Probabilidad y Estadística para Inteligencia Artificial

Dr. Ing. Pablo Briff

Laboratorio de Sistemas Embebidos - FIUBA

pbriff@fi.uba.ar

25 de Julio de 2020

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Tabla de Contenidos I

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ
- Un estimador es una v.a. de la forma $\hat{X} = g(X)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ
- Un estimador es una v.a. de la forma $\hat{X} = g(X)$
- La función $g(X)$ define los distintos estimadores de X

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ
- Un estimador es una v.a. de la forma $\hat{X} = g(X)$
- La función $g(X)$ define los distintos estimadores de X
- Notamos que al ser $\hat{X}$ una función de la v.a X adquiere propiedades estadísticas de X

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ
- Un estimador es una v.a. de la forma $\hat{X} = g(X)$
- La función $g(X)$ define los distintos estimadores de X
- Notamos que al ser $\hat{X}$ una función de la v.a X adquiere propiedades estadísticas de X
- Ejemplos de estimadores son:

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ
- Un estimador es una v.a. de la forma $\hat{X} = g(X)$
- La función $g(X)$ define los distintos estimadores de X
- Notamos que al ser $\hat{X}$ una función de la v.a X adquiere propiedades estadísticas de X
- Ejemplos de estimadores son:
- La media muestral $\bar{X} = \frac{1}{N} \sum_{i=0}^{N-1} x_i$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimador

- Sea X una v.a. de la cual queremos estimar un parámetro μ
- Un estimador es una v.a. de la forma $\hat{X} = g(X)$
- La función $g(X)$ define los distintos estimadores de X
- Notamos que al ser $\hat{X}$ una función de la v.a X adquiere propiedades estadísticas de X
- Ejemplos de estimadores son:
- La media muestral $\bar{X} = \frac{1}{N} \sum_{i=0}^{N-1} x_i$
- Calcular la esperanza y varianza del estimador $\bar{X}$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:
- Esperanza: $E[\bar{X}] = E[g(X)]$

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:
- Esperanza: $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias): $b = E[\bar{X}] - \mu$

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:
- Esperanza: $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias): $b = E[\bar{X}] - \mu$
- Varianza: $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:
- Esperanza: $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias): $b = E[\bar{X}] - \mu$
- Varianza: $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$
- Error cuadrático medio (o mean squared error, MSE): $mse = \text{var}[\bar{X}] + b^2$

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:
- Esperanza: $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias): $b = E[\bar{X}] - \mu$
- Varianza: $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$
- Error cuadrático medio (o mean squared error, MSE): $mse = \text{var}[\bar{X}] + b^2$
- Estimadores insesgados: $b = 0$

Esperanza, Sesgo, Varianza y MSE

- Definimos las siguientes propiedades de un estimador $\bar{X}$ del parámetro real μ:
- Esperanza: $E[\bar{X}] = E[g(X)]$
- Sesgo (o bias): $b = E[\bar{X}] - \mu$
- Varianza: $\text{var}[\bar{X}] = E[(\bar{X} - E[\bar{X}])^2]$
- Error cuadrático medio (o mean squared error, MSE): $mse = \text{var}[\bar{X}] + b^2$
- Estimadores insesgados: $b = 0$
- Qué es mejor? Poco sesgo o poca varianza?

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a. X dada una *medición* Y

Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a. X dada una *medición* Y
- Por ej. Y puede ser una versión de X contaminada por ruido

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a. X dada una *medición* Y
- Por ej. Y puede ser una versión de X contaminada por ruido
- Sea $\hat{X}$ un estimador de X

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a. X dada una *medición* Y
- Por ej. Y puede ser una versión de X contaminada por ruido
- Sea $\hat{X}$ un estimador de X
- Nos interesa encontrar un estimador tal que el error cuadrático medio sea mínimo (least squares estimation, LSE):

