Universidad de Buenos Aires
Facultad de Ingenieria
Año 2018 - 2er Cuatrimestre

Aprendizaje Estadístico, Teoría y aplicación

Resúmen de la materia y devolutiva

Sbruzzi, José Ignacio - Ingeniería Informática #97452
jose.sbru@gmail.com

Índice

1. Clase 1 (31/8)

La primera parte de esta clase fue un repaso de diversos temas que serán útiles durante la cursada:

- Teorema de pitágoras
- espacios euclídeos
- Ortogonalidad
- Relación entre el producto interno y el coseno
- Desigualdad Cauchy-Schwartz
- Norma inducida
- Proyección Ortogonal
- Definición de esperanza
- El espacio algebraico de variables aleatorias
- Desigualdad de Markov
- Desigualdad de Chebyshev
- Desigualdad de Chernoff
- Desigualdad de Jensen
- Función convexa
- Esperanza condicional

La segunda parte de la clase se habló del problema de la comunicación digital para ilustrar la lógica por detrás de la construcción de un clasificador bayesiano. Siendo $\delta(r)$ una función que predice el dígito (0 o 1) emitido a partir del recibido $r \in \{0, 1\}$. $P(S = s|R = r)$ es la probabilidad de que se haya emitido el dígito s dado que se recibió el dígito r.

$$\delta(r) = \mathbb{1}\{\mathbb{P}(S = 1|R = r) > \mathbb{P}(S = 0|R = r)\}$$

Así, este clasificador toma la mejor decisión posible para la información que se tiene disponible (r), con lo cual es un clasificador bayesiano.

2. Clase 2 (7/9)

2.1. Definiciones iniciales

2.1.1. Clasificador

$$g : \mathbb{R}^d \to \{1, 2, ..., M\}$$

$g(x)$ representa una conjetura respecto de la naturaleza de la distribución de las x. El clasificador se equivoca cuando $g(x) \neq y$.

2.1.2. Calidad del clasificador

Sea $(X, Y) \in \mathbb{R}^d \times \{1, 2, ..., M\}$ un par donde X es una variable aleatoria que representa las propiedades observables y Y la característica a predecir. Así, se define la pérdida de un clasificador como $L(g) = \mathbb{P}(g(X) \neq Y)$.

2.1.3. Clasificador bayesiano

Es el mejor clasificador, definido por

$$argmin_{g:\mathbb{R}^d \to \{1,...,M\}}\{\mathbb{P}(g(X) \neq Y)\} = g^*$$

$$L^* = L(g^*)$$

No se da siempre que $L^* = 0$ porque Y podría no ser una función de X.

2.1.4. Dataset de entrenamiento

Se denota como $(X_i, Y_i), i = 1, 2, ..., n$; donde las parejas (X_i, Y_i) son observaciones independientes e identicamente distribuidas, al igual que (X, Y).

$$D_n = \{(X_i, Y_i), i = 1, 2, ..., n\}$$

Así, en realidad, cuando aplicamos algoritmos de machine learning tenemos una g denotada como:

$$g(X, (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$$

Donde X es una nueva observación.

Es decir,

$$g_n : \mathbb{R}^d \times (\mathbb{R}^d \times \{1, ..., M\})^n \to \{1, ..., M\}$$

Así, tenemos

$$L_n = L(g_n) = \mathbb{P}(g(X, (X_1, Y_1), ..., (X_n, Y_n)) \neq Y | (X_1, Y_1), ..., (X_n, Y_n))$$

Con lo cual L_n es una variable aleatoria dependiente de las observaciones.

2.2. Clasificador bayesiano para M=2

Sean (con $A \subset \mathbb{R}^d$, $x \in \mathbb{R}^d$, $y \in \{0, 1\}$):

$$\mu(A) = \mathbb{P}(x \in A)$$
$$\eta(x) = \mathbb{P}(Y = 1 | X = x) = \mathbb{E}[Y | X = x]$$

Así,

$$\eta(x) = \int_C \mathbb{P}(Y = 0 | X = x)\mu(dx) + \int_C \mathbb{P}(Y = 1 | X = x)\mu(dx)$$

Siendo $C = \mathbb{R}^d \times \{0, 1\}$.

