Machine Learning / Aprendizaje Automático



Diego M. Jiménez Bravo dmjimenez@usal.es 16/04/2019



Índice





- I. Quiénes somos
- II. Introducción
- III. Métodos de validación
- IV. Aprendizaje supervisado
- V. Aprendizaje no supervisado
- VI. Aprendizaje por refuerzo
- VII. Ejemplos en código
- VIII.Referencias

Quiénes somos

Salamanca















Universidad de Salamanca

VNIVERSIDAD
BSALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

- 800 años de historia
- Entre las 10 más antiguas del mundo









ESALab

Expert Systems and Applications Lab

 Grupo de investigación multidisciplinar reconocido desde septiembre de 2018

- 16 investigadores
 - Ingenieros mecánicos
 - Ingenieros eléctricos
 - Ingenieros textiles
 - Ingenieros informáticos
 - Estadísticos
 - Técnicos electrónicos







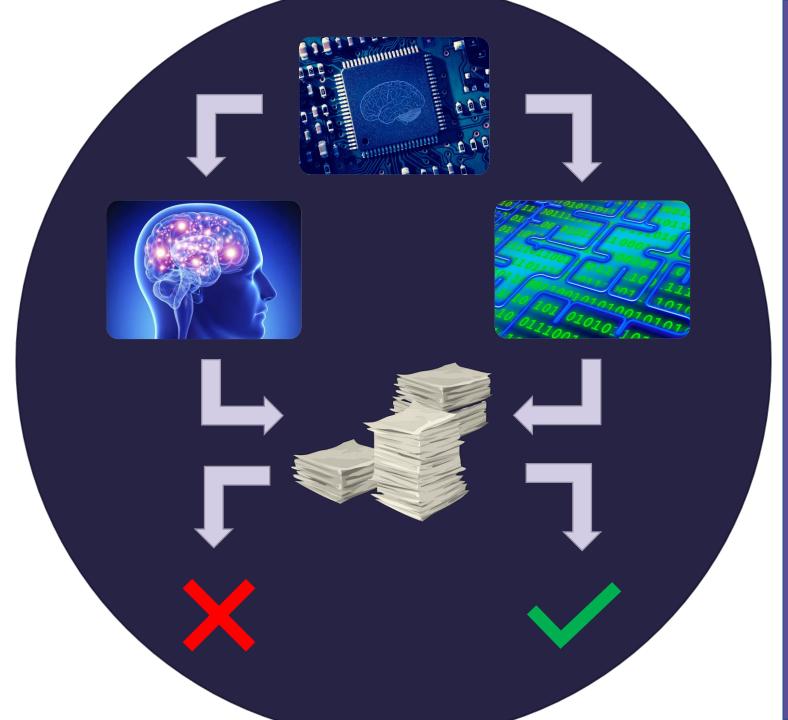


Introducción

¿Qué es el machine learning?

Desarrollo de técnicas permiten que los ordenadores "aprendan". Es decir, se trata de generar programas informáticos de generalizar capaces comportamientos a través de información suministrada.

"Hacer que el **ordenador logre** tomar decisiones como lo hace un humano."









Algoritmos en machine learning





Aprendizaje no supervisado

Aprendizaje por refuerzo

圕



Aprendizaje supervisado



Pelota



Mesa



Ordenador



Pelota



Mesa



Ordenador



Pelota

Introducción

Aprendizaje no supervisado





















Aprendizaje no supervisado









F



Aprendizaje no supervisado













Reducción de la dimensionalidad







Aprendizaje por refuerzo





Algoritmos de aprendizaje automático

Aprendizaje supervisado

- Regresión lineal
- Regresión logística
- Árboles de decisión
- Redes neuronales
- Clasificador de Bayes

Aprendizaje no supervisado

- K-means
- Jerárquico
- Análisis de componentes principales

Aprendizaje por refuerzo

Q Learning

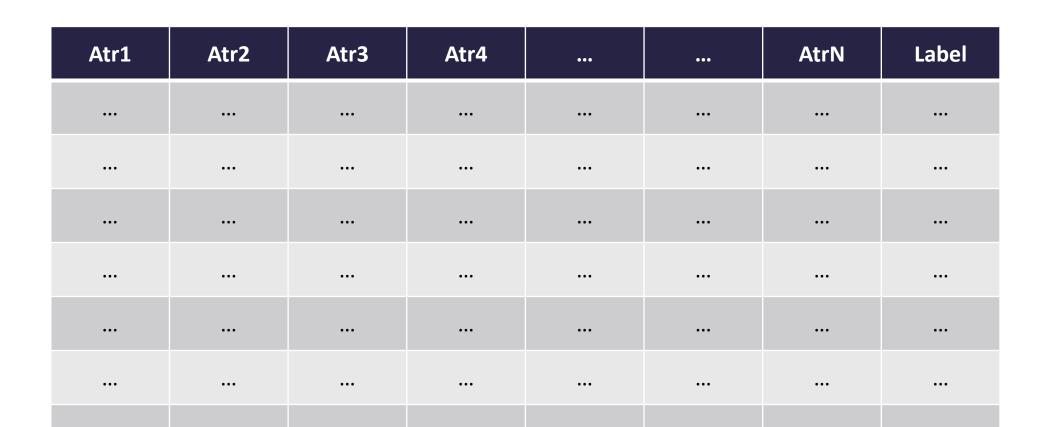
Métodos de validación

Datos

• • •

• • •

• • •



•••

• • •

• • •

• • •

• • •

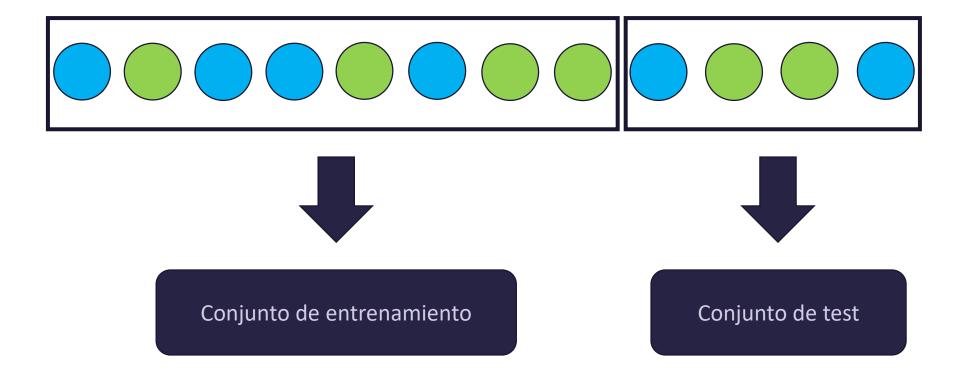




Holdout

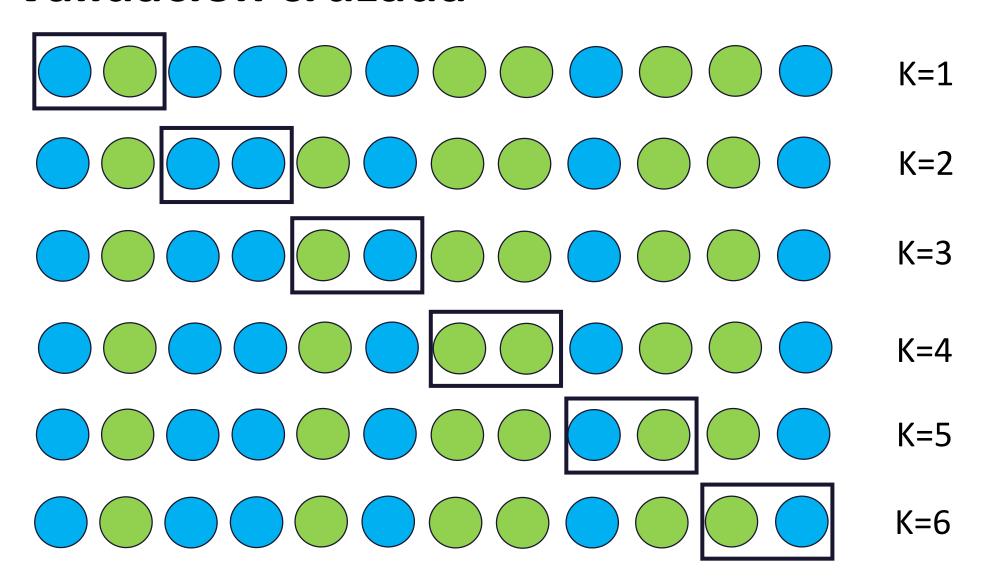






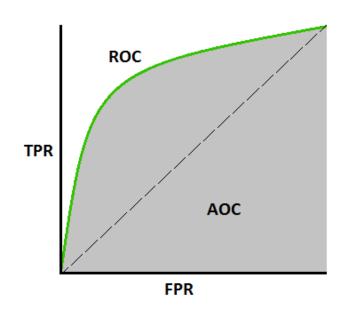


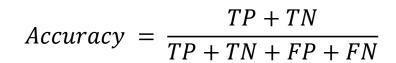
Validación cruzada



Métricas

	Real +	Real -
Estimado +	True Positives (TP)	False Positives (FP)
Estimado -	False Negatives (FN)	True Negatives (TN)





$$Precission = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$TPR = \frac{TP}{TP + FN}$$
 $FPR = \frac{FP}{FP + TN}$

$$FNR = \frac{FN}{TP + FN}$$
 $TNR = \frac{TN}{FP + TN}$





Aprendizaje supervisado



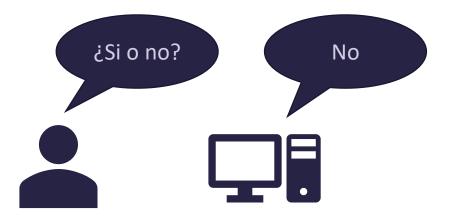
Introducción

Regresión lineal



Clasificación





Regresión lineal







Coche #1 2 asientos 500 €



Coche #2 3 asientos 700 €



Coche #3 4 asientos 900 €



Coche #4 5 asientos 1100 €



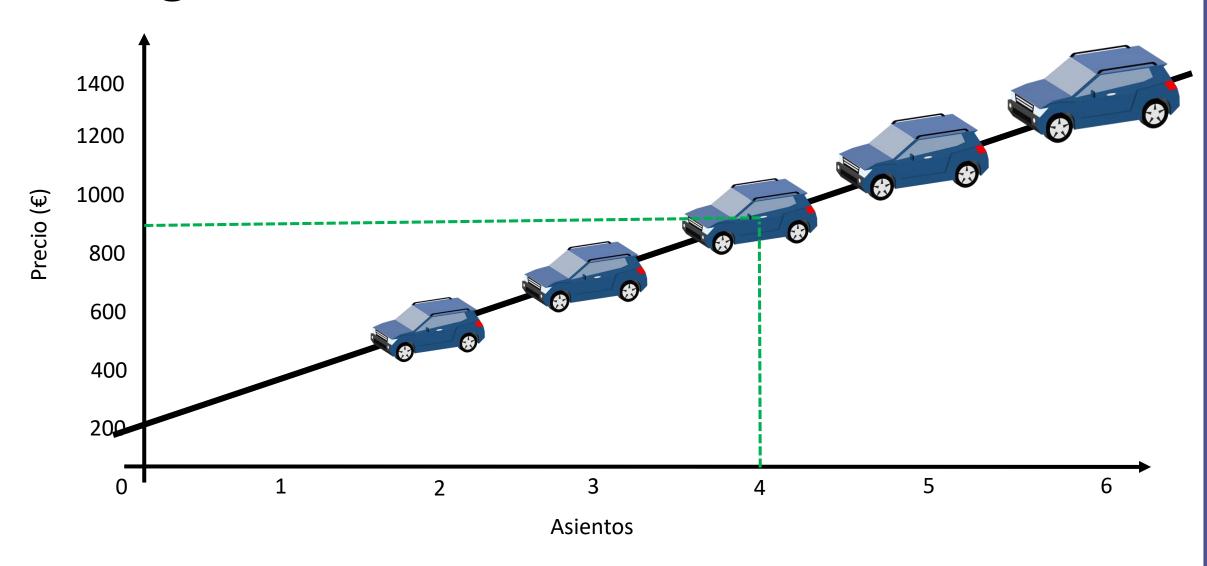
Coche #5 6 asientos 1300 €

Precio = 100 € + 200 € · (nº asientos)