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a. X dada una *medición* Y
- Por ej. Y puede ser una versión de X contaminada por ruido
- Sea $\hat{X}$ un estimador de X
- Nos interesa encontrar un estimador tal que el error cuadrático medio sea mínimo (least squares estimation, LSE):
- min $E[(X - \hat{X})^2]$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Es común querer estimar el valor de una v.a. X dada una *medición* Y
- Por ej. Y puede ser una versión de X contaminada por ruido
- Sea $\hat{X}$ un estimador de X
- Nos interesa encontrar un estimador tal que el error cuadrático medio sea mínimo (least squares estimation, LSE):
- min $E[(X - \hat{X})^2]$
- Demonstramos a continuación que el mejor estimador de LSE es $\hat{X} = E[X]$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$
- Vemos que $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$
- Vemos que $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$
- Vemos que $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición $E[X - \mu] = 0$, entonces

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$
- Vemos que $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición $E[X - \mu] = 0$, entonces
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + (\mu - \hat{X})^2$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$
- Vemos que $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición $E[X - \mu] = 0$, entonces
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + (\mu - \hat{X})^2$
- Notando que $E[(X - \mu)^2] = \text{var}[X]$ no depende de $\hat{X}$, entonces para tener mínimo $E[(X - \hat{X})^2]$ debemos minimizar el último término:

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos

- Llamamos $\mu = E[X]$
- Vemos que $E[(X - \hat{X})^2] = E[((X - \mu) + (\mu - \hat{X}))^2]$
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + 2E[X - \mu](\mu - \hat{X}) + (\mu - \hat{X})^2$
- Por definición $E[X - \mu] = 0$, entonces
- $E[(X - \hat{X})^2] = E[(X - \mu)^2] + (\mu - \hat{X})^2$
- Notando que $E[(X - \mu)^2] = \text{var}[X]$ no depende de $\hat{X}$, entonces para tener mínimo $E[(X - \hat{X})^2]$ debemos minimizar el último término:
- $\therefore \hat{X} = \mu \quad \square$

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y
- Para cualquier estimador $g(Y)$ función de la observación se cumple:

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y
- Para cualquier estimador $g(Y)$ función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y
- Para cualquier estimador $g(Y)$ función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y
- Para cualquier estimador $g(Y)$ función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:
- $E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2]$

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y
- Para cualquier estimador $g(Y)$ función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:
- $E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2]$
- El resultado se puede extender a n v.a. condicionales:

Estimación de Cuadrados Mínimos Condicional

- Sobre el mismo proceso anterior observamos $Y = y$
- Siguiendo el razonamiento anterior:
- $E[X|Y = y]$ minimiza $E[(X - \hat{X})^2|Y = y]$
- $E[X|Y = y]$ es la estimación de cuadrados mínimos de X dada la observación y
- Para cualquier estimador $g(Y)$ función de la observación se cumple:
- $E[(X - E[X|Y])^2|Y] \leq E[(X - g(Y))^2|Y]$
- Usando la ley de esperanzas iteradas llegamos a:
- $E[(X - E[X|Y])^2] \leq E[(X - g(Y))^2]$
- El resultado se puede extender a n v.a. condicionales:
- $E[(X - E[X|Y_1, Y_2, \ldots, Y_n])^2] \leq E[(X - g(Y_1, Y_2, \ldots, Y_n))^2]$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como $\tilde{X} = X - \hat{X}$, donde $\hat{X} = E[X|Y]$ es el estimador óptimo

Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como $\tilde{X} = X - \hat{X}$, donde $\hat{X} = E[X|Y]$ es el estimador óptimo
- Notamos que $E[\tilde{X}] = 0$

Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como $\tilde{X} = X - \hat{X}$, donde $\hat{X} = E[X|Y]$ es el estimador óptimo
- Notamos que $E[\tilde{X}] = 0$
- También se cumple que $E[\tilde{X}|Y = y] = 0$ para todo y

Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como $\tilde{X} = X - \hat{X}$, donde $\hat{X} = E[X|Y]$ es el estimador óptimo
- Notamos que $E[\tilde{X}] = 0$
- También se cumple que $E[\tilde{X}|Y = y] = 0$ para todo y
- Además, el error de estimación es descorrelacionado con la estimación $\hat{X}$, es decir $E[\tilde{X}\hat{X}] = 0$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como $\tilde{X} = X - \hat{X}$, donde $\hat{X} = E[X|Y]$ es el estimador óptimo
- Notamos que $E[\tilde{X}] = 0$
- También se cumple que $E[\tilde{X}|Y = y] = 0$ para todo y
- Además, el error de estimación es descorrelacionado con la estimación $\hat{X}$, es decir $E[\tilde{X}\hat{X}] = 0$
- Se cumple la siguiente ley de varianzas:

Error de Estimación de Cuadrados Mínimos

- Definimos el error de estimación como $\tilde{X} = X - \hat{X}$, donde $\hat{X} = E[X|Y]$ es el estimador óptimo
- Notamos que $E[\tilde{X}] = 0$
- También se cumple que $E[\tilde{X}|Y = y] = 0$ para todo y
- Además, el error de estimación es descorrelacionado con la estimación $\hat{X}$, es decir $E[\tilde{X}\hat{X}] = 0$
- Se cumple la siguiente ley de varianzas:
- $\text{var}[X] = \text{var}[\hat{X}] + \text{var}[\tilde{X}]$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$
- Para $n = 1$ tenemos $g(Y) = aY + b$

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$
- Para $n = 1$ tenemos $g(Y) = aY + b$
- Minimizar: $E[(X - aY - b)^2]$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$
- Para $n = 1$ tenemos $g(Y) = aY + b$
- Minimizar: $E[(X - aY - b)^2]$
- Si fijamos a, es como tener que estimar una v.a. $X - aY$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$
- Para $n = 1$ tenemos $g(Y) = aY + b$
- Minimizar: $E[(X - aY - b)^2]$
- Si fijamos a, es como tener que estimar una v.a. $X - aY$
- Entonces $b = E[X - aY] = E[X] - aE[Y]$

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$
- Para $n = 1$ tenemos $g(Y) = aY + b$
- Minimizar: $E[(X - aY - b)^2]$
- Si fijamos a, es como tener que estimar una v.a. $X - aY$
- Entonces $b = E[X - aY] = E[X] - aE[Y]$
- Reemplazando todo en función de a queda

Cuadrados Mínimos Lineal

- Calcular $E[X|Y]$ para todos los $g(Y)$ es en general complicado
- Por simplicidad nos limitamos a estimadores lineales $g(Y) = a_1 Y_1 + \ldots + a_n Y_n + b$
- Para $n = 1$ tenemos $g(Y) = aY + b$
- Minimizar: $E[(X - aY - b)^2]$
- Si fijamos a, es como tener que estimar una v.a. $X - aY$
- Entonces $b = E[X - aY] = E[X] - aE[Y]$
- Reemplazando todo en función de a queda
- $E[((X - E[X]) - a(Y - E[Y]))^2] = \sigma_X^2 + a^2\sigma_Y^2 - 2a\,\mathrm{cov}(X, Y)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Cuadrados Mínimos Lineal

- Minimizando en función de a queda

Cuadrados Mínimos Lineal

- Minimizando en función de a queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$ con $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$

Cuadrados Mínimos Lineal

- Minimizando en función de a queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$ con $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
- El estimador de cuadrados mínimos lineal de X basado en Y es:

Cuadrados Mínimos Lineal

- Minimizando en función de a queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$ con $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
- El estimador de cuadrados mínimos lineal de X basado en Y es:
- $\hat{X} = E[X] + \frac{\text{cov}(X,Y)}{\sigma_Y^2}(Y - E[Y])$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Cuadrados Mínimos Lineal

- Minimizando en función de a queda
- $a = \rho \frac{\sigma_X}{\sigma_Y}$ con $\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
- El estimador de cuadrados mínimos lineal de X basado en Y es:
- $\hat{X} = E[X] + \frac{\text{cov}(X,Y)}{\sigma_Y^2}(Y - E[Y])$
- Es decir necesitamos conocimiento previo de las medias, varianzas y covarianzas de las v.a.