Bajo estas condiciones,

$$g^*(x) = \mathbb{1}\{\eta(x) > \frac{1}{2}\}$$

3. Clase 3 (14/9)

3.1. Plug-in decision

Una "plug-in decision"(decisión .enchufada") es una función g definida por medio de una cierta función $\tilde{\eta}(x)$. Así, la función de decisión plug-in se define como:

$$g(x) = \mathbb{1}\{\tilde{\eta}(x) > \frac{1}{2}\}$$

En clase se demostró un teorema que establece que

$$L(g) - L^*(g) \leq \int_{\mathbb{R}^d} |\eta(x) - \tilde{\eta}(x)| \mu(dx) = 2\mathbb{E}[\eta(X) - \tilde{\eta}(X)]$$

Es decir, que si las funciones $\eta(x)$ y $\tilde{\eta}(x)$ son funciones similares (lo cual se ve más claramente en el miembro central de la fórmula anterior), los errores cometidos también serán similares. Es decir que, cuanto más se parezca η a $\tilde{\eta}$, más cerca estará el error de g del menor error posible (que es el de g^*).

3.2. Convergencia debil y fuerte

Una regla de clasificación g_n es consistente si, para ciertas distribuciones de (X, Y), se cumple:

$$\mathbb{E}[L_n] = \mathbb{P}(g_n(X, D_n) \neq Y) \to L^* \text{ cuando } n \to \infty$$

Y es fuertamente consistente si

$$\lim_{n\to\infty} L_n = L^* \text{ con probabilidad 1}$$

Una regla de clasificación es **universalmente consistente** si es fuertemente consistente para cualquier distribución de (X, Y).

3.3. Reglas basadas en particiones

Muchas reglas de clasificación particionan el espacio en celdas disjuntas A_i, de forma que

$$\mathbb{R}^d = \bigcup_{i=1}^{\infty} A_i$$

La regla se basa en la "mayoría electoral", es decir, si x pertenece a cierto A_i, entonces g le asignará el valor más común de y_i para los x_i pertenecientes a A_i. Es decir,

$$g_n(x) = \mathbb{1}\left\{ \sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\}\mathbb{1}\{X_i \in A(x)\} \geq \sum_{i=1}^{n} \mathbb{1}\{Y_i = 0\}\mathbb{1}\{X_i \in A(x)\} \right\}$$

donde $A(x)$ es el A_i al que pertenece x. Sea el diámetro de un conjunto contenido en $\mathbb{R}^d$ definido como:

$$diam(A) = \sup_{x,y\in A} ||x - y||$$

Y sea la cantidad de X_i presentes en la misma celda que x definida como:

$$N(x) = \sum_{i=1}^{n} \mathbb{1}\{X_i \in A(x)\}$$

La regla g_n definida más arriba es consistente cuando se cumplen las siguientes condiciones:

$$diam(A(X)) \to 0 \text{ en probabilidad}$$

$$N(X) \to \infty \text{ en probabilidad}$$

Es decir, los A_i deben ser tales que su tamaño decrece a medida que crece n pero la cantidad de puntos que contiene crece junto con n: deben ir reduciendo su tamaño pero no demasiado rápido, no deben tender a "vaciarse".

3.4. La regla del histograma

La regla del histograma es un caso especial de la regla de clasificación de la sección anterior en la que los A_i son hipercubos de dimensión d y de lado h_n.

Esta regla es universalmente consistente si se cumplen las siguientes condiciones:

$$h_n \to 0 \text{ cuando } n \to \infty$$

$$nh_n^d \to \infty \text{ cuando } n \to \infty$$

Estas condiciones son análogas a las de la sección anterior, con la diferencia de que cuando el espacio se parte en hipercubos se obtiene consistencia universal.

4. Clase 4 (28/9)

4.1. El teorema de Stone

El teorema de Stone indica condiciones bajo las cuales un clasificador que podría verse como una generalización de los de la clase 3 converge universalmente.

