Universidad de Alcalá

Escuela Politécnica Superior

Universidad de Alcalá

Fundamentos de Ciencia de Datos

Juan J. Cuadrado Gallego

Martes 10:00 – 12:00

Grado en Ingeniería Informática – Curso 2019/2020

Marcos Barranquero Fernández – 51129104N

TEMA 1 – PROBABILIDAD Y ESTADÍSTICA

FRECUENCIA

F. Absoluta(f_i)	**F. Relativa** (f_{ri})	**F. Abs/Rel Acum.**
$Nº\ Apariciones\ elemento$	$\frac{Nº\ apariciones\ elemento}{Elementos\ totales}$	$\sum_{i=0}^{indice\ elem.} f_i/f_{ri}(elemento_i$

MEDIA

Aritmética $(\overline{x_a})$	**Geométrica** $(\overline{x_g})$	**Armónica** $((\overline{x_a})^{-1})$:
$\frac{\sum_{i=0}^{n} x_i}{n}$	$\overline{x_g} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$	$\frac{\sum_{i=0}^{n} \frac{1}{x_i}}{n}$

CLASES DE EQUIVALENCIA

Rango	**Límite**	**Amplitud**	**Marca**
$v_{sup} - v_{inf}$	• Superior: l_{sup} • Inferior: l_{inf}	$l_{sup} - l_{inf}$	$(l_{sup} + l_{inf})/2$

MODA

La **moda** es el **dato que aparece más veces** en un conjunto.

MEDIDAS DE DISPERSIÓN

Desviación típica o estándar (σ)	**Desviación media** (s)	**Varianza** (σ^2)	**CV. de Pearson**
$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$	$\frac{\sum_{i=1}^{n}\lvert(x_i - \bar{x})\rvert}{n}$	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$	$cv_{\bar{x}} = \frac{\sigma}{\bar{x}}$

TEOREMA DE TCHEBYCHEV

Pertenencia	**Amplitud**	**Intervalo**
$pertenencia = 1 - \frac{1}{k^2}$	$amplitud = k \cdot \sigma$	$[\bar{x} - amplitud, \bar{x} + amplitud]$

El porcentaje de datos $pertenencia$ se encontrará representado en el $intervalo$.

MEDIDAS DE ORDENACIÓN

Cuartil	Decil	Percentil	Centil
$n_c = \frac{1}{4} n_t$	$n_c = \frac{1}{10} n_t$	$n_c = \frac{1}{100} n_t$	$n_c = \frac{1}{n} n_t$

- Si n_c es un número natural:
  - Si n_t es un número impar: $Q_i = x_{nc}$
  - Si n_t es un número par: $Q_i = \frac{x_{nc\downarrow} + x_{nc\uparrow}}{2}$
- Si n_c es un número real: $Q_i = \frac{x_{nc\downarrow} + x_{nc\uparrow}}{2}$

REGLAS DE PROBABILIDAD

- $0 \leq P(A) \leq 1$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $Si\ A \subset B \rightarrow P(A) \geq P(B)$
- $P(E) = 1$
- $P(\emptyset) = 0$

REGLAS DE CONJUNTOS

- $B \cup C = \{elementos(B) + elementos(C)\}$
- $B \cap C = \{elementos\ compartidos(B, C)\}$
- $B - C = \{elementos(B) - elementos\ compartidos(B, C)\}$
- $B\ \Delta\ C = \{\ elementos(B) + elementos(C) - elementos\ compartidos(B, C)\ \}$

TEMA 2 – ASOCIACIÓN

SOPORTE Y CONFIANZA

Soporte	Confianza
$S(C) = \frac{numero\ de\ apariciones(C)}{tamaño\ del\ espacio\ muestral}$	$C(C_1 \rightarrow C_2) = \frac{numero\ de\ apariciones(C_1 \cup C_2)}{numero\ de\ apariciones(C_1)}$

