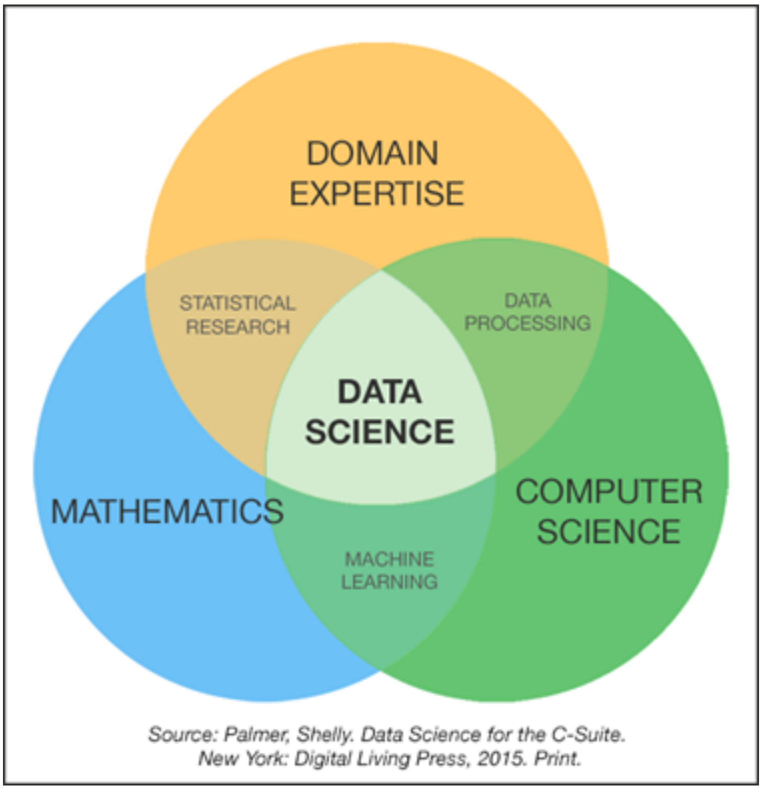
UNIVERSIDAD NACIONAL DE INGENIERÍA

CURSOS PIT
CTIC UNI

CIENCIA DE DATOS:
APRENDE LOS FUNDAMENTOS DE MANERA PRÁCTICA



SESION 03
APRENDIZAJE SUPERVISADO
EDA II

Juan Antonio Chipoco Vidal
jchipoco@gmail.com

*Cuadros e imágenes extraidos de internet solo con fines educativos

UNIVERSIDAD NACIONAL DE INGENIERÍA

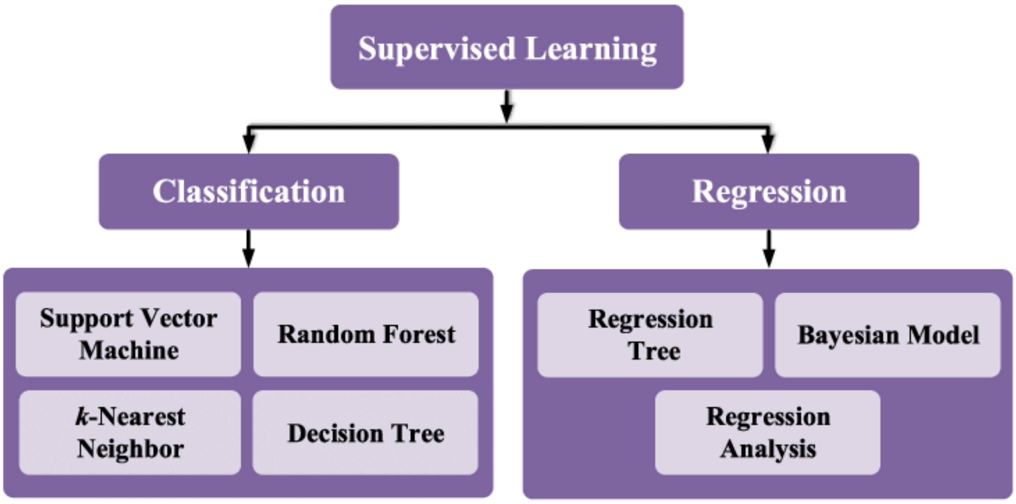
CURSOS PIT
CTIC UNI

ÍNDICE

Objetivo

El objetivo de esta sesion es profundizar en el analisis de la correlacion lineal de dos variables, la cual cuantifica que tan relacionadas estan las mismas. Esta tecnica esta estrechamente relacionada con la ***regresion lineal*** la cual da lugar a una ecuacion que describe dicha relacion en terminos matematicos.