Se definen:

$$\eta_n(x) = \sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\} W_{ni}(x)$$

Siendo:

$$\sum_{i=1}^{n} W_{ni}(x) = 1$$

Y se define la regla de clasificación como:

$$g_n(x) = \mathbb{1}\left\{ \sum_{i=1}^{n} \mathbb{1}\{Y_i = 1\} W_{ni}(x) \geq \mathbb{1}\{Y_i = 0\} W_{ni}(x) \right\}$$

g_n converge universalmente cuando se cumplen las siguientes tres condiciones:

4.1.1. Condición 1

Existe una constante c tal que, para cualquier función medible f tal que $\mathbb{E}[f(X)] < \infty$,

$$\mathbb{E}\left\{ \sum_{i=1}^{n} W_{ni}(X) f(X_i) \right\} \leq c\mathbb{E}[f(X)]$$

4.1.2. Condición 2

Para todo $a > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left\{ \sum_{i=1}^{n} W_{ni}(X) \mathbb{1}\{||X_i - X|| > a\} \right\} = 0$$

4.1.3. Condición 3

$$\lim_{n\to\infty} \mathbb{E}\left\{ \max_{1\leq i\leq n} W_{ni}(X) \right\}$$

5. Clase 5 (5/10)

5.1. Desigualdad de Hoeffding

Sean $X_1, ..., X_n$ variables aleatoreas independientes y sean a_i y b_i tales que $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$, y sea $S_n = \sum_{i=1}^{n} X_i$. Entonces, para cualquier $\epsilon > 0$ se cumplen:

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n} (b_i - a_i)^2}$$

y también

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \leq -\epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n} (b_i - a_i)^2}$$

5.2. Cómo estimar L

Sea el estimador de L:

$$\widehat{L}_{n,m} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\left\{ g_n(X_{n+i}) \neq Y_{n+i} \right\}$$

$\widehat{L}_{n,m}$ es un estimador insesgado, es decir, $\mathbb{E}[\widehat{L}_{n,m} | D_n] = L_n$
Aplicando la desigualdad de Hoeffding podemos deducir que:

$$\mathbb{P}(|\widehat{L}_{n,m} - L_n| > \epsilon | D_n) \leq 2e^{-2m\epsilon^2}$$

Es decir, independientemente de la distribución de (X, Y) se puede acotar el error que tiene el estimador de L_n.

5.3. Cómo elegir clasificadores

C es una família de funciones $\phi : (R)^d \to \{0, 1\}$
ϕ_n^* es el clasificador de C que minimiza la L_n estimada:

$$\widehat{L_n}(\phi_n^*) \leq \widehat{L_n}(\phi) \text{ para todo } \phi \in C$$

Así, según descubierto por Vapnik y Chervonenkis:

$$L(\phi_n^*) - \inf_{\phi \in C} L(\phi) \leq 2 \sup_{\phi \in C} |\widehat{L_n}(\phi) - L(\phi)|$$

$$|\widehat{L_n}(\phi_n^*) - L(\phi_n^*)| \leq \sup_{\phi \in C} |\widehat{L_n}(\phi) - L(\phi)|$$

Además, si $|C| \leq N$, tenemos que, para todo $\epsilon > 0$,

$$\mathbb{P}\left\{ \sup_{\phi \in C} |\widehat{L_n}(\phi) - L(\phi)| > \epsilon \right\} \leq 2Ne^{-2n\epsilon^2}$$

Estos teoremas exhiben que utilizar el estimador de L_n para elegir el ϕ dentro de una clase lleva a elegir funciones de decisión que minimizan L. Esto da fundamento teórico a los algoritmos de aprendizaje supervisado, en los cuales se utilizan métodos numéricos para buscar un ϕ que minimiza una estimación de la función de pérdida, la cual se calcula a partir de los datos de entrenamiento.

6. Clase 6 (12/10)

6.1. Lema Borel-Cantelli

Sea A_n con $n = 1, 2, ...$ una secuancia de ventos infinita. Si se da que

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$$

Entonces:

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m\Big) = 0$$

6.2. Teorema de Glivenko-Cantelli

Sean $Z_1, ..., Z_n$ variables aleatorias identicamente distribuidas, con función de distribución F. Sea

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z_i \leq z\}$$

$$\mathbb{P}\left\{ \sup_{z \in \mathbb{R}} |F(z) - F_n(z)| > \epsilon \right\} \leq 8(n+1)e^{-n\epsilon^2/32}$$

Por el lema Borel-Cantelli:

$$\lim_{n \to \infty} \sup_{z \in \mathbb{R}} |F(z) - F_n(z)| = 0$$

La desigualdad de Vapnik-Chervonenkis puede comprenderse como una generalización de este teorema para cuando $z \in \mathbb{R}^d$ (en este teorema se requiere F, lo cual implica que $z \in \mathbb{R}$)

