Aprendizaje automático

... con apoyo de herramientas de scripting visual

Francisco Jurado

Universidad Autónoma de Madrid

Contenidos

UAM

Contenidos

UAM

Orange Data Mining

Orange

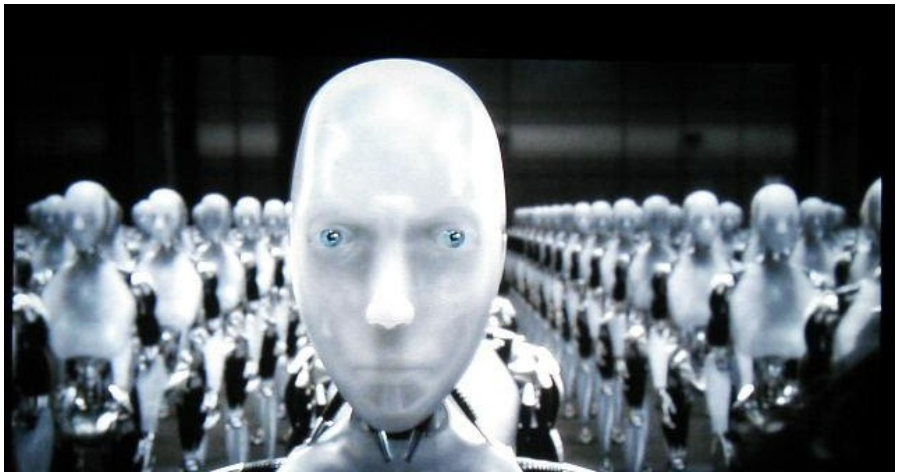
- Herramienta visual para aprendizaje automático y visualización de datos.
- Permite definir flujos de datos de forma visual o mediante programación.