En la practica de esta sesion, continuacion de la practica de la sesion anterior, finalizaremos el analisis exploratorio de datos para poder ya aplicar diversos algoritmos de ***clasificacion*** para obtener la variables objetivo buscada, en este caso la supervicencia o no de un pasajero del Titanic.



Medidas de Dispersion

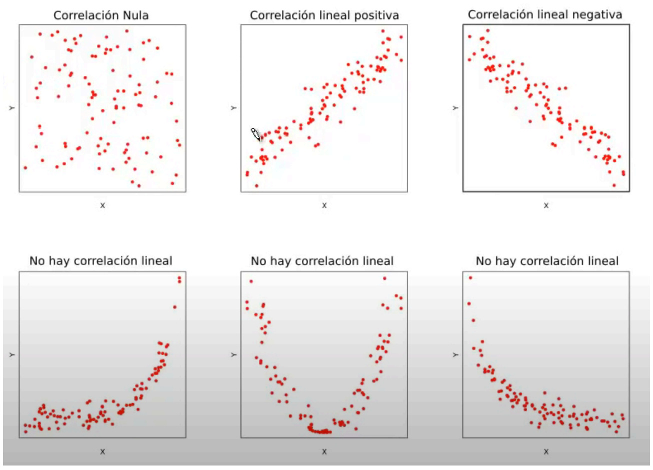
Las medidas de dispersión, se utiliza para describir la variabilidad en una muestra o población. Por lo general, se usa junto con una medida de tendencia central, como la media o la mediana, para proporcionar una descripción general de un conjunto de datos.

Medidas de Dispersion: Correlacion

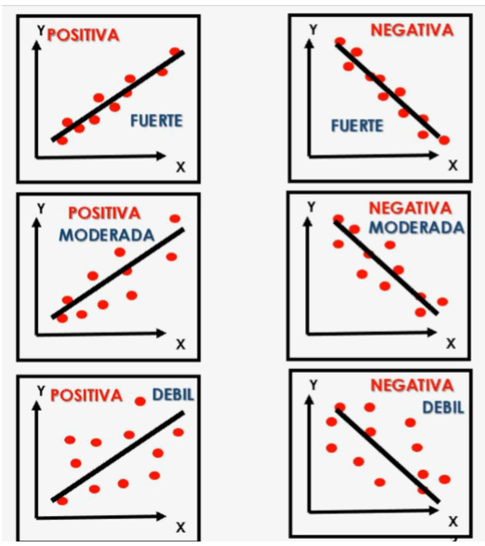
La correlacion sirve para medir la relacion que existe entre dos o mas variables.

La correlacion contesta preguntas como las siguientes:

La practica de algun deporte esta relacionada con una vida mas longeva?
Existe una relacion entre la cantidad de carne ingerida diariamente y el cancer?
Mayor estudio implica mejores notas en un examen?



Si la correlacion es lineal su direccion puede ser positiva o negativa. Su fuerza varia entre perfecta y nula.



Medidas de Dispersion: Correlacion

Medidas de Dispersion: Coeficiente de Correlacion

Para cuantificar las relaciones anteriores tenemos el Coeficiente de Correlacion al cual se le asignara un valor entre -1 y 1.

Este coeficiente nos da una medida de la fuerza y el sentido de una relacion lineal entre variables cuantitativas.

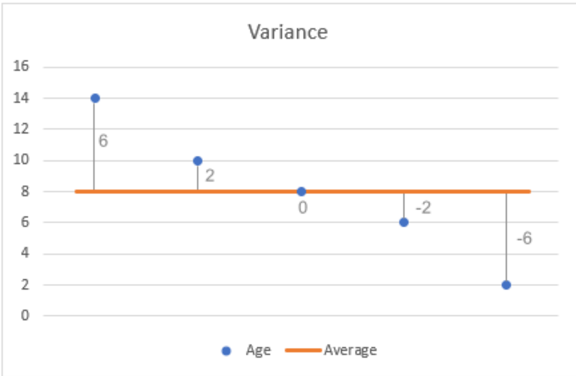
Cuando el signo es positivo la asociacion lineal es positiva lo que implica que cuando el valor de una variable x aumenta tambien aumenta el valor de la otra variable y.

Cuando el signo es negativo la asociacion lineal es negativa lo que implica que cuando el valor de una variable x aumenta el valor de la otra variable y disminuye.