6.3. Desigualdad Vapnik-Chervonenkis

6.3.1. Definiciones previas

Sea $\mathbb{A}$ un conjunto de conjuntos medibles. Sean $z_1, ..., z_n \in \mathbb{R}^d$. Sea:

$$N_{\mathbb{A}}(z_1, ..., z_n) = \left| \left\{ \{z_1, ..., z_n\} \cap A; \cap A \in \mathbb{A} \right\} \right|$$

Sea entonces el N-avo coeficiente de destrozo/shattering/astillado de $\mathbb{A}$:

$$s(\mathbb{A}, n) = \max_{(z_1, ..., z_n) \in \{\mathbb{R}^d\}^n} N_{\mathbb{A}}(z_1, ..., z_n)$$

$s(\mathbb{A}, n)$ es la máxima cantidad de diferentes subconjuntos de n puntos que pueden ser "pescados" por la clase de conjuntos $\mathbb{A}$. En cierta forma podría decirse que $s(\mathbb{A}, n)$ es una medida de cuán "versátil", "adaptable" o "expresiva" $\mathbb{A}$ es $\mathbb{A}$.

Se definen Además

$$\nu(A) = \mathbb{P}(Z_i \in A)$$

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z_i \in A\}$$

Siendo Z_i cualquier $Z_1, Z_2, ...Z_n$ (son variables aleatorias en $\mathbb{R}^d$ identicamente distribuidas).

Entonces la desigualdad Vapnik-Chervonenkis indica que, para cualquier medida de probabilidad ν, clase de conjuntos $\mathbb{A}$, n y $\epsilon > 0$:

$$\mathbb{P}\Big\{ \sup_{A \in \mathbb{A}} |\nu_n(A) - \nu(A)| > \epsilon \Big\} \leq 8 s(\mathbb{A}, n) e^{-n\epsilon^2/32}$$

De esta desigualdad me llaman la atención dos cosas:

- Fundamenta el fenómeno de .ºverfitting", ya que la distancia entre la estimación de ν y el ν real son más distantes cuando aumenta la capacidad expresiva de $\mathbb{A}$
- Fundamenta el hecho de que el overfitting puede ser contrarrestado con más datos de entrenamiento, lo cual puede notarse en el hecho de que la cota decrece exponencialmente al crecer n.

7. Clase 7 (19/10)

En esta clase empezamos a ver regresión.

7.1. Definiciones iniciales

$$m(x) = \mathbb{E}[Y|X = x]$$
$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i$$

7.2. Algunos criterios para medir la calidad de $m_n(x)$

7.2.1. Error de norma supremo

$$||m_n - m||_\infty = \sup_{x \in C} |m_n(x) - m(x)| \text{ con } C \in \mathbb{R}^d$$

Mide el máximo error que comete m_n en C, es decir, m_n debe estar en un "tubo" que rodee a m.

7.2.2. Error LP

$$||m_n - m||_p = \int_C |m_n(x) - m(x)|^p dx$$

El error LP suma el error que comete m_n a lo largo de todo C. Generalmente se usa $p = 2$.

7.3. 4 paradigmas relacionados para estimar $m(x)$

A continuación se presentan métodos que responden a la siguiente forma:

$$m_n(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i$$

Cada $W_{n,i}(x)$ representa el peso de Y_i en el promedio ponderado.
Debe cumplirse que:

$$\sum_{i=1}^{n} W_{n,i}(x) = 1 \text{ para todo } x$$

7.3.1. Promedios locales

Consiste en asignarle a $m_n(x)$ un valor que es el promedio de los y_i que corresponden a los k puntos x_i más cercanos a x. Formalmente:

Sean los pares $(X_{(i)}, Y_{(i)})$ una permutación de los datos de entrenamiento D_n tal que

$$||x - X_{(1)}|| \leq ||x - X_{(2)}|| \leq ... \leq ||x - X_{(n)}||$$

Con esta permutación, se define m_n como:

$$m_n(x) = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}$$

Es decir que los pesos tienen los valores

$$W_{n,i} = \frac{1}{k} \mathbb{1}\{ \ X_i \text{ es uno de los } k \text{ vecinos más cercanos a } x\}$$

Este método es una adaptación del algoritmo KNN para el caso de regresión.