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean $\mathrm{x} = [x_1, \ldots, x_n]^T$, $\mathrm{y} = [y_1, \ldots, y_n]^T$ observaciones de dos variables X, Y en instantes de tiempos distintos n

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean $x = [x_1, \ldots, x_n]^T$, $y = [y_1, \ldots, y_n]^T$ observaciones de dos variables X, Y en instantes de tiempos distintos n
- Queremos encontrar una relación lineal $Y = aX + b$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean $\mathrm{x} = [x_1, \ldots, x_n]^T$, $\mathrm{y} = [y_1, \ldots, y_n]^T$ observaciones de dos variables X, Y en instantes de tiempos distintos n
- Queremos encontrar una relación lineal $Y = aX + b$
- Encontrar $\beta = [a, b]^T$ tal que $\|\mathrm{y} - \mathrm{x}\|^2$ sea mínimo

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean $x = [x_1, \ldots, x_n]^T$, $y = [y_1, \ldots, y_n]^T$ observaciones de dos variables X, Y en instantes de tiempos distintos n
- Queremos encontrar una relación lineal $Y = aX + b$
- Encontrar $\beta = [a, b]^T$ tal que $\|y - x\|^2$ sea mínimo
- Definimos la matriz: $A = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean $\mathrm{x} = [x_1, \ldots, x_n]^T$, $\mathrm{y} = [y_1, \ldots, y_n]^T$ observaciones de dos variables X, Y en instantes de tiempos distintos n
- Queremos encontrar una relación lineal $Y = aX + b$
- Encontrar $\beta = [a, b]^T$ tal que $\|\mathrm{y} - \mathrm{x}\|^2$ sea mínimo
- Definimos la matriz: $\mathrm{A} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$
- Si A tiene rango completo, entonces

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Regresión Lineal

- Extendemos el concepto de LSE a una regresión lineal
- Sean $\mathrm{x} = [x_1, \ldots, x_n]^T$, $\mathrm{y} = [y_1, \ldots, y_n]^T$ observaciones de dos variables X, Y en instantes de tiempos distintos n
- Queremos encontrar una relación lineal $Y = aX + b$
- Encontrar $\beta = [a, b]^T$ tal que $\|\mathrm{y} - \mathrm{x}\|^2$ sea mínimo
- Definimos la matriz: $\mathrm{A} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$
- Si A tiene rango completo, entonces
- $\beta = (\mathrm{A}^T\mathrm{A})^{-1}\mathrm{A}^T\mathrm{y}$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Sean $\mathcal{X} = \{x_1, \ldots, x_N\}$ muestras obtenidas a partir de una realización de las v.a. i.i.d $X = \{X_1, \ldots, X_N\}$

Máxima Verosimilitud

- Sean $\mathcal{X} = \{x_1, \ldots, x_N\}$ muestras obtenidas a partir de una realización de las v.a. i.i.d $X = \{X_1, \ldots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s) desconocido(s) θ

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Sean $\mathcal{X} = \{x_1, \ldots, x_N\}$ muestras obtenidas a partir de una realización de las v.a. i.i.d $X = \{X_1, \ldots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s) desconocido(s) θ
- Es decir, $x_n \sim p(\mathcal{X}|\theta)$

Máxima Verosimilitud

- Sean $\mathcal{X} = \{x_1, \ldots, x_N\}$ muestras obtenidas a partir de una realización de las v.a. i.i.d $X = \{X_1, \ldots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s) desconocido(s) θ
- Es decir, $x_n \sim p(\mathcal{X}|\theta)$
- La idea en Máxima Verosimilitud (MV) es encontrar θ tal que la probabilidad de obtener $\mathcal{X}$ a partir de $p(\mathcal{X}|\theta)$ sea máxima

Máxima Verosimilitud

- Sean $\mathcal{X} = \{x_1, \ldots, x_N\}$ muestras obtenidas a partir de una realización de las v.a. i.i.d $X = \{X_1, \ldots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s) desconocido(s) θ
- Es decir, $x_n \sim p(\mathcal{X}|\theta)$
- La idea en Máxima Verosimilitud (MV) es encontrar θ tal que la probabilidad de obtener $\mathcal{X}$ a partir de $p(\mathcal{X}|\theta)$ sea máxima
- La verosimilitud del producto, dada la independencia de las v.a. es el producto de las verosimilitudes