± 0.96 , ± 1.0	PERFECTA
± 0.85 , ± 0.95	FUERTE
± 0.70 , ± 0.84	SIGNIFICATIVA
± 0.50 , ± 0.69	MODERADA
± 0.20 , ± 0.49	DÉBIL
± 0.10 , ± 0.19	MUY DÉBIL
± 0.09 , ± 0.0	NULA

Medidas de Dispersion: Coeficiente de Determinacion R^2

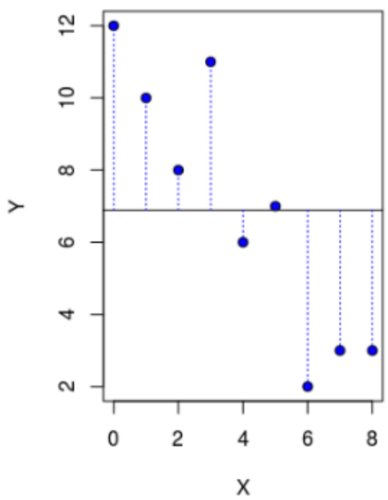
Recordemos que la varianza es la medida de la variabilidad de un conjunto de datos que indica hasta qué punto se distribuyen los diferentes valores. Matemáticamente, se define como la suma de los cuadrados de las diferencias entre una variable y su media, dividido entre el numero de datos.

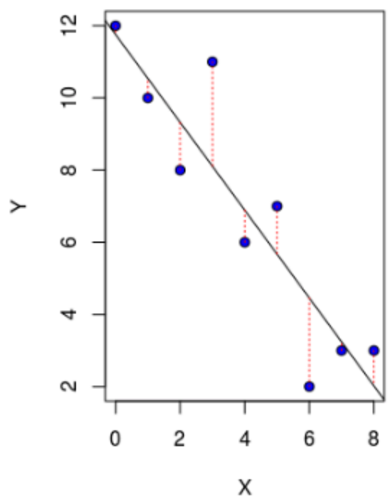


$$Mean = \frac{14 + 10 + 8 + 6 + 2}{5} = 8$$

$$Variance = \frac{6^2 + 2^2 + 0^2 + (-2)^2 + (-6)^2}{5} = 16$$

El 16 nos da una idea de la dispersion de los datos. Un valor de 0 indica que no hay variabilidad, mayor el valor, mayor la dispersion de los datos.

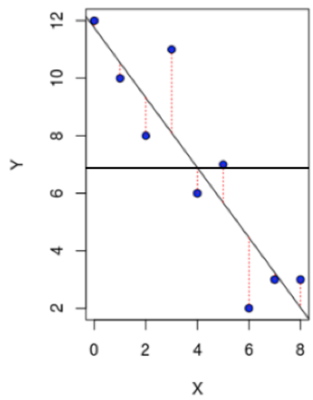




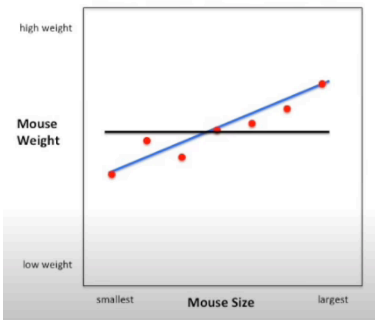
$\text{Var(mean)} = \Sigma\,(y_i - \bar{y})^2$ $\text{Var(line)} = \Sigma\,(y_i - (mx_i + b))^2$

En la grafica anterior trataremos de averiguar que tan bien se ajusta la recta del lado derecho al conjunto de datos. ¿Cual es la bondad del ajuste?.

Medidas de Dispersion: Coeficiente de Determinacion R^2



¿Es este ajuste mejor que el ajuste con la media? Si es asi, ¿Qué tan mejor es? ¿Cómo cuantificamos esta diferencia?



Variables correlacionadas

La suma total de cuadrados de los residuos de la imagen anterior Var(line) representa la variacion del modelo ajustado, o variacion no explicada por el modelo (recta de regresion).

Supongamos que Var(mean) = 32 y Var(line) = 6

Por lo que Var(line)/Var(mean) nos indicara que porcentaje de la variacion total en y (peso del raton) no esta explicada por la variacion en x (tamaño del raton).