7.3.2. Estimador kernel de Nadaraya - Watson

sea $K : \mathbb{R}^d \to \mathbb{R}^+$ un kernel. Se definen los pesos para este caso:

$$W_{n,i}(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)}$$

Algunos kernels útiles son:

$$K(x) = \mathbb{1}\{||x|| \leq 1\}$$

$$K(x) = e^{-||x||^2}$$

Estos estimadores son una versión más interesante del caso anterior, ya que permiten establecer el peso de cada observación en función de su distancia a x, lo cual no hace el método anterior.

7.3.3. Particiones

Sea $\{A_{n,1}, A_{n,2}, ...\}$ una partición de $\mathbb{R}^d$

$$W_{n,i}(x) = \frac{\mathbb{1}\{X_i \in A_{n,i}\}}{\sum_{j=1}^{n} \mathbb{1}\{X_j \in A_{n,i}\}}$$

Este método le asigna a los x de una misma partición el mismo valor, que es el promedio de los Y_i correspondientes, es similar al método de promedios locales pero en este caso las particiones pueden ser independientes de D_n.

7.3.4. Modelado local

Según esta família de reglas, g_n está definida con l parámetros, es decir:

$$g(\cdot, \{a_k\}_{k=1}^{l}) : \mathbb{R}^d \to \mathbb{R}$$

Y los parámetros están elegidos de la siguiente forma:

$$\{a_k(x)\}_{k=1}^{l} = \underset{\{b_k\}_{k=1}^{l}}{argmin} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \cdot \left(Y_i - g\big(x, \{b_k\}_{k=1}^{l}\big)\right)$$

Así, se permite "ajustar" el promedio que genera el método Nadaraya-Watson por medio de una función con parámetros.

Así, g_n podría pertenecer a una família de polinomios:

$$g_n(x, \{a_k\}_{k=1}^{l}) = \sum_{k=1}^{l} a_k \cdot x^{k-1}$$

7.3.5. Modelado global

Se tiene una partición del espacio como en el método de particiones. El modelado globales una generalización del método de particiones.

Así, definimos:

$$\mathbb{F}_n = \left\{ \sum_j a_j \mathbb{1}\{x \in A_{n,j} : a_j \in \mathbb{R}\} \right\}$$

$$m_n = \underset{f \in \mathbb{F}_n}{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \right\}$$

7.3.6. Minimos cuadrados penalizados

Se define un término de penalización $J_n(f) \geq 0$ que penaliza a f según cuán violentamente cambia, permitiendo así elegir funciones f que no hagan overfit sobre los datos de entrenamiento:

$$m_n = \underset{f}{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - Y_i\big)^2 + J_n(f) \right\}$$

Para el caso $d = 1$, se utiliza como función de penalización

$$J_n(f) = \lambda_n \int |f^{(k)}(t)|^2 dt$$

Se puede generalizar J_n a campos vectoriales, la fórmula utiliza derivadas parciales. El objetivo del término $J_n(f)$ es penalizar a las f que tienen curvas violentas, lo cual es usual cuando se ajustan datos de entrenamiento a un polinomio.

7.4. La maldición de la dimensionalidad

La maldición de la dimensionalidad consiste en el hecho de que, cuantas más dimensiones tienen los datos, menos significativa es la distancia entre ellos. Así, si se tiene un hipercubo de dimensión d con puntos distribuidos de forma uniforme sobre cada dimensión, la mayoría de los puntos se ubicarán sobre los bordes del cubo y estarán a una distancia muy similar los unos de los otros. Este efecto es cada vez más pronunciado a medida que crece d.

7.5. Balance sesgo-varianza

$$\mathbb{E}\Big\{ |m_n(x) - m(x)|^2 \Big\} = \mathbb{E}\Big\{ |m_n(x) - \mathbb{E}\big\{m_n(x)\big\}|^2 \Big\} + \Big| \mathbb{E}\big\{m_n(x)\big\} - m(x) \Big|^2$$

$$\mathbb{E}\Big\{ |m_n(x) - m(x)|^2 \Big\} = Var(m_n(x)) + |sesgo(m_n(x))|^2$$

Así, *var* es la varianza de la variable aleatoria $m_n(x)$, es decir, cuánto "le cree.ª los datos: un algoritmo de machine learning con alta varianza es

más propenso a sufrir overfitting. Una alta varianza de $m_n(x)$ implica que el modificar D_n cambia $m_n(x)$ de forma significativa.