F



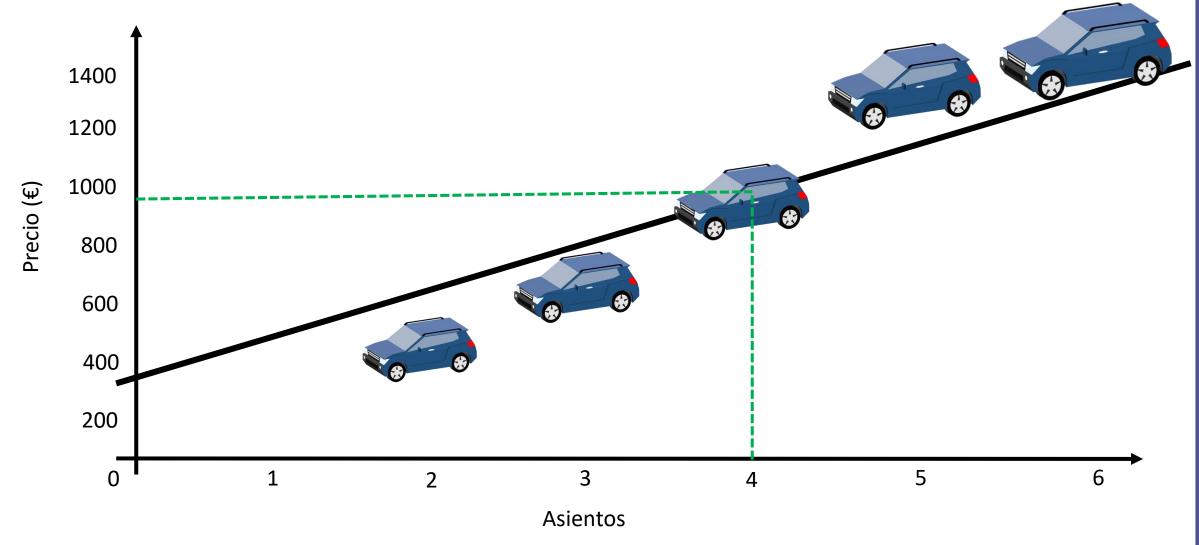
Regresión lineal



F



Regresión lineal



Regresión lineal







Coche #1 2 asientos 80 CV 500 Km 700 €



Coche #2
3 asientos
75 CV
500 Km
825 €



Coche #3
4 asientos
100 CV
400 Km
1600 €



Coche #4 5 asientos 120 CV 700 Km 1500 €



Coche #5 6 asientos 180 CV 1000 Km 2000 €



Regresión logística







No spam

Hola Juan, nos vemos a las 18:00 en el parque.
Saludos.





Es spam si aparecen "¡!" y la palabra "oferta".



Spam

¡Hola! Aprovecha esta magnífica oferta. ¡¡No la dejes escapar!!



Posibles reglas



Regresión logística

Si aparecen los símbolos "¡!", entonces spam. Si no, no spam.

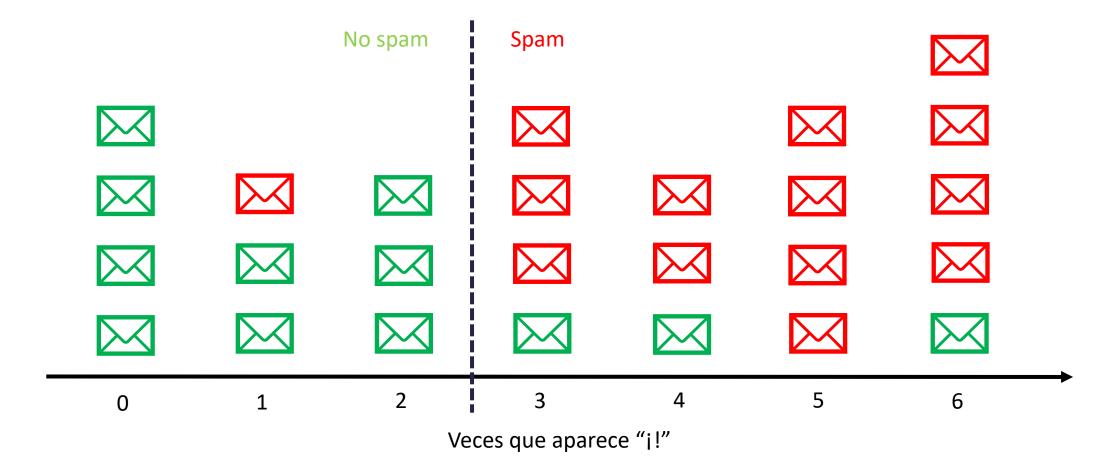
Si aparece la palabra "oferta", entonces spam. Si no, no spam.

Si aparece la palabra "oferta" tres veces y los símbolos "¡!" cuatro veces, entonces spam. Si no, no spam.

Si el número de veces que aparece "oferta" más el número de apariciones de "¡!" es mayor que cuatro, entonces spam. Si no, no spam.

Regresión logística

Si aparecen los símbolos "¡!" tres o más veces, entonces spam. Si no, no spam.





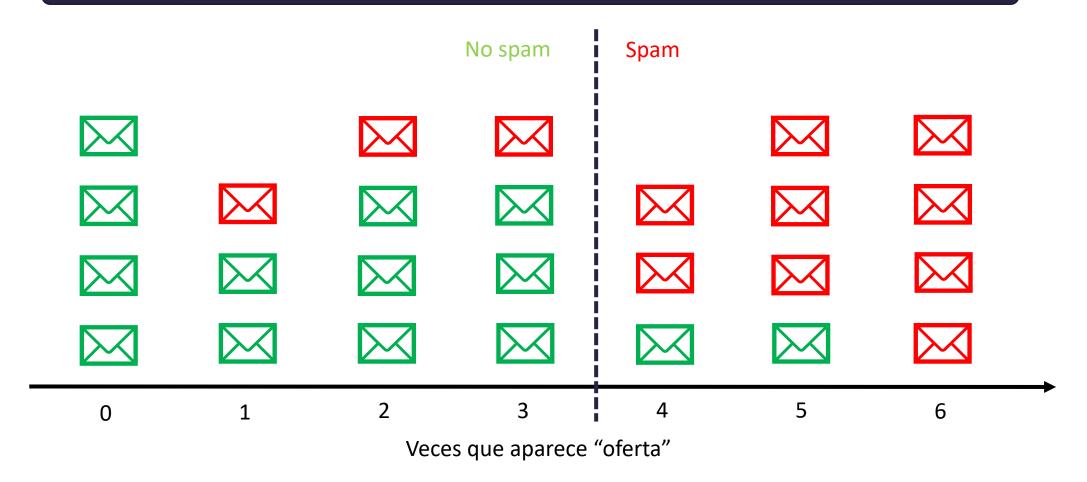


F



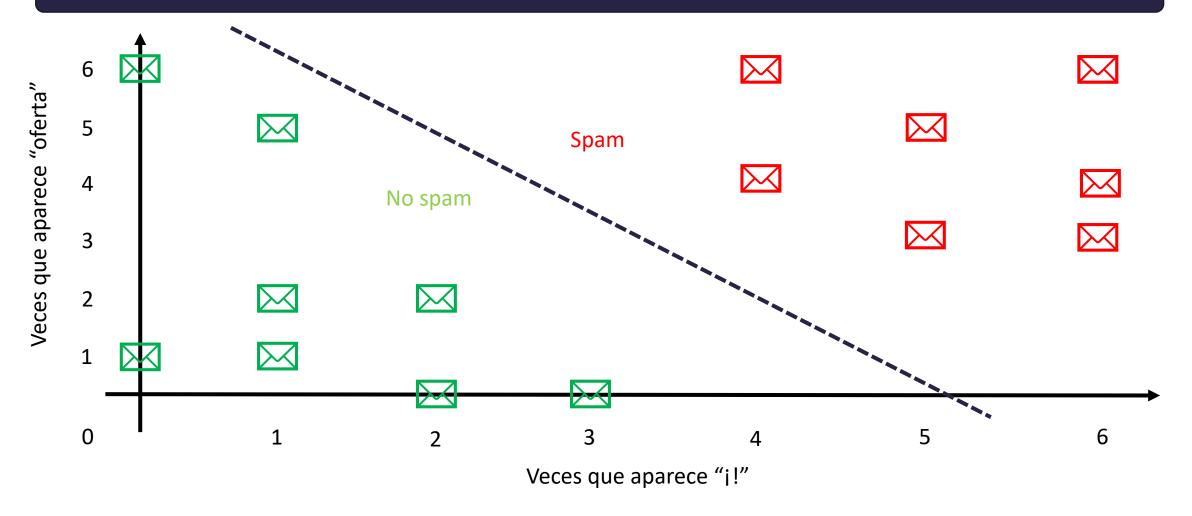
Regresión logística

Si aparece la palabra "oferta" aparece cuatro o más veces, entonces spam. Si no, no spam.



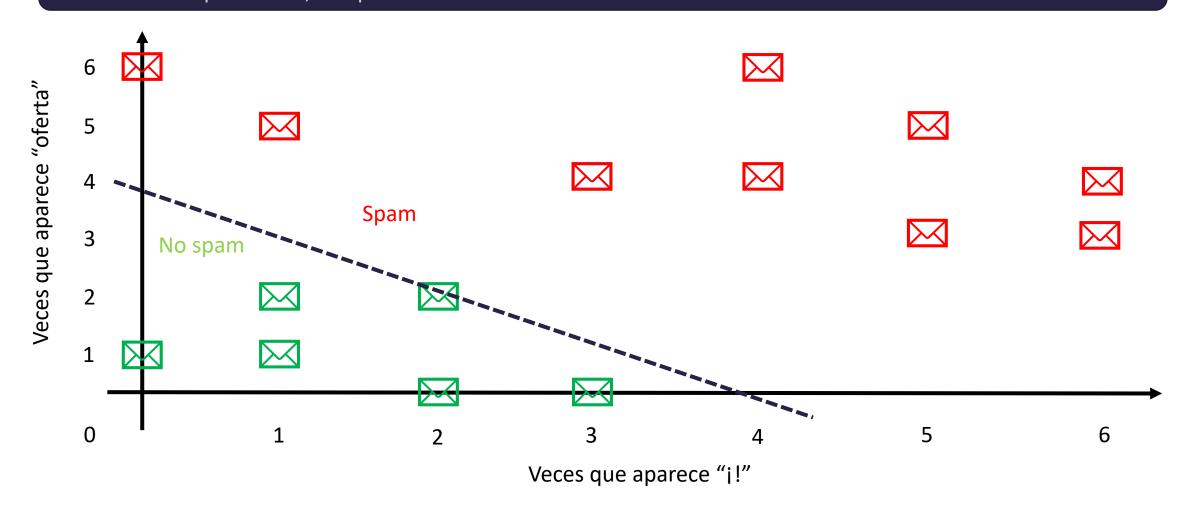
Regresión logística

Si aparece la palabra "oferta" tres veces y los símbolos "¡!" cuatro veces, entonces spam. Si no, no spam.