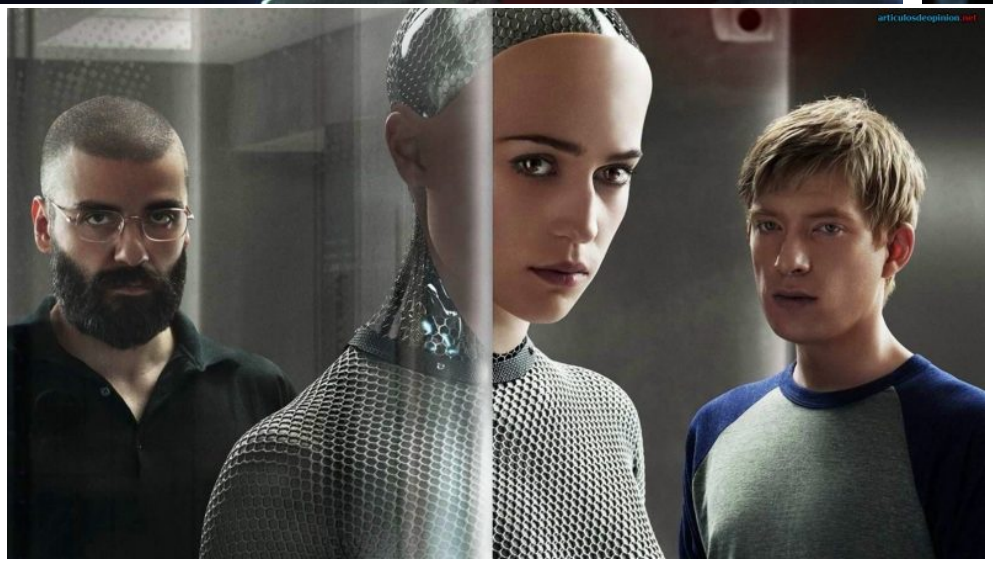
https://orange.biolab.si/

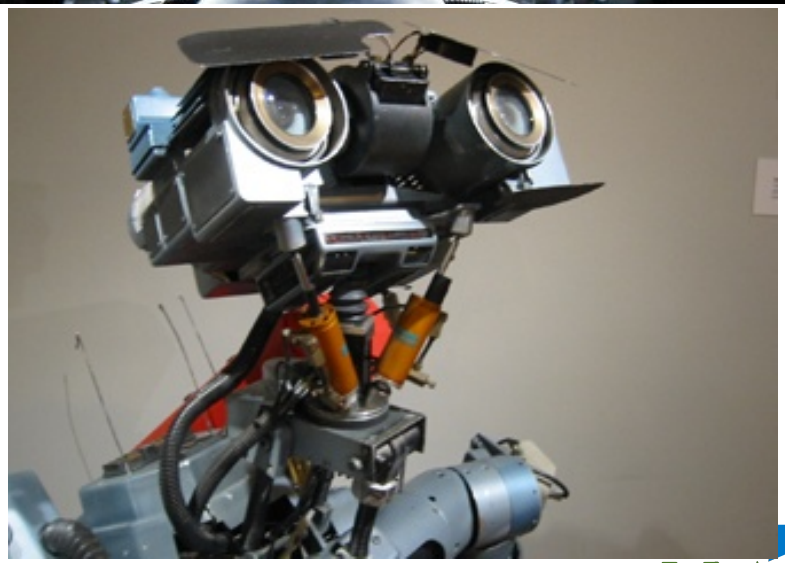
Contenidos

UAM

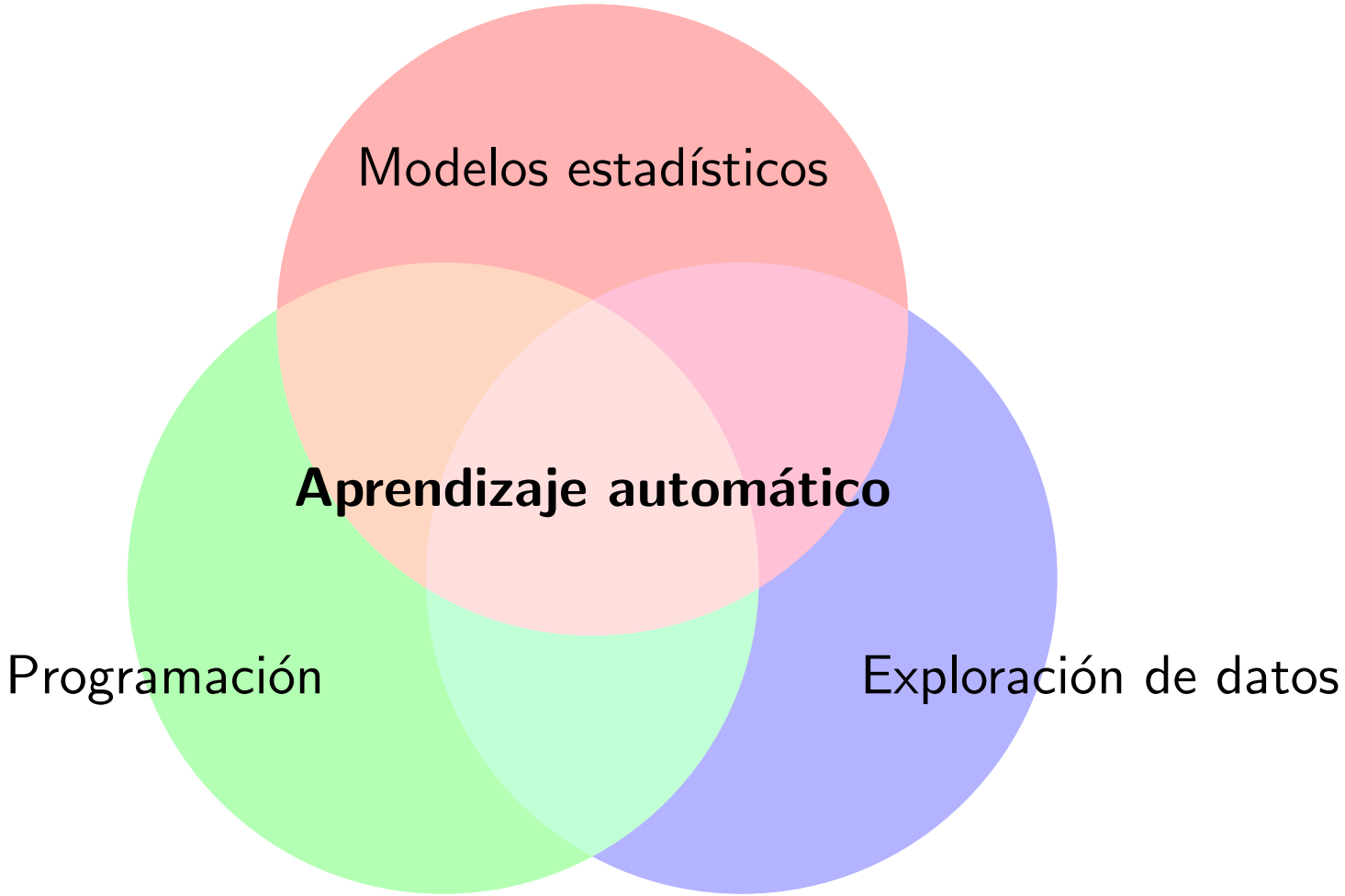
Qué no es Aprendizaje Automático







Aprendizaje automático: dónde situarlo



Algunas aplicaciones

De forma general

