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 .^{en}chufada") 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

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

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:

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

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

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