Regresión logística

Si el número de veces que aparece "oferta" más el número de apariciones de "¡!" es mayor que cuatro, entonces spam. Si no, no spam.



Árboles de decisión

Sexo	Edad	Producto
F	18	
М	38	
М	23	
F	30	
F	44	
M	16	

Mujer de 37 años, ¿qué se le recomienda?









Árboles de decisión

Sexo	Edad	Producto
F	18	
М	38	
M	23	
F	30	
F	44	
M	16	

Entre sexo y edad, ¿Cuál es más descriptivo?

Aprendizaje supervisado

Árboles de decisión

Sexo	Edad	Producto
F	18	
M	38	
M	23	
F	30	Q
F	44	
M	16	

Entre sexo y edad, ¿Cuál es más descriptivo?





Aprendizaje supervisado

Árboles de decisión



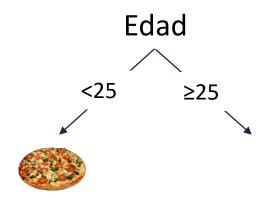


Sexo	Edad	Producto
F	18	
M	38	
M	23	
F	30	W
F	44	
M	16	

Entre sexo y edad, ¿Cuál es más descriptivo?

Edad

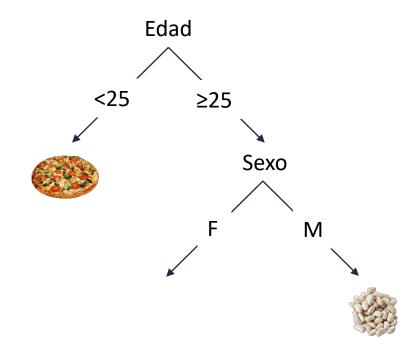
Sexo	Edad	Producto
F	18	
M	38	
M	23	
F	30	Q
F	44	
M	16	







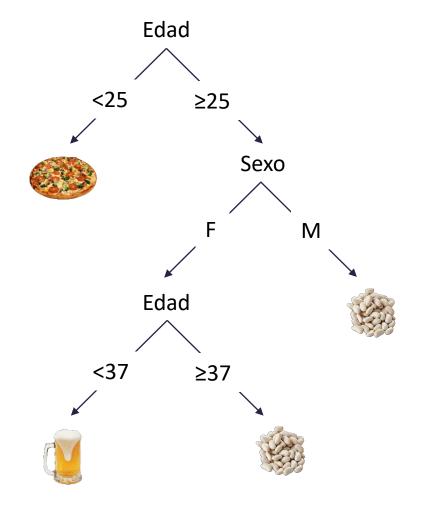
Sexo	Edad	Producto
М	38	
F	30	Q
F	44	







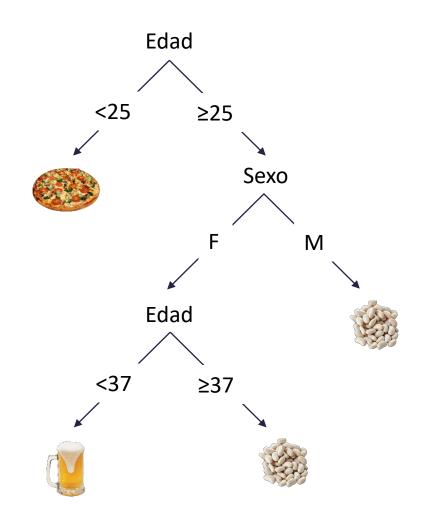
Sexo	Edad	Producto
F	30	
F	44	

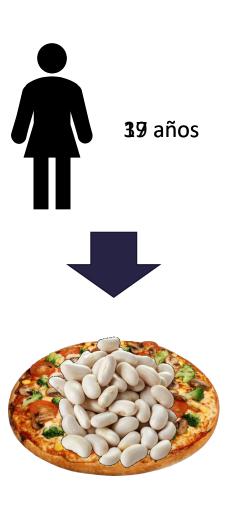












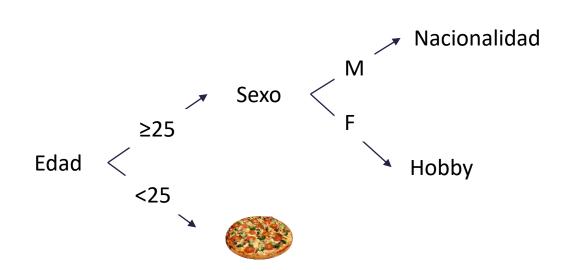


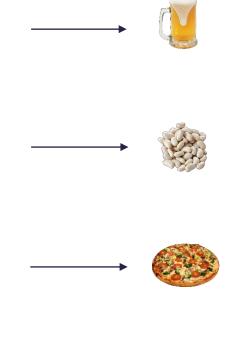


Sexo	Edad	Nacionalidad	Ocupación	Hobby	Producto
F	18	España	Estudiante	Videojuegos	
M	38	Brasil	Jornada completa	Fútbol	
M	23	Francia	Media jornada	Baloncesto	
F	30	Alemania	Jornada completa	Tenis	
F	44	Brasil	Media jornada	Fútbol	
M	16	Panamá	Estudiante	Videojuegos	

旧







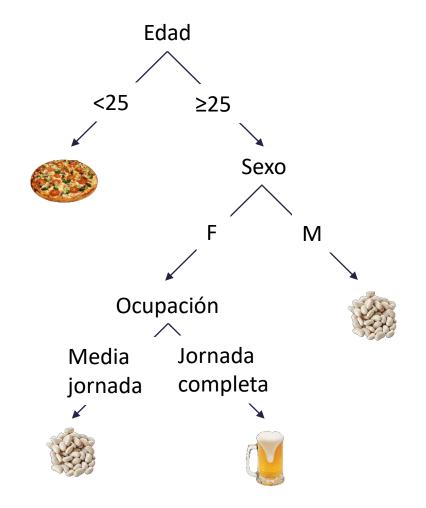


Sexo	Edad	Nacionalidad	Ocupación	Hobby	Producto
F	18	España	Estudiante	Videojuegos	
M	38	Brasil	Jornada completa	Fútbol	
М	23	Francia	Media jornada	Baloncesto	
F	30	Alemania	Jornada completa	Tenis	
F	44	Brasil	Media jornada	Fútbol	
M	16	Panamá	Estudiante	Videojuegos	





Sexo	Edad	Nacionalidad	Ocupación	Hobby	Producto
F	18	España	Estudiante	Videojuegos	
M	38	Brasil	Jornada completa	Fútbol	
M	23	Francia	Media jornada	Baloncesto	
F	30	Alemania	Jornada completa	Tenis	
F	44	Brasil	Media jornada	Fútbol	
M	16	Panamá	Estudiante	Videojuegos	



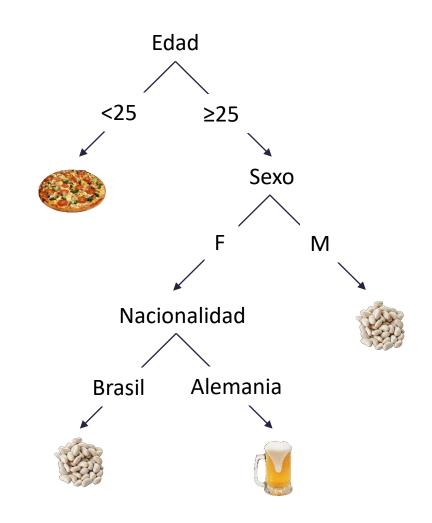




F



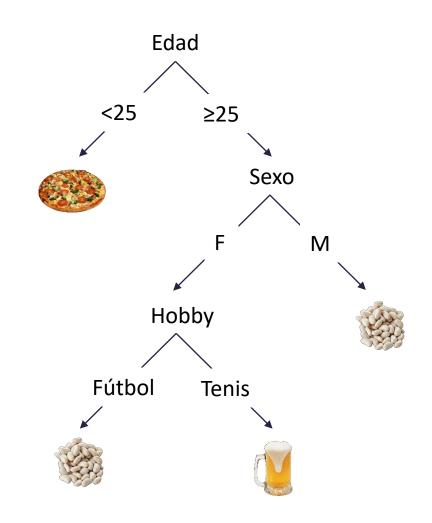
Sexo	Edad	Nacionalidad	Ocupación	Hobby	Producto
F	18	España	Estudiante	Videojuegos	
M	38	Brasil	Jornada completa	Fútbol	
M	23	Francia	Media jornada	Baloncesto	
F	30	Alemania	Jornada completa	Tenis	
F	44	Brasil	Media jornada	Fútbol	
M	16	Panamá	Estudiante	Videojuegos	



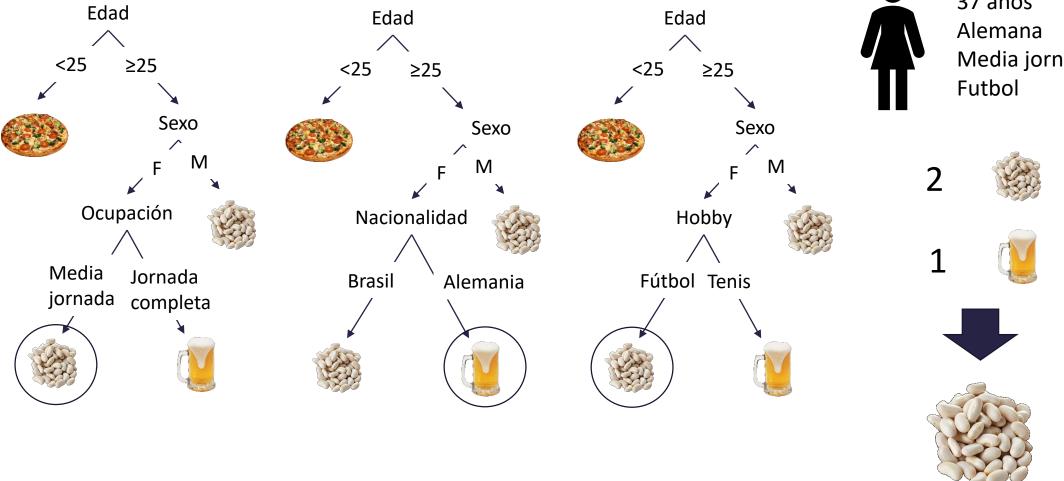
F

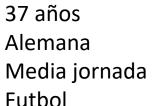


Sexo	Edad	Nacionalidad	Ocupación	Hobby	Producto
F	18	España	Estudiante	Videojuegos	
M	38	Brasil	Jornada completa	Fútbol	
M	23	Francia	Media jornada	Baloncesto	
F	30	Alemania	Jornada completa	Tenis	
F	44	Brasil	Media jornada	Fútbol	
M	16	Panamá	Estudiante	Videojuegos	



