Estadística en Analítica

2023-2

Pablo A. Saldarriaga
psaldar2@eafit.edu.co

UNIVERSIDAD EAFIT

Programación del curso

Evaluación:

- ✓ Parcial ***(25%)***
- ✓ Seguimiento ***(40%)***
  - ✓ Taller 1 ***(20%)***
  - ✓ Taller 2 ***(20%)***
- ✓ Proyecto Integrador ***(35%)***

Sesiones de Clase

Fecha	Comentario
Octubre 20	Inicio del curso
Octubre 21	
Octubre 27	
Octubre 28	
Noviembre 3	Entrega Taller 1
Noviembre 4	
Noviembre 10	
Noviembre 11	Ultima clase teorica
Noviembre 17	Parcial y Entrega Taller 2

Texto guía: *The Elements of Statistical Learning (2009)*

Contenido del Curso

0. Preliminares y pre-procesamiento de datos

- ✓ Repaso de probabilidad
- ✓ Variables aleatorias
- ✓ Inferencia Estadistica (estimadores y propiedades)
- ✓ Intervalos de confianza y pruebas de hipotesis
- ✓ Estadísticos univariados y multivariados

1. Técnicas Supervisadas

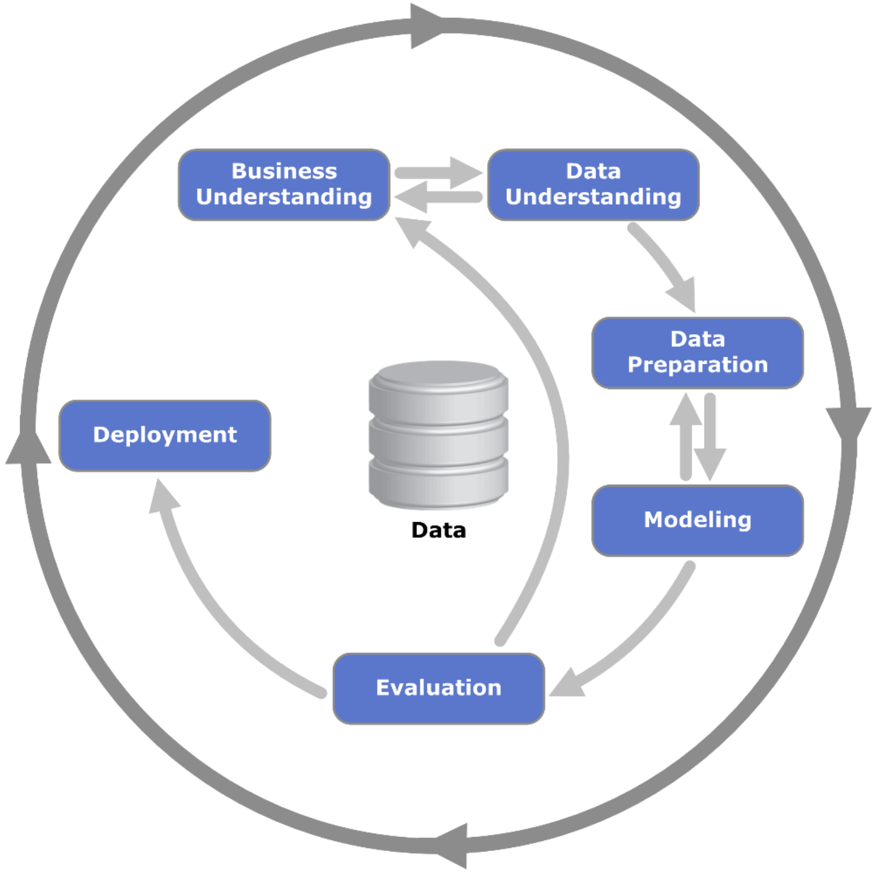
- ✓ Modelos de regresión
- ✓ Modelos de clasificación
- ✓ Estrategia de regularización
- ✓ Métricas de desempeño de regresión y clasificación
- ✓ Sesgo VS Varianza
- ✓ Validación Cruzada

2. Técnicas No Supervisadas

- ✓ Reduccion de Dimensionalidad
  - ✓ PCA
  - ✓ Umap

- ✓ Análisis de cluster
  - ✓ Métodos jerárquicos
  - ✓ Métodos no jerárquicos
  - ✓ Selección de cantidad de grupos
  - ✓ Métricas de validación de cluster

Metodología CRISP-DM
(Cross Industry Standard Process for Data Mining)



Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. Step-by-step data mining guide (2000).