Por otro lado, el *sesgo*, cuantifica cuán "terco.[es] m_n. Un alto sesgo implica que incluso si la distribución de D_n cambia, m_n permanece similar. Esto está relacionado al fenómeno de underfitting.

El error de un m_n es la suma de ambas componentes, ambas crecen en condiciones opuestas.

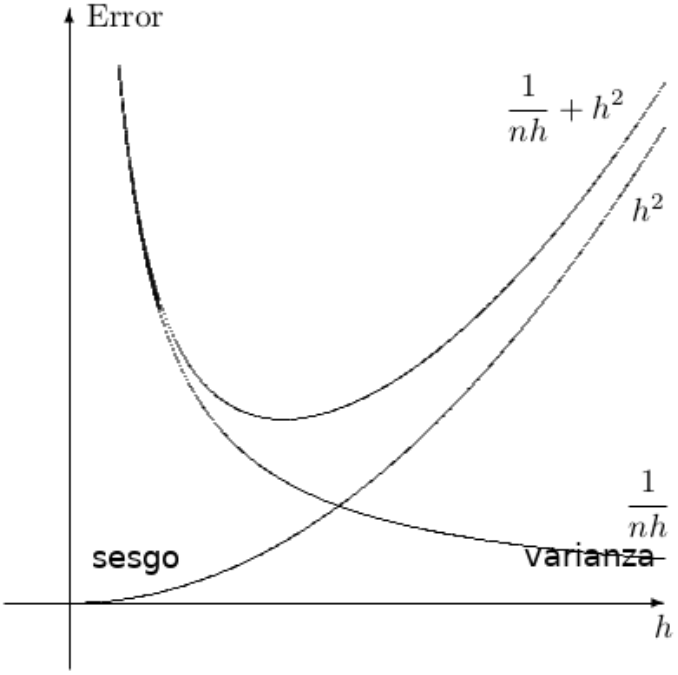

Figure 2.4. Bias–variance tradeoff.

Ese gráfico está extraído de Gyorfi, muestra cómo se comportan el sesgo y la varianza al cambiar cuán detallado es un g_n construído a partir de un kernel Nadaraya-Watson.

7.6. Cómo comparar g_n cuando no se dispone de las distribuciones

7.6.1. Método de resustitución

En el método de resustitución se eligen los parámetros p de manera de que la función de regresión $m_{n,p}(x)$ minimice

$$L(D_n) = \frac{1}{n}\sum_{i=1}^{n} |m_{n,p}(X_i) - Y_i|^2$$

Esto lleva a resultados demasiado optimistas, ya que no es otra cosa que usar los mismos datos para entrenar el algoritmo y para medir cuán bueno es. Así, g_n funcionará peor que lo esperado al predecir nuevos pares (X, Y).

7.6.2. Otro método de resustitución

Se divide el conjunto de datos D_n en D_{n_1} y D_{n_2} de forma que $D_n = D_{n_1} \cap D_{n_2}$. Luego se elige p minimizando $L(D_{n_1})$ y se utiliza $g(\cdot, D_{n_2})$ para realizar predicciones.

7.6.3. K-fold cross-validation (dividido de forma secuencial)

Se definen k conjuntos de pares $D_{n,k}$ de forma que $|D_{n,k}| = \frac{n}{k}$ y que $\bigcup_{l=1}^{k} D_{n,l} = D_n$.

Luego se eligen los parámetros p minimizando

$$\frac{1}{k}\sum_{l=1}^{k} \frac{1}{n/k} \sum_{i=\frac{n}{k}(l-1)+1}^{\frac{n}{k}l} |m_{n-\frac{n}{k},p}(X_i, D_{n,l} - Y_i)|^2$$

Así, se tiene en cuenta el promedio de los errores de las k particiones al mismo tiempo para obtener p.