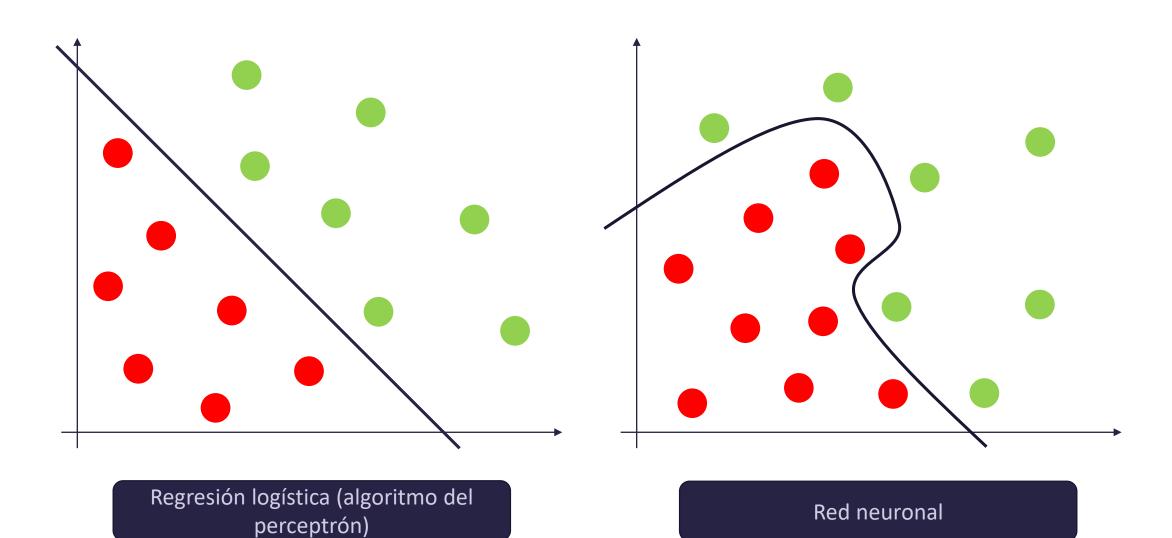






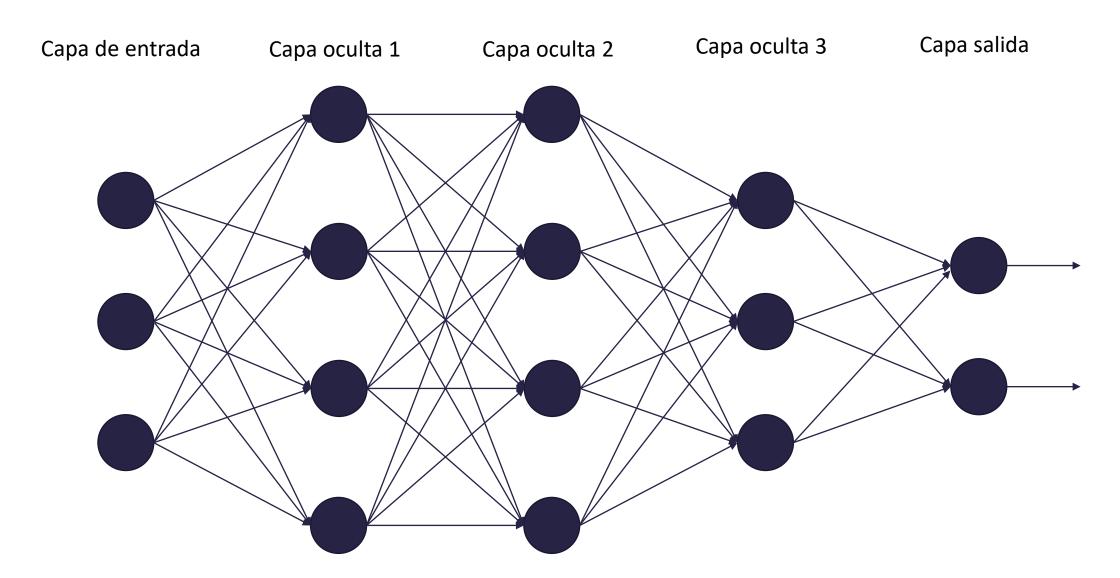


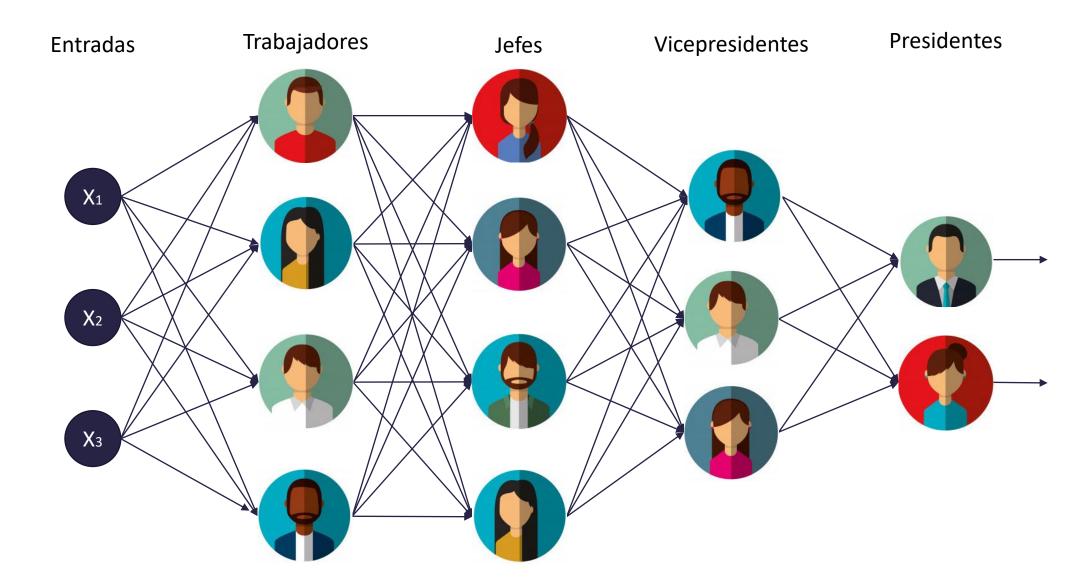




48

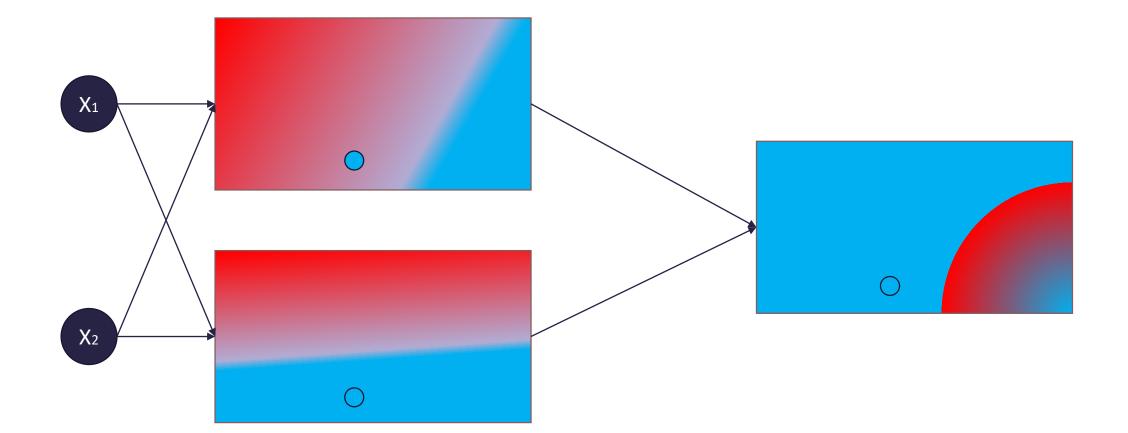






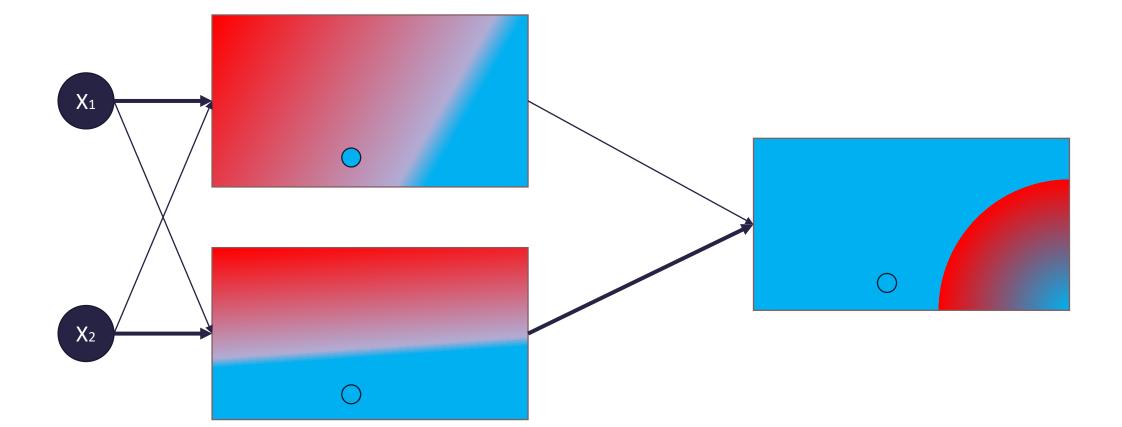






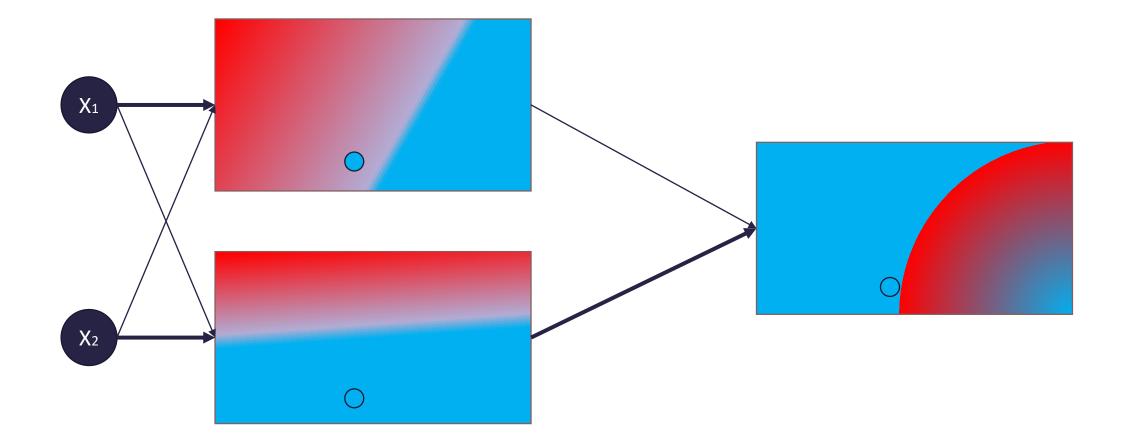






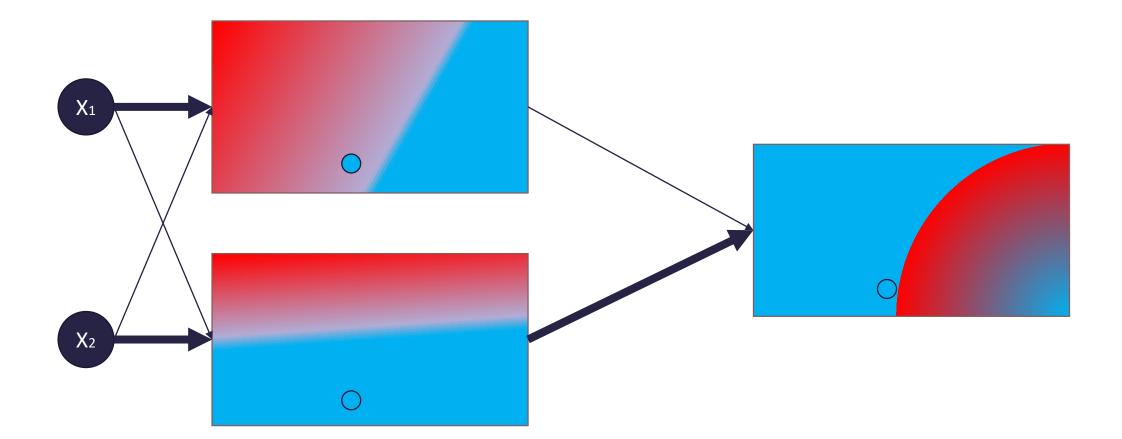








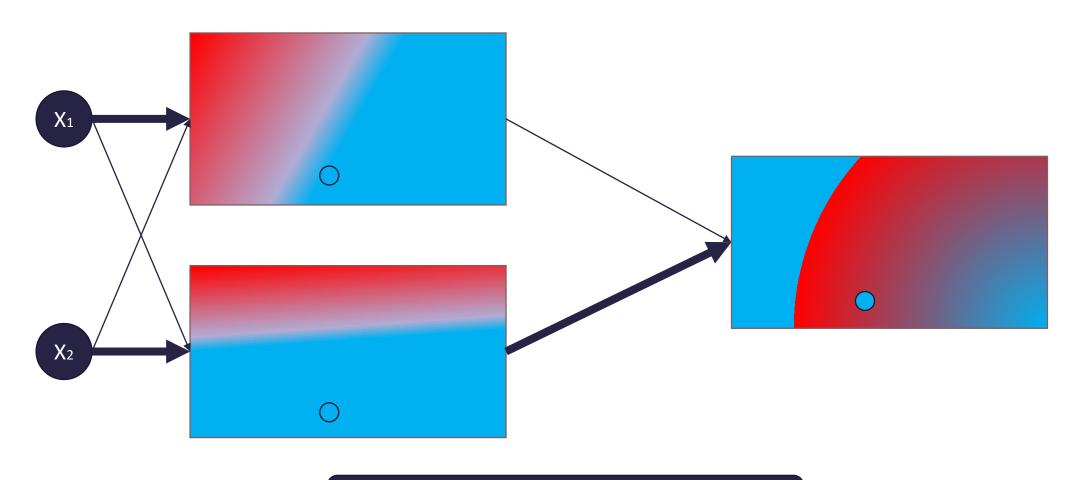








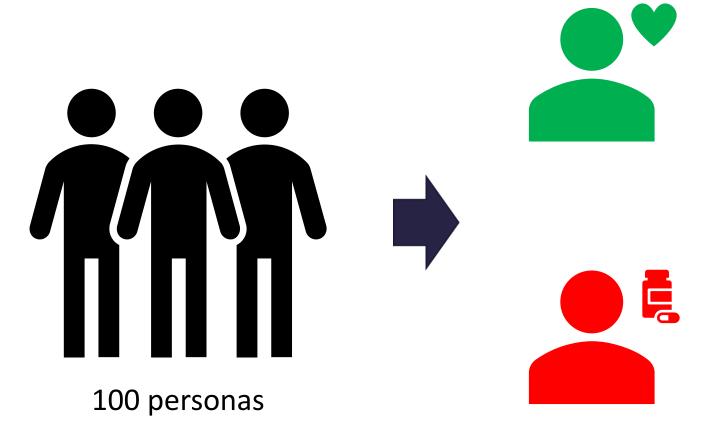




Backpropagation









Fiebre alta Presión arterial alta Dolores de estomago

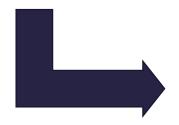
F



Clasificador de Bayes

	Enfermo		Sano	
Total	25		7	' 5
Fiebre alta	20 4/5		6	2/25
Presión arterial alta	10	2/5	20	4/15
Fiebre alta y presión arterial alta	8	8/25	8/5	8/375





Principio de Bayes: todas las variables son independientes entre sí



	Enfermo		Sa	no
Total	25		7	5
Fiebre alta	20 4/5		6	2/25
Presión arterial alta	10 2/5		20	4/15
Fiebre alta y presión arterial alta	8	8/25	8/5	8/375

$$\frac{8}{8+\frac{8}{5}} = \frac{40}{48} = 83,33\%$$

	Enfe	ermo	Sano		
Total	otal 25 75		5		
