



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

<b>1. Clase 1 (31/8)</b>	<b>2</b>
<b>2. Clase 2 (7/9)</b>	<b>3</b>
2.1. Definiciones iniciales . . . . .	3
2.1.1. Clasificador . . . . .	3
2.1.2. Calidad del clasificador . . . . .	3
2.1.3. Clasificador bayesiano . . . . .	3
2.1.4. Dataset de entrenamiento . . . . .	3
2.2. Clasificador bayesiano para $M=2$ . . . . .	4
<b>3. Clase 3 (14/9)</b>	<b>4</b>
3.1. Plug-in decision . . . . .	4
3.2. Convergencia debil y fuerte . . . . .	5
3.3. Reglas basadas en particiones . . . . .	5
3.4. La regla del histograma . . . . .	6
<b>4. Clase 4 (28/9)</b>	<b>6</b>
4.1. El teorema de Stone . . . . .	6
4.1.1. Condición 1 . . . . .	7
4.1.2. Condición 2 . . . . .	7
4.1.3. Condición 3 . . . . .	7
<b>5. Clase 5 (5/10)</b>	<b>7</b>
5.1. Desigualdad de Hoeffding . . . . .	7
5.2. Cómo estimar $L$ . . . . .	7
5.3. Cómo elegir clasificadores . . . . .	8
<b>6. Clase 6 (12/10)</b>	<b>8</b>
6.1. Lema Borel-Cantelli . . . . .	8
6.2. Teorema de Glivenko-Cantelli . . . . .	9
6.3. Desigualdad Vapnik-Chervonenkis . . . . .	9
6.3.1. Definiciones previas . . . . .	9
<b>7. Clase 7 (19/10)</b>	<b>10</b>

## 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

$$g : \mathbb{R}^d \rightarrow \{1, 2, \dots, M\}$$

$g(x)$  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  $(X, Y) \in \mathbb{R}^d \times \{1, 2, \dots, M\}$  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

$$\operatorname{argmin}_{g: \mathbb{R}^d \rightarrow \{1, \dots, M\}} \{\mathbb{P}(g(X) \neq Y)\} = g^*$$

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

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  $(X_i, Y_i), i = 1, 2, \dots, n$ ; donde las parejas  $(X_i, Y_i)$  son observaciones independientes e idénticamente distribuidas, al igual que  $(X, Y)$ .

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

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

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

Donde  $X$  es una nueva observación.

Es decir,

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

Así, tenemos

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

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  $\eta$  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:

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

Y es fuertamente consistente si

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

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:

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

$N(X) \rightarrow \infty$  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 "vaciar".

### 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:

$$\begin{aligned} h_n &\rightarrow 0 \text{ cuando } n \rightarrow \infty \\ nh_n^d &\rightarrow \infty \text{ cuando } n \rightarrow \infty \end{aligned}$$

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  $a > 0$ ,

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

#### 4.1.3. Condición 3

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

### 5. Clase 5 (5/10)

#### 5.1. Desigualdad de Hoeffding

Sean  $X_1, \dots, X_n$  variables aleatorias 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 familia de funciones  $\phi : (R)^d \rightarrow \{0, 1\}$   
 $\phi_n^*$  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  $\phi$  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  $\phi$  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  $n = 1, 2, \dots$  una secuencia de eventos infinita. Si se da que

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

Entonces:

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

## 6.2. Teorema de Glivenko-Cantelli

Sean  $Z_1, \dots, Z_n$  variables aleatorias idénticamente 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:

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

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  $z_1, \dots, z_n \in \mathbb{R}^d$ . Sea:

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

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

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

$s(\mathbb{A}, n)$  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  $Z_1, Z_2, \dots, Z_n$  (son variables aleatorias en  $\mathbb{R}^d$  idénticamente distribuidas).

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

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

De esta desigualdad me llaman la atención dos cosas:

- Fundamenta el fenómeno de "overfitting", ya que la distancia entre la estimación de  $\nu$  y el  $\nu$  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)

esta es complicada creo