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Sean $\mathcal{X} = \{x_1, \ldots, x_N\}$ muestras obtenidas a partir de una realización de las v.a. i.i.d $X = \{X_1, \ldots, X_N\}$
- Las muestras surgen de una pdf conocida con parámetro(s) desconocido(s) θ
- Es decir, $x_n \sim p(\mathcal{X}|\theta)$
- La idea en Máxima Verosimilitud (MV) es encontrar θ tal que la probabilidad de obtener $\mathcal{X}$ a partir de $p(\mathcal{X}|\theta)$ sea máxima
- La verosimilitud del producto, dada la independencia de las v.a. es el producto de las verosimilitudes
- $l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{i=1}^{N} p_{X_i|\Theta}(x_i|\theta)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Tomamos el logaritmo de $l(\theta|\mathcal{X})$ para transformar el producto en sumas

Máxima Verosimilitud

- Tomamos el logaritmo de $l(\theta|\mathcal{X})$ para transformar el producto en sumas
- Log es una función monótona y no cambia la propiedades de optimalidad

Máxima Verosimilitud

- Tomamos el logaritmo de $l(\theta|\mathcal{X})$ para transformar el producto en sumas
- Log es una función monótona y no cambia la propiedades de optimalidad
- Entonces la verosimilitud logarítmica (log-likelihood, LL) es:

Máxima Verosimilitud

- Tomamos el logaritmo de $l(\theta|\mathcal{X})$ para transformar el producto en sumas
- Log es una función monótona y no cambia la propiedades de optimalidad
- Entonces la verosimilitud logarítmica (log-likelihood, LL) es:
- $\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_{i=1}^{N} \log p_{X_i|\Theta}(x_i|\theta)$

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para $\mathcal{X} = \{x_i\}$ con pdf como la anterior, la LL:

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para $\mathcal{X} = \{x_i\}$ con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para $\mathcal{X} = \{x_i\}$ con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de $\mathcal{L}$ a cero

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para $\mathcal{X} = \{x_i\}$ con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de $\mathcal{L}$ a cero
- $\nabla\mathcal{L} = [\frac{\partial\mathcal{L}}{\partial\mu}\frac{\partial\mathcal{L}}{\partial\sigma}] = [0\,0]$ lo cual da:

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para $\mathcal{X} = \{x_i\}$ con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de $\mathcal{L}$ a cero
- $\nabla\mathcal{L} = [\frac{\partial\mathcal{L}}{\partial\mu}\frac{\partial\mathcal{L}}{\partial\sigma}] = [0\,0]$ lo cual da:
- $\mu = \frac{\sum_i x_i}{N}$, estimador de máxima verosimilitud de la media (insesgado)

Máxima Verosimilitud

- Tomamos el ejemplo de la distribución Gaussiana
- $X \sim \mathcal{N}(\mu, \sigma^2)$
- La pdf es $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Entonces para $\mathcal{X} = \{x_i\}$ con pdf como la anterior, la LL:
- $\mathcal{L}(\mu, \sigma) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_i (x_i-\mu)^2}{2\sigma^2}$
- Para encontrar el estimador de máxima verosimilitud debemos igualar el gradiente de $\mathcal{L}$ a cero
- $\nabla\mathcal{L} = [\frac{\partial\mathcal{L}}{\partial\mu}\frac{\partial\mathcal{L}}{\partial\sigma}] = [0\,0]$ lo cual da:
- $\mu = \frac{\sum_i x_i}{N}$, estimador de máxima verosimilitud de la media (insesgado)
- $s^2 = \frac{\sum_i (x_i-\mu)^2}{N}$, estimador de máxima verosimilitud de la varianza (sesgado)

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- El estimador de varianza anterior se suele denominar s_N^2

Máxima Verosimilitud

- El estimador de varianza anterior se suele denominar s_N^2
- Es el estimador óptimo de máxima verosimilitud para la distribución Gaussiana

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Máxima Verosimilitud

- El estimador de varianza anterior se suele denominar s_N^2
- Es el estimador óptimo de máxima verosimilitud para la distribución Gaussiana
- Es asintóticamente insesgado, es decir el sesgo tiende a cero cuando $N \to \infty$