Fiebre alta	20 4/5 6			2/25	
Presión arterial alta	10 2/5		20	4/15	
Dolor de estomago	10 2/5		10	2/15	
Fiebre alta y presión arterial alta	16/5	16/125	16/75	16/5625	

$$\frac{\frac{16}{5}}{\frac{16}{5} + \frac{16}{75}} = \frac{1200}{1280} = 93,75\%$$









Teorema de Bayes

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

E: Enfermo

S: Sano F: Fiebre

Pr: Presión

$$= \frac{\frac{\frac{4}{5} \cdot \frac{2}{5} \cdot \frac{23}{100}}{\frac{4}{5} \cdot \frac{2}{5} \cdot \frac{25}{100} + \frac{2}{25} \cdot \frac{4}{15} \cdot \frac{75}{100}} = 83,33 \%$$

Aprendizaje no supervisado

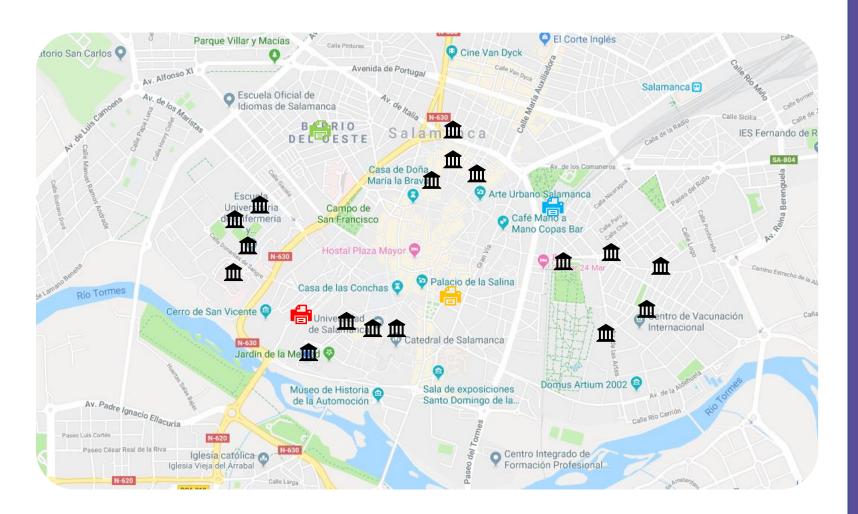








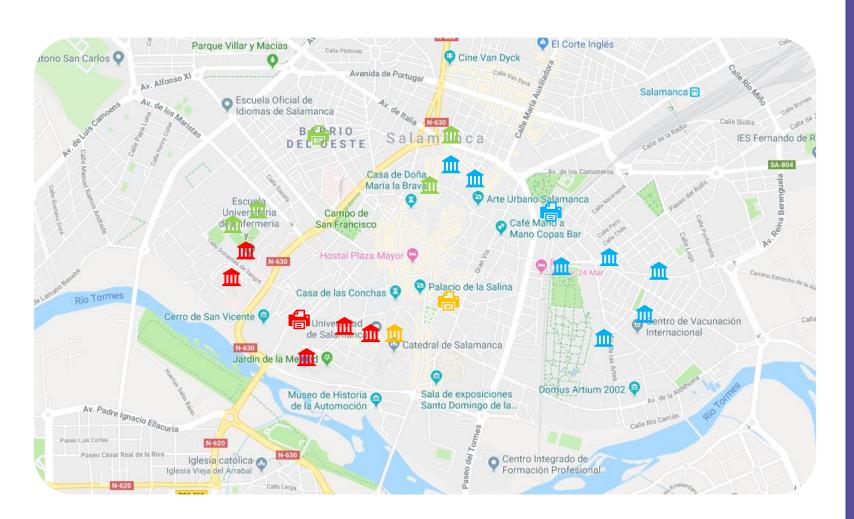
1. Posiciones aleatorias







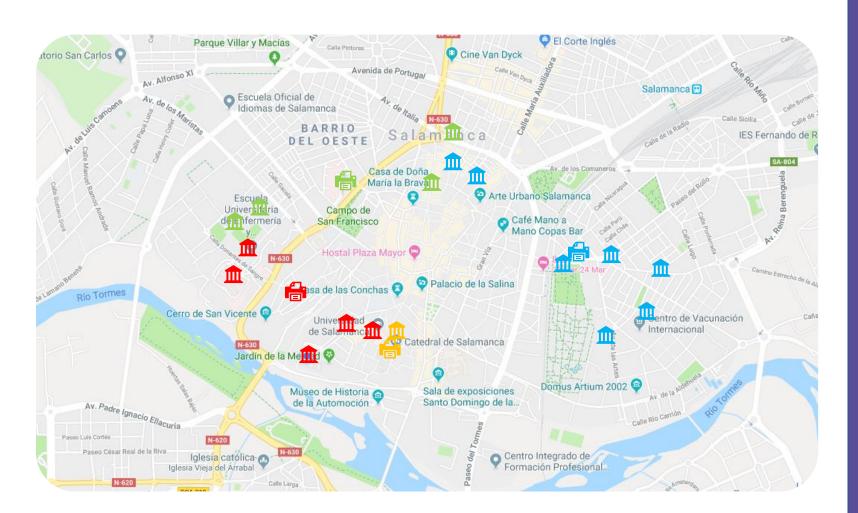
- 1. Posiciones aleatorias
- 2. Bucle
 - Asignación de las instancias a los centros más cercanos







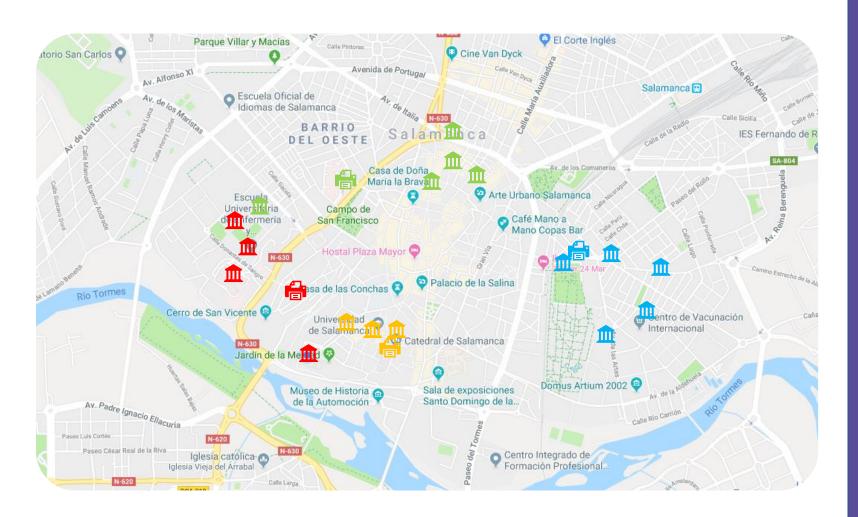
- 1. Posiciones aleatorias
- 2. Bucle
 - 1. Asignación de las instancias a los centros más cercanos
 - Reubicación de los centros