APRIORI

1. Se calcula soporte de los sucesos elementales de ese espacio muestral, y se descartan sucesos elementales que no superen el umbral de soporte.
2. Se generan sucesos candidatos empleando Apiori-Gen. Partiendo de los sucesos elementales, para generar casos de tamaño k+1, se toman sucesos de tamaño k que compartan k-1 elementos.
3. Se calcula soporte para los sucesos candidatos generados, y se descartan aquellos que no superen el umbral de soporte.
4. Se generan todas las asociaciones (permutaciones) posibles partiendo de los sucesos candidatos. Para cada asociación, se calcula la confianza y se descartan las asociaciones que no superen el umbral de confianza.
5. Nos quedamos con los casos no descartados.

PROPIEDADES

$soporte(caso) \geq soporte(subcaso\ que\ compone\ el\ caso)$
$Si\ C(A \rightarrow (B - A)) \leq\ umbral\ de\ confianza, entonces$ $C(A' \rightarrow (B - A')) \leq\ umbral\ de\ confianza$ $donde\ A' \subseteq A$

TEMA 3 – CLASIFICACIÓN SUPERVISADA

ÁRBOLES DE DECISIÓN – ALGORITMO DE HUNT

1. Obtener suceso elemental en nodo inicial (no puede ser el clasificador).
2. Clasificar sucesos en base a nodo elegido.
    a. Si podemos clasificar todos, hemos finalizado.
    b. Si no, obtener suceso elemental en nodo intermedio y ejecutar paso 2.

GANANCIA DE INFORMACIÓN

Cálculo de la impureza		
$\Delta Impureza = I_{padre} \cdot freq_{padre} - I_{hijos} \cdot freq_{hijos}$		
$I_{hijos} = \sum_{j=1}^{k} \frac{N_{sucesos}(Nodo_j)}{N_{total\ sucesos}} I(Nodo_j)$		
Entropía	**Error**	**Gini**
$-\sum(freq_i(nodo_i) \cdot log_2(freq_i(nodo_i)))$	$1 - max(freq_i(nodo_i))$	$1 - \sum(freq_i(nodo_i))^2$

REGRESIÓN – COVARIANZA Y CORRELACIÓN

Covarianza	**Correlación [0,1]**
$S_{xy} = \frac{(\sum_{i=0}^{n} x_i \cdot y_i)}{n} - (\bar{x} \cdot \bar{y})$	$r_{xy} = \frac{S_{xy}}{s_x \cdot S_y}$
• $S_{xy} > 0$ →Dependencia directa positiva. • $S_{xy} = 0$ →No existe relación lineal. • $S_{xy} > 0$ →Dependencia inversa negativa.	• (0,1] → relación ascendente. • 0 → no hay correlación. • [-1,0) → relación descendente. Se consideran fuertemente correlacionadas a partir de 0'8.

REGRESIÓN POR MÍNIMOS CUADRADOS

$y = a + bx$	• $a = \bar{y} - b \cdot x$ • $b = \frac{Sxy}{(s_x)^2}$

ANÁLISIS REGRESIÓN

$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})$	• $\hat{y}_i$ = valor calculado con la recta de regresión, tomando su x_i asociado. • $\bar{y}$ = media de y.
$SSy = \sum_{i=1}^{n} (y_i - \bar{y})$	• y_i = valor original del dato. • $\bar{y}$ = media de y.
$r^2 = \frac{SSR}{SSy}$	• $0 \leq r \leq 1$; Conforme más se aproxime a 1, más correcta es la recta de regresión.

ANÁLISIS DEL ERROR

$Sr = \sqrt{\frac{\sum_{i=1}^{n}((y_i - \hat{y}_i)^2)}{n}}$	• $\hat{y}_i$ = valor calculado con la recta de regresión, tomando su x_i asociado. • y_i = valor original del dato.

Cuanto más cercano a 0, mejor.