8. Clase 8 (26/10): Ley de los grandes números

Sea $f \in \mathbb{F}_n$, y sea f la función de regresión. Queremos minimizar el riesgo empírico L_2, es decir

$$\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2$$

Se definen $Z = (X, Y)$, $Z_i = (X_i, Y_i)$, $i = 1, ..., n$, $g_f(x, y) = |f(x) - y|^2$, $\mathbb{G}_n = \{g_f : f \in \mathbb{F}_n\}$

El estimador del riesgo L_2 sólo es consistente sii

$$\sup_{g \in \mathbb{G}_n} \left| \frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\} \right| \to 0 \text{ cuando } (n \to \infty) \text{ a.s.}$$

8.1. Desigualdades exponenciales básicas

Se cumple

$$\mathbb{P}\left\{ \sup_{g\in\mathbb{G}_n} \left| \frac{1}{n}\sum_{j=1}^{n} g(Z_j) - \mathbb{E}\{g(Z)\} \right| > \epsilon \right\} \leq 2|\mathbb{G}_n| e^{-\frac{2n\epsilon^2}{B^2}}$$

Cuando $\mathbb{G}_n$ es una clase finite y cumple

$$\sum_{n=1}^{\infty} |\mathbb{G}_n| e^{-\frac{2n\epsilon^2}{B^2}} < \infty$$

Para cualquier $\epsilon > 0$

Cuando $\mathbb{G}_n$ no es una clase finita, puede ser posible encontrar un $\mathbb{G}_{n,\epsilon}$ finito que cumpla que todos los elementos de la família infinita están a una distancia supremo menor que ϵ que cualquier elemento de la família finita.

8.2. Herramientas métricas

Sea d una distancia entre funciones.

Sean las bolas $B_\epsilon(g) = \{g' \in g : d(g', g) < \epsilon\}$.

Un cubrimiento de G es cualquier $G' = \{g_1, ..., g_N\}$ finito tal que $G' \subset G$ y $G \subset \bigcup_{i=1}^{N} B_\epsilon(g_i)$.

Sea el número de ϵ cubrimiento de G:

$$\mathbb{N}(\epsilon, G, d) = \mathbb{N}_d(\epsilon, G) = \begin{cases} \infty & \text{si no existe ningún } \epsilon \text{ cubrimiento de } G \text{ con respecto a } d \\ H & \text{en otro caso} \end{cases}$$

$H = min\{n \text{ natural } : \text{ existe un } \epsilon \text{ cubrimiento de tamaño n de } G \text{ con respecto a d } \}$

Cualquier colección finita $\{g_1, g_2, ..., g_n\} \subset G$ tal que $d(g_i, g_j) \geq \epsilon \forall i \neq j$ es un ϵ empaquetado de G con respecto a d.

Se define el número de ϵ empaquetado de G con respecto a d como $M(\epsilon, G, d) = M_d(\epsilon, G)$, que vale infinito para cualquier n natural existe un ϵ empaquetado de G de tamaño n. De lo contrario, vale

$max\left\{n \in \text{ naturales } : \{g_1, ..., g_n\} \text{ es un } \epsilon\text{-empaquetado de } G \text{ con respecto a } d\right\}$

8.3. Distancias posibles

8.3.1. Distancia supremo

$$d_{||\cdot||_\infty}(f,g) = \underset{z\in\mathbb{R}^d}{sup}\,|f(z)-g(z)|$$

8.3.2. Distancia LP

$$d_{L_p(\nu)}(f,g) = \left(\int |f(z)|^p \nu(dz)\right)^{\frac{1}{p}}$$

8.4. Lema

Sea n natural, y sea G_n un conjunto de funciones $g : \mathbb{R}^d \to [0, B]$ y $\epsilon > 0$. Entonces:

$$\mathbb{P}\left\{ \underset{g\in\mathbb{G}_n}{sup} \left|\frac{1}{n}\sum_{j=1}^{n} g(Z_j) - \mathbb{E}\{g(Z)\}\right| > \epsilon \right\} \leq 2\mathbb{N}_\infty(\epsilon/3, \mathbb{G}_n) e^{-\frac{2n\epsilon^2}{9B^2}}$$

Además, si se cumple

$$\sum_{n=1}^{\infty} \mathbb{N}_\infty(\epsilon/3, \mathbb{G}_n) e^{-\frac{2n\epsilon^2}{9B^2}}$$

entonces también es se cumple

$$\underset{g\in\mathbb{G}_n}{sup} \left|\frac{1}{n}\sum_{i=1}^{n} g(Z_i) - \mathbb{E}\{g(Z)\}\right| \to 0 \text{ cuando } (n\to\infty) \text{ a.s.}$$

Este lema indica que con una família $\mathbb{G}_n$ discreta se pueden lograr funciones de regresión consistentes.