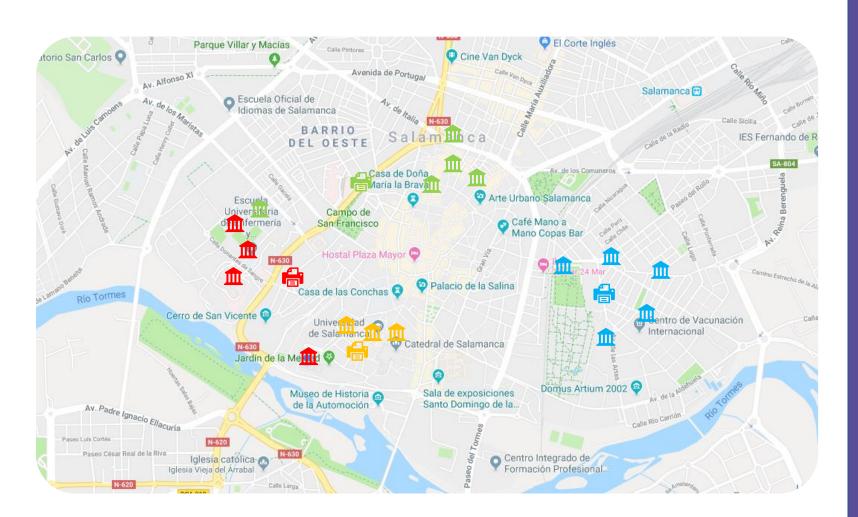
- 1. Posiciones aleatorias
- 2. Bucle
 - 1. Asignación de las instancias a los centros más cercanos
 - 2. Reubicación de los centros







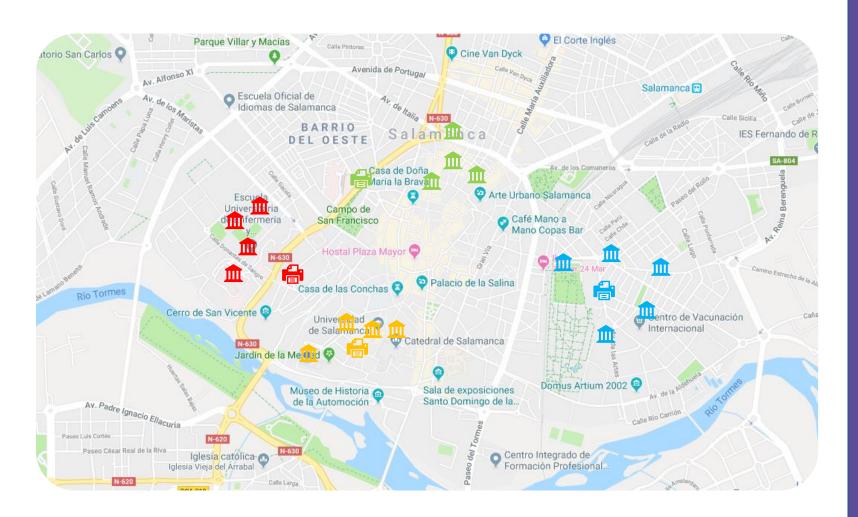
- 1. Posiciones aleatorias
- 2. Bucle
 - 1. Asignación de las instancias a los centros más cercanos
 - Reubicación de los centros







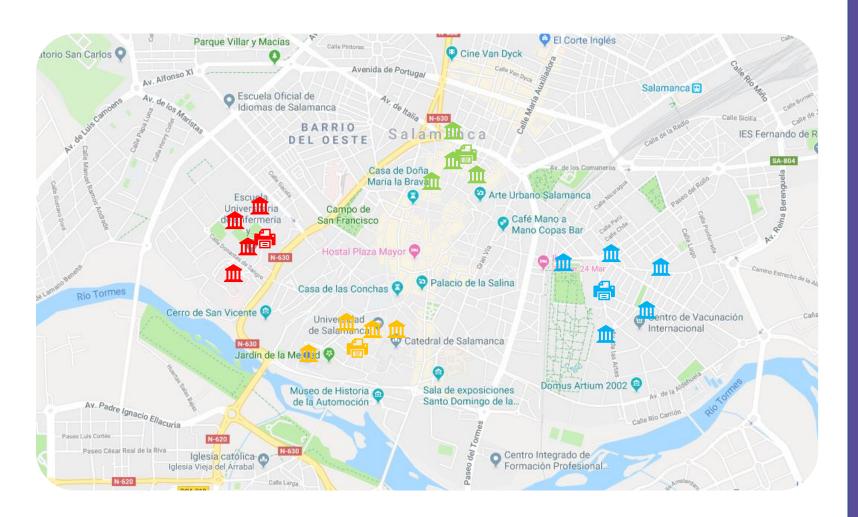
- 1. Posiciones aleatorias
- 2. Bucle
 - 1. Asignación de las instancias a los centros más cercanos
 - 2. Reubicación de los centros







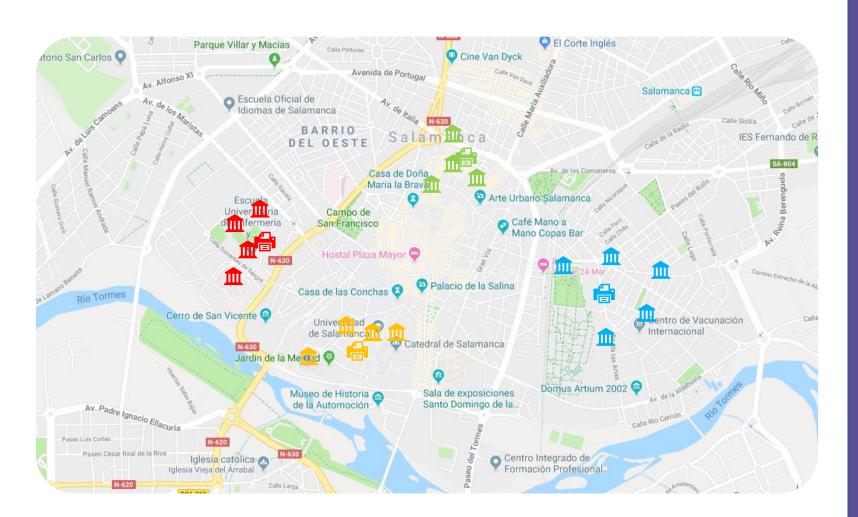
- 1. Posiciones aleatorias
- 2. Bucle
 - Asignación de las instancias a los centros más cercanos
 - Reubicación de los centros







- 1. Posiciones aleatorias
- 2. Bucle
 - 1. Asignación de las instancias a los centros más cercanos
 - 2. Reubicación de los centros







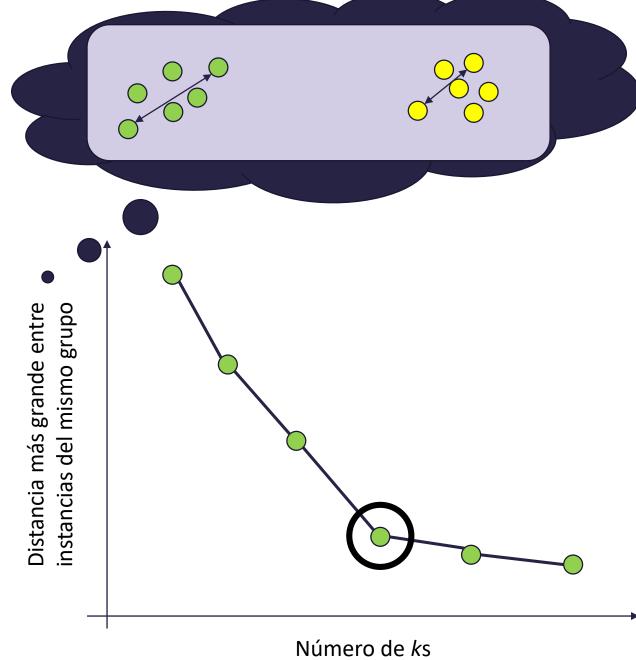


¿Qué valor de *k* es más apropiado?



Método de elbow





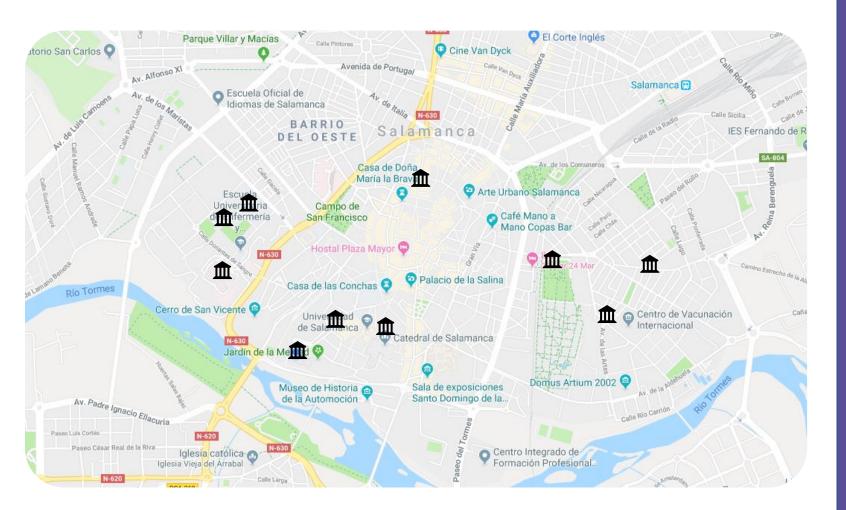
Agrupamiento jerárquico







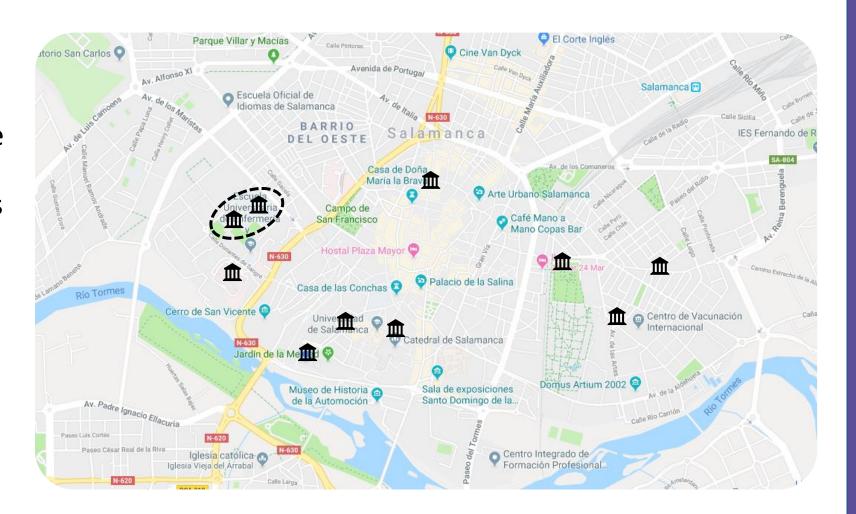
- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo







- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo

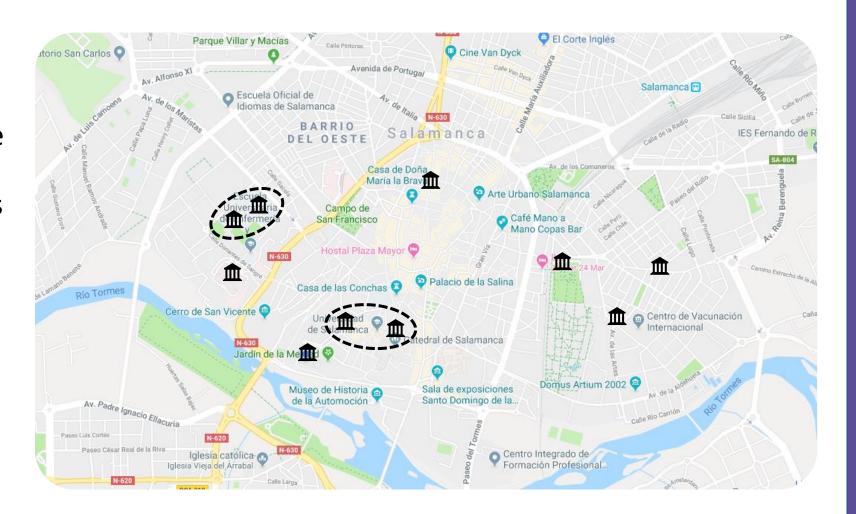




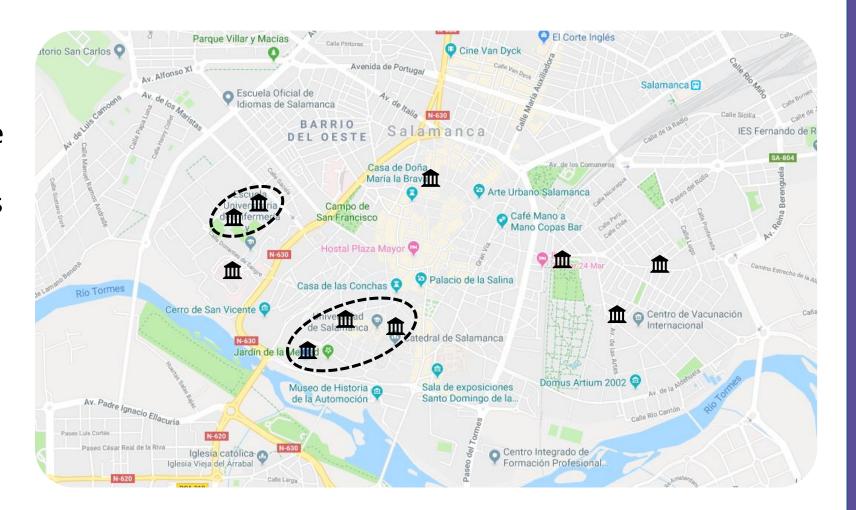




- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo



- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo

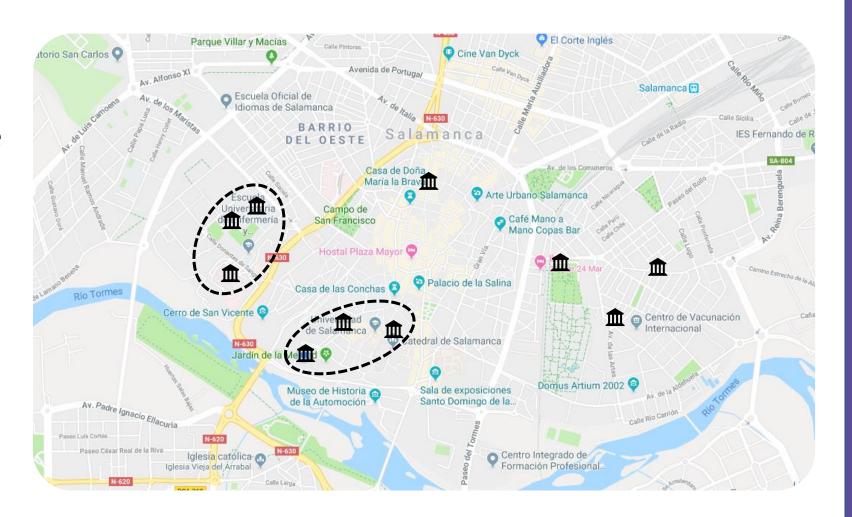




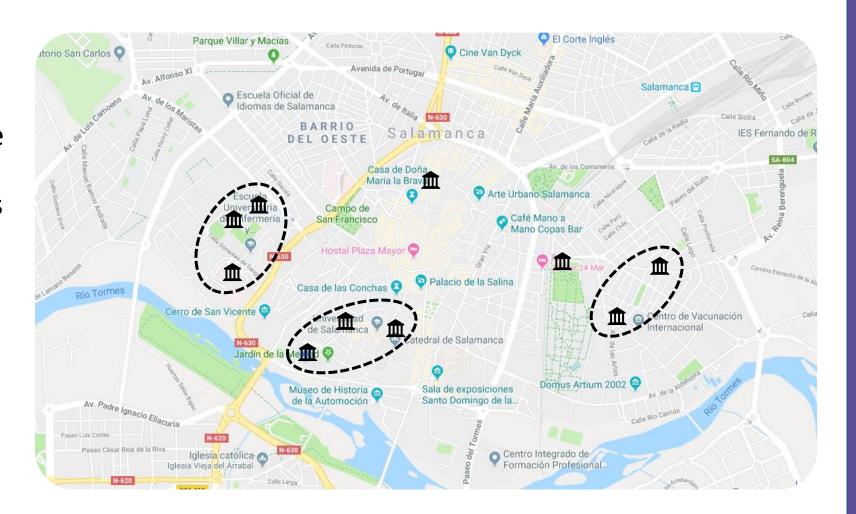




- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo



- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo

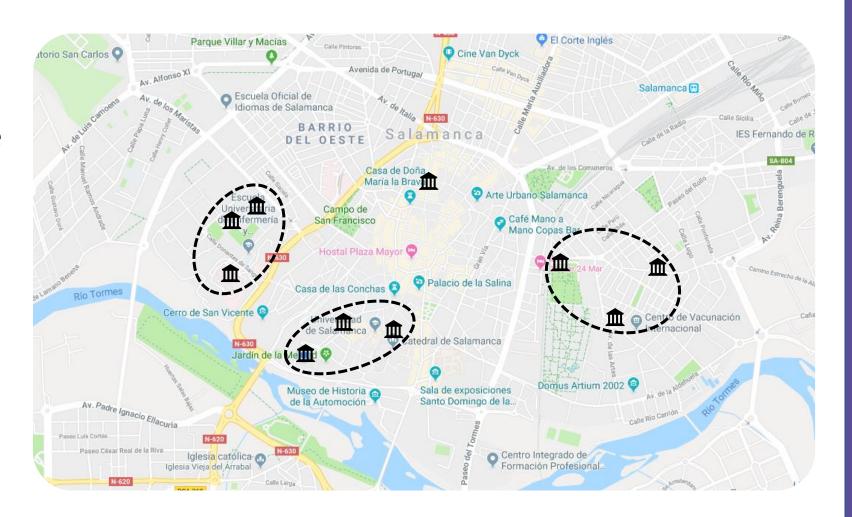




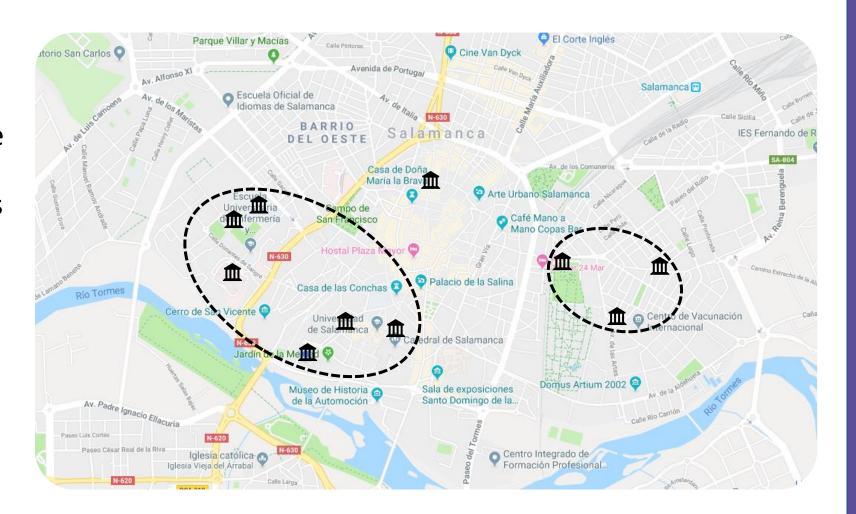




- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo



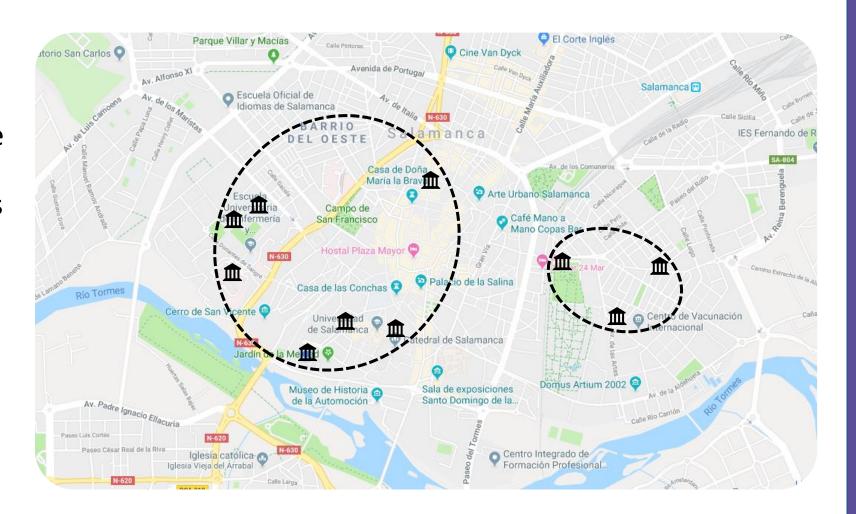
- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo







- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo

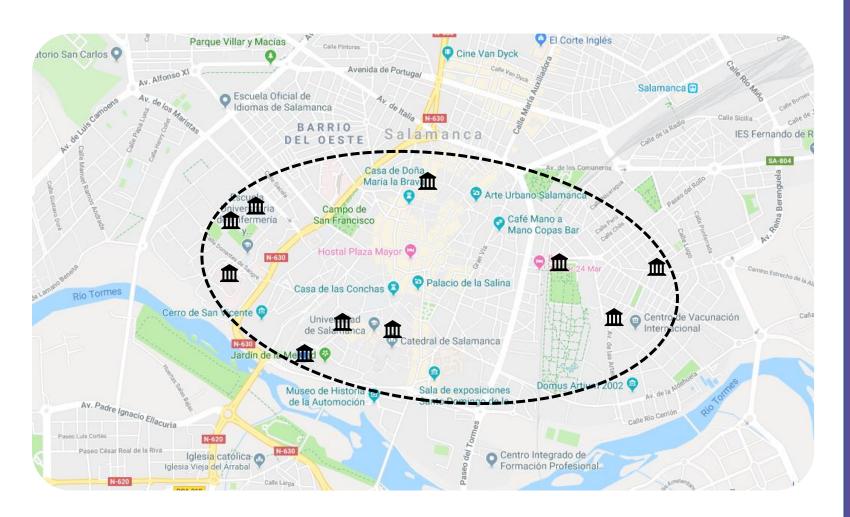


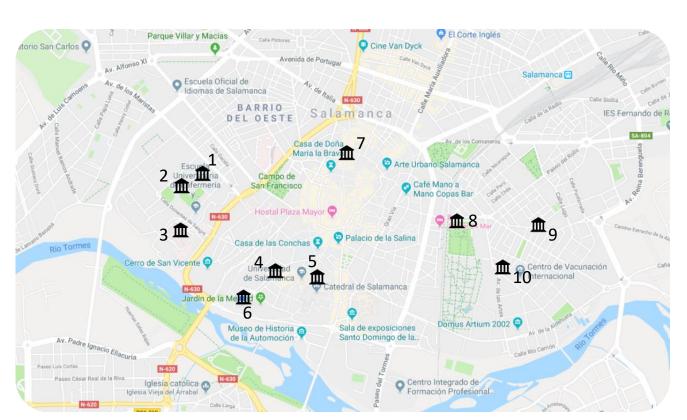


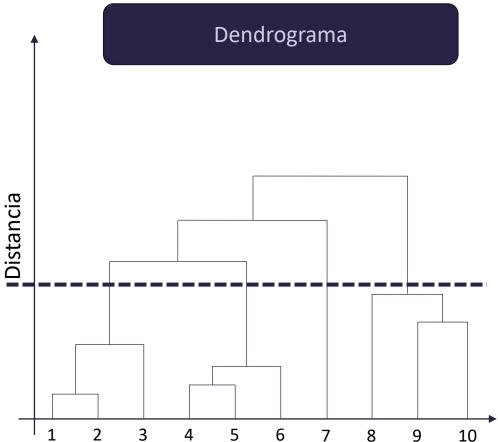




- Calcular la distancia entre los puntos/grupos
- La distancia más pequeña formará un nuevo grupo