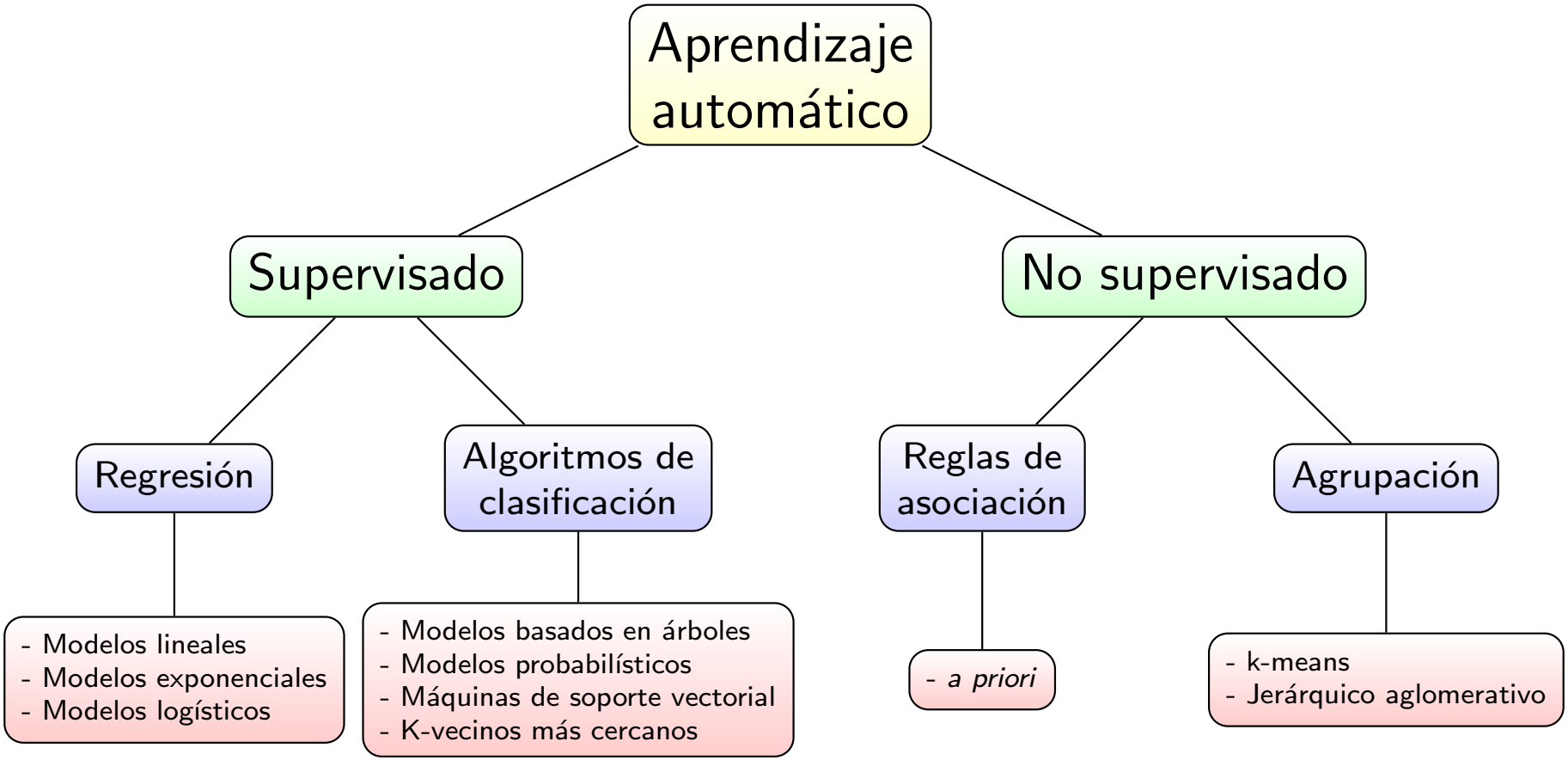
- Análisis de mercado
- Sistemas de recomendación
  - Amazon, Netflix, etc.
- Reconocimiento de patrones
  - Reconocimiento facial y huellas en móviles
- Procesamiento de lenguaje natural
  - Análisis de sentimiento/polaridad
  - Modelado de topics
- Recuperación de información