Preliminares: Probabilidad e Inferencia

Probabilidad

Qué tan frecuente ocurre algo

 # de registros que cumplen la condición en estudio

¡Enfoque frecuentista!

$$\frac{\#\ Casos\ Posibles}{\#\ Casos\ Totales}$$

Tamaño de la muestra

Probabilidad Condicional

Es el cálculo de una probabilidad acotando el espacio muestral que se considera

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

	Ocurre B	No ocurre B
Ocurre A	N1	N2
No ocurre A	N3	N4

$$\frac{N1/(N1+N2+N3+N4)}{(N1+N3)/(N1+N2+N3+N4)}$$

$$\frac{N1}{N1+N3}$$

Probabilidad Condicional

	Credito Rechazado	Credito Aprobado	Totales
Cliente Nuevo	210.000	140.000	**350.000**
Cliente Recurrente	1.500	28.500	**30.000**
Totales	**211.500**	**168.500**	**380.000**

Probabilidad Total

Sean $A_1, A_1, \ldots, A_1$ eventos excluyentes, tal que $U_{i=1}^{n} A_i = S$, además *B* es un evento de *S*

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B|A_i)$$

Teorema de Bayes

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\boxed{P(B)}}$$

Probabilidad Total

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$$

Variables Aleatorias

Una variable aleatoria X es una función real definida en espacio de probabilidad asociado a un experimento aleatorio

$$X\colon \Omega \to \mathbb{R}$$

Función de densidad y distribución

Para cualquier distribución de probabilidad discreta, lo siguiente debe ser verdadero:

1. $0 \leq p(y) \leq 1$ para toda y.
2. $\sum_y p(y) = 1$, donde la sumatoria es para todos los valores de y con probabilidad diferente de cero.

Propiedades de una función de densidad Si $f(y)$ es una función de densidad para una variable aleatoria continua, entonces

1. $f(y) \geq 0$ para toda y, $-\infty < y < \infty$.
2. $\int_{-\infty}^{\infty} f(y)\, dy = 1$.

Las probabilidades puntuales son 0

Distribuciones de Probabilidad

Distribuciones

Discretas

- ✓ *Uniforme*
- ✓ *Binomial*
- ✓ *Poisson*
- ✓ *Otras!*

Continuas

- ✓ *Uniforme*
- ✓ *Normal*
- ✓ *Exponencial*
- ✓ *Otras!*

Función Generadora de Momentos

Promedio

$$\mu = E[X]$$

Varianza

$$\sigma^2 = E[X^2] - (E[X])^2$$

$$M_x(t) = E[e^{tx}]$$

$$M_x(t) = \sum_{x=1}^{\infty} e^{tx} P(X = x)$$

$$M_x(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

Dist. Discreta: Uniforme

$X \sim Unif\{x1, x2, \ldots, xn\}$

$$P(X = x) = \begin{cases} \dfrac{1}{n} & si\ x = x1, x2, \ldots, x2 \\ 0 & En\ otro\ caso \end{cases}$$

$$\mu = \frac{n+1}{2}$$

$$\sigma^2 = \frac{n^2 - 1}{12}$$

Dist. Discreta: Binomial

$X \sim Bin(n, p)$

$$P(X = x) = \begin{cases} \dfrac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x} & si\ x = x1, x2, \ldots, x2 \\ 0 & En\ otro\ caso \end{cases}$$

$\mu = np$

$\sigma^2 = np(1-p)$

Modela la cantidad de éxitos en n sucesos

Dist. Discreta: Poisson

$X \sim Poisson(\lambda)$

$$P(X = x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!} & si\ x = 0, 1, \ldots \\ 0 & En\ otro\ caso \end{cases}$$

$\mu = \lambda$

$\sigma^2 = \lambda$

Modela la cantidad de eventos que ocurren en un periodo de tiempo

Dist. Continua: Uniforme

$X \sim Unif(a, b)$

$$f(x) = \begin{cases} \dfrac{1}{b-a} & si\ a < x < b \\ 0 & En\ otro\ caso \end{cases}$$

$$\mu = \frac{a+b}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Dist. Continua: Normal

$X \sim N(\mu, \sigma)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu = \mu$

$\sigma^2 = \sigma^2$

Dist. Continua: Exponencial

$X \sim Exp(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & si\ x > 0 \\ 0 & En\ otro\ caso \end{cases}$$

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

Modela tiempos de espera para la ocurrencia de eventos

De binomial a Normal

$$X \sim Bin(n, p)$$

$$X \sim Poisson(\lambda = np)$$

$$X \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$$

$$\boldsymbol{n} \rightarrow \infty$$

Inspira
Crea
Transforma