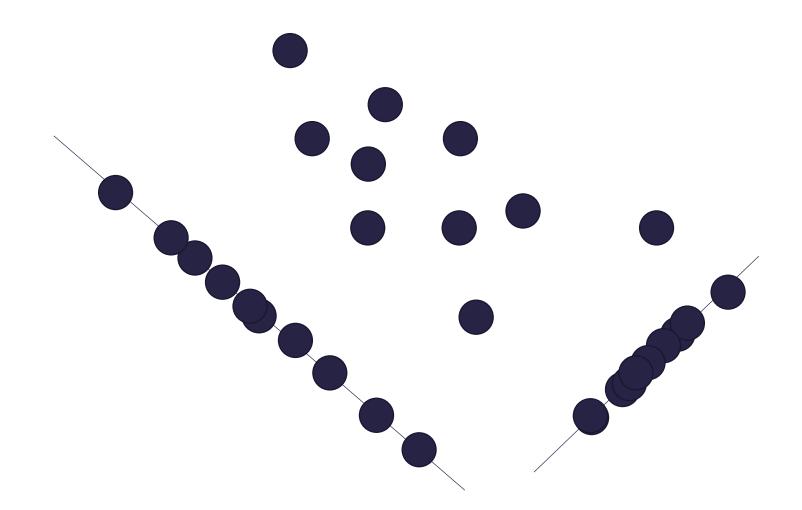
X 1	X 2	Х3	•••	Xn-1	Xn
a ₁	a 2	a 3	•••	a n-1	a n
b ₁	b ₂	b 3	•••	b n-1	bn
•••	•••	•••	•••	•••	•••
•••	•••	•••	•••	•••	•••
•••	•••	•••	•••	•••	•••
•••	•••	•••	•••	•••	•••



Xi	Xj
ai	aj
bi	bj
•••	•••
•••	•••
•••	•••
•••	•••







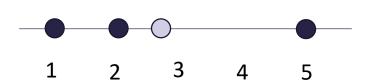






Promedio

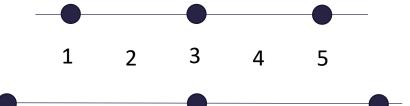
$$\bar{X} = \frac{\sum_{i=0}^{n} x_i}{n}$$



$$\frac{1+2+5}{3} = \frac{8}{3} = 2,666...$$

Varianza

$$var = \frac{\sum_{i=0}^{n} (x_i - \overline{X})^2}{n}$$



$$\frac{3^2 + 0^2 + (-3^2)}{3} = \frac{18}{3} = 6$$

 $\frac{2^2+0^2+(-2^2)}{3} = \frac{8}{3} = 2,666...$

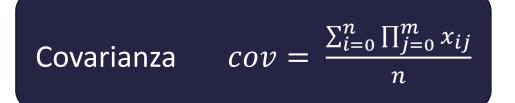
y-varianza x-varianza

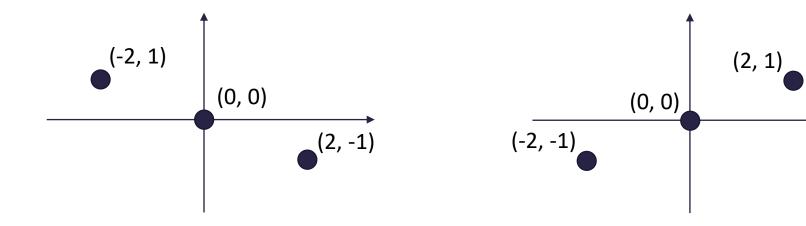




F



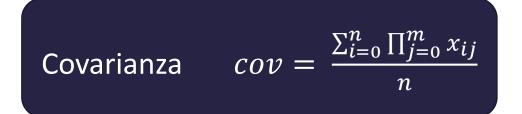


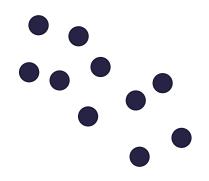


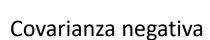
$$cov = \frac{(-2)+0+(-2)}{3} = \frac{-4}{3}$$

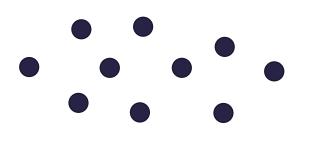
$$cov = \frac{(2)+0+(2)}{3} = \frac{4}{3}$$



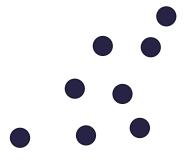






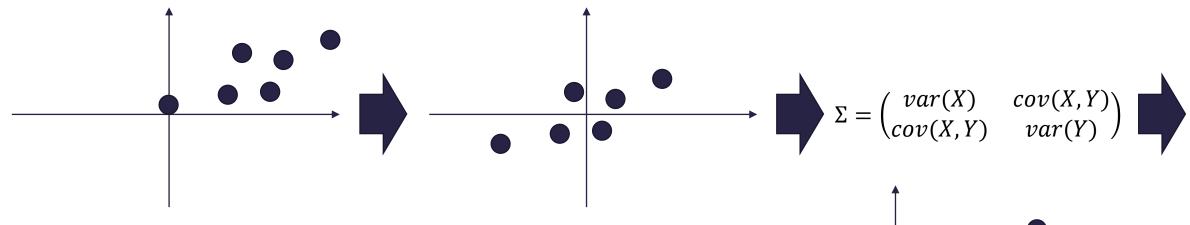


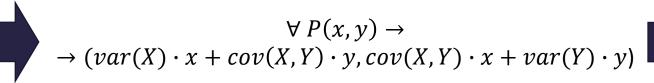
Covarianza cero (o muy cercana a cero)

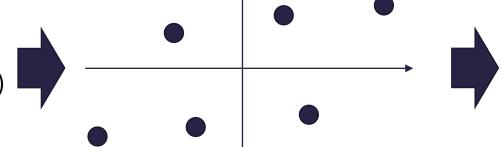


Covarianza positiva







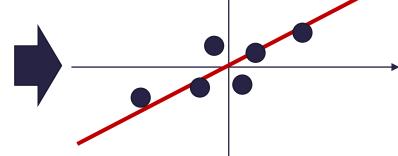




$$\lambda = \begin{vmatrix} x - var(X) & -cov(X,Y) \\ -cov(X,Y) & x - var(Y) \end{vmatrix} \qquad \forall \ \lambda_i \ , i \in [0,n] \rightarrow \Sigma \cdot v = \lambda_i \cdot v$$



$$\forall \; \lambda_i$$
 , $i \; \in [0,n] o \Sigma \cdot v = \lambda_i \; \cdot v$

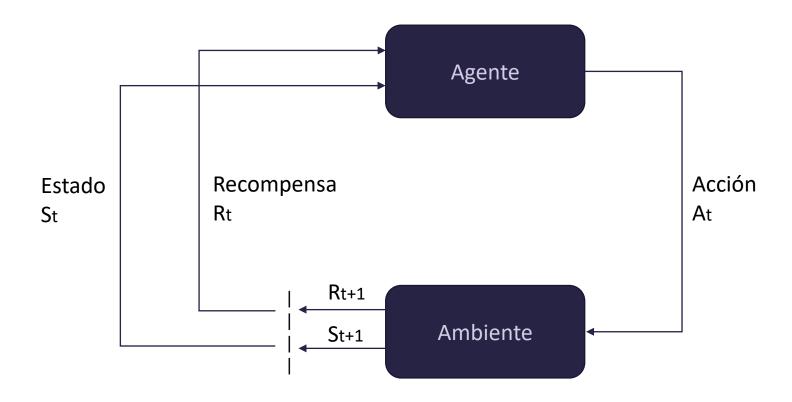


Aprendizaje por refuerzo

Introducción







Q Learning





$$Q(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$$

- Valor Q para un par (estado, acción): $Q(s_t, a_t)$
- Recompensa para el par (estado, acción): $r(s_t, a_t)$
- Refuerzo acumulado para los siguientes estados: $\gamma \in [0, 1]$

$$Q'(s_t, a_t) = (1 - v)Q(s_t, a_t) + v[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})]$$

• Factor de aprendizaje: *v*

上



Q Learning

Exploración vs. Explotación

Exploración aleatoria

Exploración mediante probabilidad

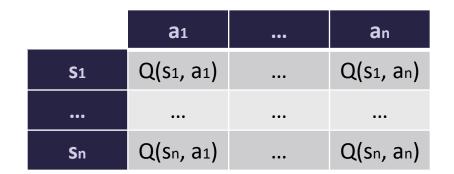
- En función del valor de Q
- En función del tiempo

$$P(a_t) = \frac{e^{E \cdot t \cdot Q(s_t, a_t)}}{\sum_{a \in A_{s_t}} e^{E \cdot t \cdot Q(s_t, a)}}$$

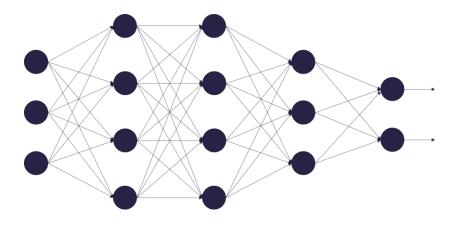
Q Learning

VS.

Continuos



No continuos







Ejemplos en código

Ejemplos en código

- Regresión lineal
- Árboles de decisión
- Redes neuronales
- K-means



diegomjb94/MachineLearningAlgorithms

Referencias

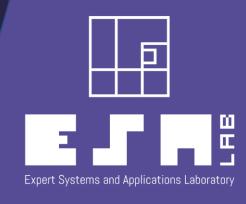




Referencias

- The Elements of Statistical Learning. Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, 2001.
- Machine Learning for Hackers. Drew Conway, John Myles White, 2012.
- Introduction to Machine Learning with Python: A Guide for Data Scientists. Andreas C. Müller, Sarah Guido, 2016.
- Artificial Intelligence with Python. Prateek Joshi, 2017,
- Hands-On Machine Learning with Scikit-Learn and TensorFlow. Aurélien Géron, 2017.
- Deep Learning with TensorFlow. Md. Rezaul Karim, Ahmed Menshawy, Giancarlo Zaccone, 2017.
- Hands-On Reinforcement Learning with Python. Sudharsan Ravichandiran, 2018.
- Reinforcement Learning with TensorFlow. Sayon Dutta, 2018.

Machine Learning / Aprendizaje Automático



Diego M. Jiménez Bravo dmjimenez@usal.es 16/04/2019