Ccia. forense y ciberseguridad

- Reconocimiento de patrones
  - Análisis de perfiles de usuario
  - Análisis de malware
  - Identificación de ataques
  - Reconocimiento y análisis de imágenes
- Procesamiento de lenguaje natural
  - Discurso de odio
  - Cyberstalking
  - Cyberbulling
  - Doxing

UAM

Categorización de algunos de los algoritmos más conocidos



Aprendizaje supervisado vs. no supervisado

Aprendizaje supervisado

- Datos etiquetados $\rightarrow$ **clases**.
- Tratan de construir un modelo que prediga datos futuros no etiquetados.
- Análisis predictivo.
- Algoritmos de **clasificación y regresión**

Aprendizaje no supervisado

- Datos NO etiquetados.
- Tratan de identificar grupos atendiendo a las características de los datos.
- Análisis descriptivo.
- Algoritmos de **agrupación (clustering)**

Contenidos

Regresión

Tipos de datos

Trabaja con datos numéricos tanto en la entrada como en la salida.

Regresión

Tipos de datos

Trabaja con datos numéricos tanto en la entrada como en la salida.

Objetivo

Modelar una variable continua Y (respuesta) como función matemática de una variable X (predictora), de manera que podamos usar este modelo de regresión para predecir la Y cuando sólo se conoce la X.

Regresión

Tipos de datos

Trabaja con datos numéricos tanto en la entrada como en la salida.

Objetivo

Modelar una variable continua Y (respuesta) como función matemática de una variable X (predictora), de manera que podamos usar este modelo de regresión para predecir la Y cuando sólo se conoce la X.