Máxima Verosimilitud

- El estimador de varianza anterior se suele denominar s_N^2
- Es el estimador óptimo de máxima verosimilitud para la distribución Gaussiana
- Es asintóticamente insesgado, es decir el sesgo tiende a cero cuando $N \to \infty$
- En el infinito (N grande), el estimador s_N^2 y el estimador insesgado $s_{N-1}^2 = \frac{\sum_i (x_i - \mu)^2}{N-1}$ coinciden

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es $B_1 = [0, \frac{1}{M}], B_2 = [\frac{1}{M}, \frac{2}{M}), B_M = [\frac{M-1}{M}, 1]$

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es $B_1 = [0, \frac{1}{M}], B_2 = [\frac{1}{M}, \frac{2}{M}), B_M = [\frac{M-1}{M}, 1]$
- Para cada $x \in B_l$, la estimación de densidad es:

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es $B_1 = [0, \frac{1}{M}], B_2 = [\frac{1}{M}, \frac{2}{M}), B_M = [\frac{M-1}{M}, 1]$
- Para cada $x \in B_l$, la estimación de densidad es:
- $\hat{p}_n(x) = \frac{\text{no. observaciones en } B_l}{n} \times \frac{1}{\text{longitud bin}}$

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es $B_1 = [0, \frac{1}{M}], B_2 = [\frac{1}{M}, \frac{2}{M}), B_M = [\frac{M-1}{M}, 1]$
- Para cada $x \in B_l$, la estimación de densidad es:
- $\hat{p}_n(x) = \frac{\text{no. observaciones en } B_l}{n} \times \frac{1}{\text{longitud bin}}$
- $\hat{p}_n(x) = \frac{M}{n} \sum_{i=1}^{n} I(X_i \in B_l)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad usando Histograma

- Planteo del problema: queremos estimar la función de densidad de probabilidad $p_X(x)$
- Podemos usar el histograma para tal fin
- Por simplicidad, asumamos que tenemos n observaciones $x_i \in [0, 1]$
- Un histograma particiona el conjunto $[0, 1]$ en M porciones (bins)
- Vamos a usar el número de apariciones de la v.a. en cada bin para estimar la densidad
- Cada bin es $B_1 = [0, \frac{1}{M}], B_2 = [\frac{1}{M}, \frac{2}{M}), B_M = [\frac{M-1}{M}, 1]$
- Para cada $x \in B_l$, la estimación de densidad es:
- $\hat{p}_n(x) = \frac{\text{no. observaciones en } B_l}{n} \times \frac{1}{\text{longitud bin}}$
- $\hat{p}_n(x) = \frac{M}{n} \sum_{i=1}^{n} I(X_i \in B_l)$
- $I(X_i \in B_l)$ es la función indicador, definida por $I(X_i \in B_l) = 1$ si $X_i \in B_l$ y $I(X_i \in B_l) = 0$ si no

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins M

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins M
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins M
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins M
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador
- $\text{var}(\hat{p}_n(x)) = M\frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins M
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador
- $\text{var}(\hat{p}_n(x)) = M\frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$
- La varianza aumenta con el número de bins y decrece con el número de observaciones

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad usando Histograma

- Calculamos ahora la esperanza de la estimación
- $E[\hat{p}_n(x)] = M \times P(X_i \in B_l) = p(x^*),\ x^* \in [\frac{l-1}{M}, \frac{l}{M}]$
- El sesgo o bias del estimador es:
- $b(\hat{p}_n(x)) = E[\hat{p}_n(x)] - p(x) \propto \frac{1}{M}$
- Es decir, el bias decrece al aumentar el número de bins M
- Esto tiene sentido porque al tener más bins, tenemos mejor resolución de la densidad
- Encontramos ahora la varianza del estimador
- $\text{var}(\hat{p}_n(x)) = M\frac{p(x^*)}{n} + \frac{p^2(x^*)}{n}$
- La varianza aumenta con el número de bins y decrece con el número de observaciones
- Existe un M óptimo que minimiza el $MSE = \text{var} + b^2$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra x_i y lo transforma en una "montaña" (bump)

Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra x_i y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad

Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra x_i y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:

Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra x_i y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:
- $\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x_i - x}{h})$

Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra x_i y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:
- $\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x_i - x}{h})$
- $K(\cdot)$ es la función de kernel, gralmente. una función suave y simétrica como la Gaussiana

Estimación de Densidad de Kernel

- La idea de un estimador de kernel es suavizar cada muestra x_i y lo transforma en una "montaña" (bump)
- Luego todas las montañas se suman para conformar la estimación de densidad
- Sea el estimador de densidad de kernel:
- $\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x_i - x}{h})$
- $K(\cdot)$ es la función de kernel, gralmente. una función suave y simétrica como la Gaussiana
- $h > 0$ es el ancho de banda de filtrado

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$, es decir que sea pdf

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$, es decir que sea pdf
- $\lim_{t\to-\infty} K(x) = \lim_{t\to+\infty} K(x) = 0$

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$, es decir que sea pdf
- $\lim_{t\to-\infty} K(x) = \lim_{t\to+\infty} K(x) = 0$
- Funciones kernel comunes:

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$, es decir que sea pdf
- $\lim_{t\to-\infty} K(x) = \lim_{t\to+\infty} K(x) = 0$
- Funciones kernel comunes:
- Gaussiana: $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$, es decir que sea pdf
- $\lim_{t\to-\infty} K(x) = \lim_{t\to+\infty} K(x) = 0$
- Funciones kernel comunes:
- Gaussiana: $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$
- Uniforme: $K(x) = I(-1 \leq x \leq 1)$

Estimación de Densidad de Kernel

- Cuando el ancho de banda h es chico, las curvas tienen picos de alta frecuencia (*undersmoothing*)
- Cuando h es muy grande, se filtran los detalles y se pierde información (*oversmoothing*)
- Para elegir una función de kernel $K(\cdot)$ debemos considerar que:
- Sea simétrica
- $\int K(x)dx = 1$, es decir que sea pdf
- $\lim_{t\to-\infty} K(x) = \lim_{t\to+\infty} K(x) = 0$
- Funciones kernel comunes:
- Gaussiana: $K(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$
- Uniforme: $K(x) = I(-1 \leq x \leq 1)$
- Epanechnikov (mínimo MSE): $K(x) = \frac{3}{4} \max\{1 - x^2, 0\}$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado, x_0

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado, x_0
- El sesgo está dado por:

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado, x_0
- El sesgo está dado por:
- $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado, x_0
- El sesgo está dado por:
- $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$
- Obviando la demostración tenemos:
  $b(\hat{p}_n(x_0)) = \frac{1}{2}h^2 p''(x_0)\int y^2 H(y)dy + \mathcal{O}(h^3)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado, x_0
- El sesgo está dado por:
- $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$
- Obviando la demostración tenemos:
  $b(\hat{p}_n(x_0)) = \frac{1}{2}h^2 p''(x_0)\int y^2 H(y)dy + \mathcal{O}(h^3)$
- Esto significa que cuando $h \to 0$ el bias decrece como $\mathcal{O}(h^2)$

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Como dijimos antes, queremos estimar la pdf $p_X(x)$
- Estudiamos el sesgo, la varianza y MSE del estimador en un punto dado, x_0
- El sesgo está dado por:
- $b(\hat{p}_n(x_0)) = E[\hat{p}_n(x_0)] - p(x_0) = E\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) - p(x_0)$
- Obviando la demostración tenemos:
  $b(\hat{p}_n(x_0)) = \frac{1}{2}h^2 p''(x_0)\int y^2 H(y)dy + \mathcal{O}(h^3)$
- Esto significa que cuando $h \to 0$ el bias decrece como $\mathcal{O}(h^2)$
- El término $p''(x_0)$ indica que la curvatura de la pdf (desconocida) incrementa el sesgo porque es estimador de kernel suaviza las curvas

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) =$
  $\frac{1}{nh} p(x_0) \int K^2(y)dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$

Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y)dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa $\mathcal{O}\left(\frac{1}{nh}\right)$ cuando $n \to \infty, h \to 0$

Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y)dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa $\mathcal{O}\left(\frac{1}{nh}\right)$ cuando $n \to \infty, h \to 0$
- Por último, estudiamos el MSE del estimador:

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y)dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa $\mathcal{O}\left(\frac{1}{nh}\right)$ cuando $n \to \infty, h \to 0$
- Por último, estudiamos el MSE del estimador:
- $\text{MSE}(\hat{p}_n(x_0)) = b^2 + \text{var} = \mathcal{O}(h^4) + \mathcal{O}\left(\frac{1}{nh}\right)$

Estimación de Densidad de Kernel

- Analizamos ahora la varianza de la estimación de kernel
- $\text{var}(\hat{p}_n(x_0)) = \text{var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right)\right) = \frac{1}{nh} p(x_0) \int K^2(y)dy + \mathcal{O}\left(\left(\frac{1}{nh}\right)^2\right)$
- Es decir la varianza se reduce a una tasa $\mathcal{O}\left(\frac{1}{nh}\right)$ cuando $n \to \infty, h \to 0$
- Por último, estudiamos el MSE del estimador:
- $\text{MSE}(\hat{p}_n(x_0)) = b^2 + \text{var} = \mathcal{O}(h^4) + \mathcal{O}\left(\frac{1}{nh}\right)$
- Podemos elegir h tal que se minimice el MSE asintóticamente, esto se logra con $h \propto n^{-1/5}$, lo cual logra un $\text{MSE} = \mathcal{O}(n^{-4/5})$

Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ($\mathcal{O}(n^{-2/3})$) pero más lenta que máxima verosimilitud ($\mathcal{O}(n^{-1})$)

Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ($\mathcal{O}(n^{-2/3})$) pero más lenta que máxima verosimilitud ($\mathcal{O}(n^{-1})$)
- Sin embargo en estimación de kernel no asumimos ninguna distribución como en máxima verosimilitud

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ($\mathcal{O}(n^{-2/3})$) pero más lenta que máxima verosimilitud ($\mathcal{O}(n^{-1})$)
- Sin embargo en estimación de kernel no asumimos ninguna distribución como en máxima verosimilitud
- Solamente asumimos que la distribución original es suave y diferenciable

Estimación de Densidad de Kernel

- La estimación de densidad de kernel es más rápida que el estimador óptimo del método con histograma ($\mathcal{O}(n^{-2/3})$) pero más lenta que máxima verosimilitud ($\mathcal{O}(n^{-1})$)
- Sin embargo en estimación de kernel no asumimos ninguna distribución como en máxima verosimilitud
- Solamente asumimos que la distribución original es suave y diferenciable
- Este es el precio a pagar por una incremento en la flexibilidad de la estimación (i.e., menos hipótesis)

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Ejercicio 1

(a) Sea X una v.a. con media μ y varianza v

(b) Sean $Y_1, \ldots, Y_n$ mediciones de la forma $Y_i = X + W_i$

(c) Sean W_i v.a. con media cero y varianza v_i

(d) Asumamos que $X, W_1, \ldots, W_n$ son independientes

(e) Demostrar que el estimador lineal de cuadrados mínimos de X en base a las mediciones $Y_1, \ldots, Y_n$ es:

(f) $\hat{X} = \frac{(\mu/v)+\sum_{i=1}^{n}(Y_i/v_i)}{(1/v)+\sum_{i=1}^{n}(1/v_i)}$

(g) Simular para $n = 10$, $n = 1000$, $v = 0.1$, $v = 100$

(h) Qué conclusión se puede sacar para valores de n y/o v grandes?

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Ejercicio 2

(a) Se tira una moneda 100 veces, y salen 55 cecas.

(b) Encontrar el estimador de máxima verosimilitud de la probabilidad de ceca p

(c) Simular el experimento y encontrar por computadora el valor de la estimación de máxima verosimilitud de p

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Ejercicio 3

(a) Una v.a. continua X responde a un proceso Gaussiano de media cero y $\sigma^2 = 1$

(b) Simular varias realizaciones de X y estimar la pdf usando el método del histograma

(c) Variar la cantidad de bins y sacar conclusiones acerca de la calidad de estimación

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Ejercicio 4

(a) Repetir el ejercicio 3 usando un estimador de kernel con una función Gaussiana

(b) Comparar resultados entre ambos ejercicios

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Bibliografía I