9. Clase 9 (2/11)

9.1. Dimensión Vapnik-Chervonenkis

Retomando el concepto de "N-avo coeficiente de destrozo / shattering / astillado", se añade la definición de "Dimensión Vapnik - Chervonenkis":

$$dim_{VC}(\mathbb{A}) = sup\left\{n \text{ natural } : s(\mathbb{A}, n) = 2^n\right\}$$

$s(\mathbb{A}, n) = 2^n$ implica que la família $\mathbb{A}$ puede destrozar algún conjunto de n puntos, es decir, que existen conjuntos pertenecientes a $\mathbb{A}$ que permiten elegir cada una de las 2^n combinaciones posibles de n puntos.

Así, la dimensión Vapnik-Chervonenkis indica la cantidad máxima de puntos que $\mathbb{A}$ puede destrozar. Así, esta dimensión es un indicador de la "capacidad expresiva" o la "versatilidad" de $\mathbb{A}$.

9.2. Lema de Sawer

$$s(\mathbb{A}, n) = \begin{cases} 2^n & \text{si } n \leq dim_{VC}(\mathbb{A}) \\ \sum_{i=0}^{dim_{VC}(\mathbb{A})} \binom{n}{i} & \text{si } n > dim_{VC}(\mathbb{A}) \end{cases}$$

Tal como explicado en los párrafos anteriores, $\mathbb{A}$ destroza todos los conjuntos de cardinalidad menor a $dim_{VC}(\mathbb{A})$, y el coeficiente de destrozo más allá de ese ümbral"sólo crece de forma polinómica (esto, si la dimensión Vapnkik-Chervonenkis es finita).

9.3. Lema

Sea $\mathbb{G}$ una clase de funciones $g : R^d \to [0, B]$ y sea $\mathbb{G}^+ = \big\{\{(z, t) \in R^d \times R; t \leq g(z)\} : g \in \mathbb{G}\big\}$. Así, es posible referirnos a la dimensión Vapnik-Chervonenkis de $\mathbb{G}^+$.

Si $dim_{VC}(\mathbb{G}^+) \geq 2$, $p \geq 1$, $0 < \epsilon < \frac{B}{4}$, con ν una medida de probabilidad sobre $\mathbb{R}^d$; entonces:

$$\mathbb{M}(\epsilon, \mathbb{G}, ||\cdot||_{L_p(\nu)}) \leq 3\left(\frac{2eB^p}{\epsilon^p} log \frac{3eB^p}{\epsilon^p}\right)^{dim_{VC}(\mathbb{G}^+)}$$

10. Clase 10 (9/11)

10.1. Teorema

Si $\mathbb{A}$ es un conjunto de subconjuntos de $\mathbb{R}^d$ con dimensión Vapnik - Chervonenkis finita, entonces para todo n vale que:

$$s(\mathbb{A}, n) \leq (n + 1)^{dim_{VC}(\mathbb{A})}$$

Este teorema también demuestra que el coeficiente de destrozo crece como mucho polinomicamente.

10.2. Princípio de mínimos cuadrados

Elegir un f entre todas las funciones posibles lleva a overfitting, lo cual no es consistente. Así, se debe buscar un f perteneciente a una clase de funciones "adecuadas" $\mathbb{F}_n$.

Sea $\mathbb{F}_n = \mathbb{F}_n(D_n)$ una família de funciones $f : \mathbb{R}^d \to \mathbb{R}$ dependeientes en los datos D_n. Si m_n cumple

$$m_n(\cdot) = \underset{f \in \mathbb{F}_n}{argmin} \frac{1}{n} \sum_{j=1}^{n} |f(X_j) - Y_j|^2$$

entonces

$$\int |m_n(x) - m(x)|^2 \mu(dx) \leq A + B$$

$$A = 2 \underset{f \in \mathbb{F}_n}{sup} \left| \frac{1}{n} \sum_{j=1}^{n} |f(X_j) - Y_j|^2 - \mathbb{E}[(f(X) - Y)^2] \right|$$

$$B = \underset{f \in \mathbb{F}_n}{inf} \int |f(x) - m(x)|^2 \mu(dx)$$

11. Clase 11 (16/11)

Copiar el teorema 9.1 pag 136 y después algo de la pag 152