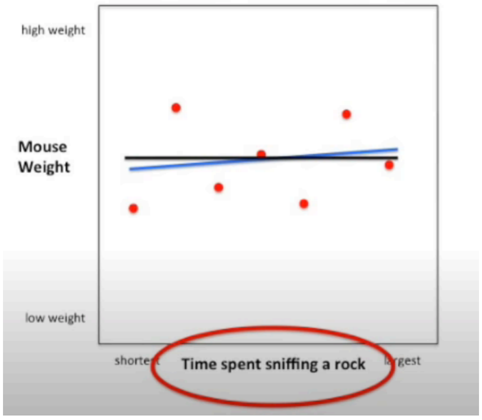
Var(line)/Var(mean) = 6/32 = 19%

Asi pues para saber que porcentaje de la variacion total en y (peso del raton) esta explicada por la variacion en x (tamaño del raton) usamos 1 - Var(line)/Var(mean) = 81%

En otras palabras la relacion entre las dos variables explica el 81% de la variacion de los datos. Esta relacion es significativa.

A este ultimo resultado se el conoce como coeficiente de determinacion R^2

Medidas de Dispersion: Coeficiente de Determinacion R^2



Variables no correlacionadas

Var(mean) = 32 y Var(line) = 30

Var(line)/Var(mean) nos indicara que porcentaje de la variacion total en y (pero del raton) no esta explicada por la variacion en x (tiempo oliendo una roca).

Var(line)/Var(mean) = 30/32 = 94%

Asi pues para saber que porcentaje de la variacion total en y (peso del raton) esta explicada por la variacion en x (tiempo oliendo una roca) usamos 1 - Var(line)/Var(mean) = 6%

En otras palabras la relacion entre las dos variables explica el 6% de la variacion de los datos. Esta relacion no es significativa.

Si el coeficiente de correlacion R = 0.9 entonces el coeficiente de determinacion $R^2 = 0.81$, la relacion entre las dos variables explica el 81% de la variacion de los datos.

R^2 es mas facil de interpretar, por ejemplo que tan mejor es R = 0.7 que R = 0.5

$R^2 = 0.7^2 = 0.49$
$R^2 = 0.5^2 = 0.25$

Con R^2 es facil ver que la primera correlacion es el doble mejor que la segunda correlacion.

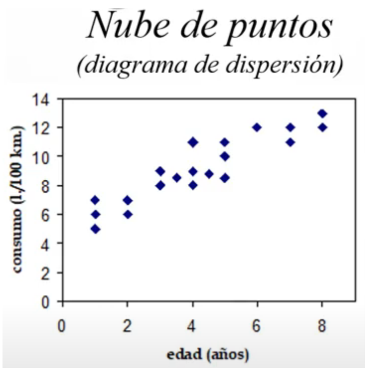
Covarianza y el coeficiente de correlacion

En el siguiente grafico tenemos informacion de la edad de 20 automoviles asi como el consumo de gasolina en litros por cada 100 km según la edad del automovil.

Edad y consumo de gasolina de 20 automóviles

edad auto (años)	Consumo (l/100km)
7	11
5	10
3	8
2	7
7	12
8	12
5	11
4	11
4	8
8	13
1	7
6	12
1	6
3	9
2	6
3,5	8,5
1	5
4,5	8,75
5	8,5
4	9

Al graficar el diagrama de dispersion podemos ver que hay una relacion lineal positiva o directa entre ambas variables.Nos da informacion sobre la covariacion (variacion conjunta) y sus caracteristicas, si es lineal, su signo y su intensidad.



Ahora calculemos la covarianza y el coeficiente de correlacion para estos datos:

	edad auto x_i	consumo y_i	$x_i y_i$	x_i^2	y_i^2
	7	11	77	49	121
	5	10	50	25	100
	3	8	24	9	64
	2	7	14	4	49
	7	12	84	49	144
	8	12	96	64	144
	5	11	55	25	121
	4	11	44	16	121
	4	8	32	16	64
	8	13	104	64	169
	1	7	7	1	49
	6	12	72	36	144
	1	6	6	1	36
	3	9	27	9	81
	2	6	12	4	36
	3,5	8,5	29,75	12,25	72,25
	1	5	5	1	25
	4,5	8,75	39,375	20,25	76,5625
	5	8,5	42,5	25	72,25
	4	9	36	16	81
Totales	84	182,75	856,625	446,5	1770,0625

Covarianza y el coeficiente de correlacion

De la formula de covarianza tenemos:

$$\sigma_{XY} = \sum_{i=1}^{N} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

$$\sigma_{XY} = 4.4548$$

$$S_{XY} \begin{cases} > 0 & \Rightarrow \text{covariación lineal directa (positiva)} \\ < 0 & \Rightarrow \text{covariación lineal inversa (negativa)} \\ = 0 & \Rightarrow \text{no hay covariación lineal} \end{cases}$$

De la formula del coeficiente de correlacion:

$$\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$$

$$\rho_{XY} = 0.9194$$

Se trata entonces de una relación directa o positiva y muy fuerte.