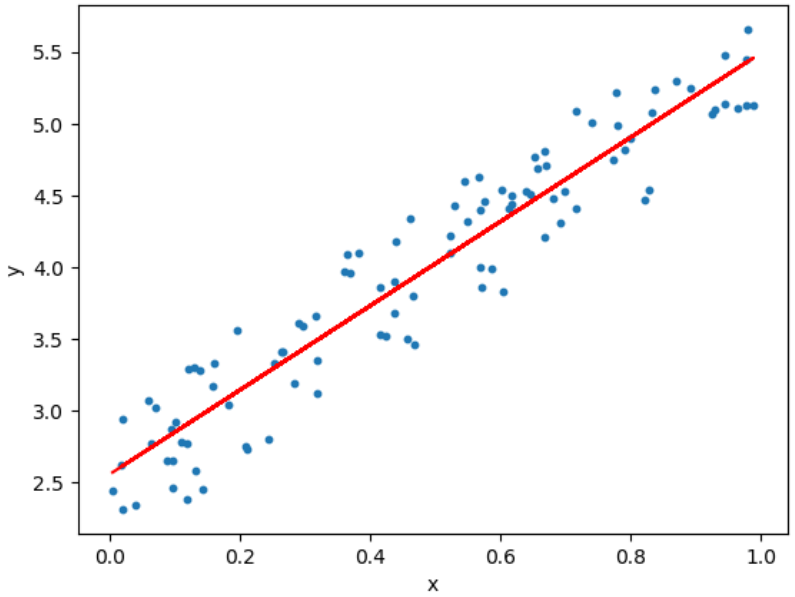
El reto

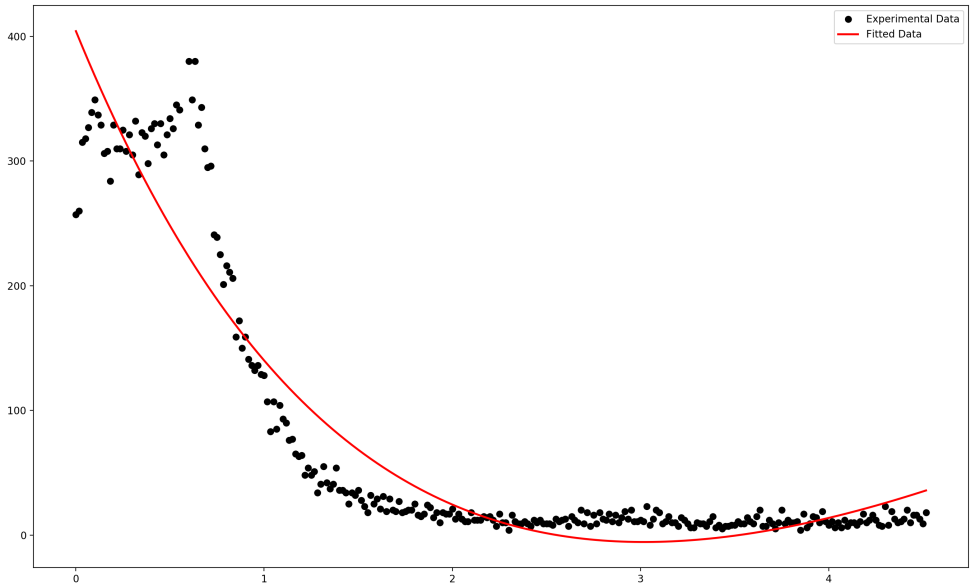
Conseguir la ecuación matemática (función de ajuste):

$$Y = f(X)$$

donde $f(X)$ puede ser una función lineal, cuadrática, logística, etc.

Regresión


5.5
5.0
4.5
4.0
3.5
3.0
2.5
y
0.0
0.2
0.4
0.6
0.8
1.0
x


Experimental Data
Fitted Data
400
300
200
100
0
0
1
2
3
4

Regresión lineal: ejemplo

- Usemos el dataset ‘cars‘ con los datos sobre la velocidad de los coches (en ‘mph‘) y las distancias empleada para detenerse (en pies ‘ft‘) registrados en la década de 1920.
- Tratemos de construir un modelo que permite predecir la Distancia (dist) estableciendo una relación lineal estadísticamente significativa con la Velocidad (velocidad).
- Haremos los cálculos con Orange

Regresión lineal: ejemplo

Tras los cálculos con Orange.

Coeficientes de regresión

- Intersección (α): -17.579
- Coeficiente para Velocidad (β): 3.932

Con lo que la función de ajuste queda:

$$dist = \alpha + \beta.speed = -17.579 + 3.932.speed$$

Coeficiente de determinación*

Proporción de observaciones que es capaz de explicar el modelo.

$$R^2 = 0.8$$

* En los modelos de regresión lineal simple el valor de R^2 se corresponde con el cuadrado del coeficiente de correlación de Pearson (r) entre X e Y.

Modelos de clasificación

Tipos de datos

Trabajan con datos "etiquetados" para buscar una función que reciba las variables de entrada y devuelva la etiqueta apropiada en su salida.

Modelos de clasificación

Tipos de datos

Trabajan con datos "etiquetados" para buscar una función que reciba las variables de entrada y devuelva la etiqueta apropiada en su salida.

Cómo se entrena

Con un "histórico" de datos etiquetados con los que "aprende" a asignar la etiqueta de salida adecuada a nuevos valores, es decir,
predice la etiqueta (clase) de salida.

Modelos basados en árboles

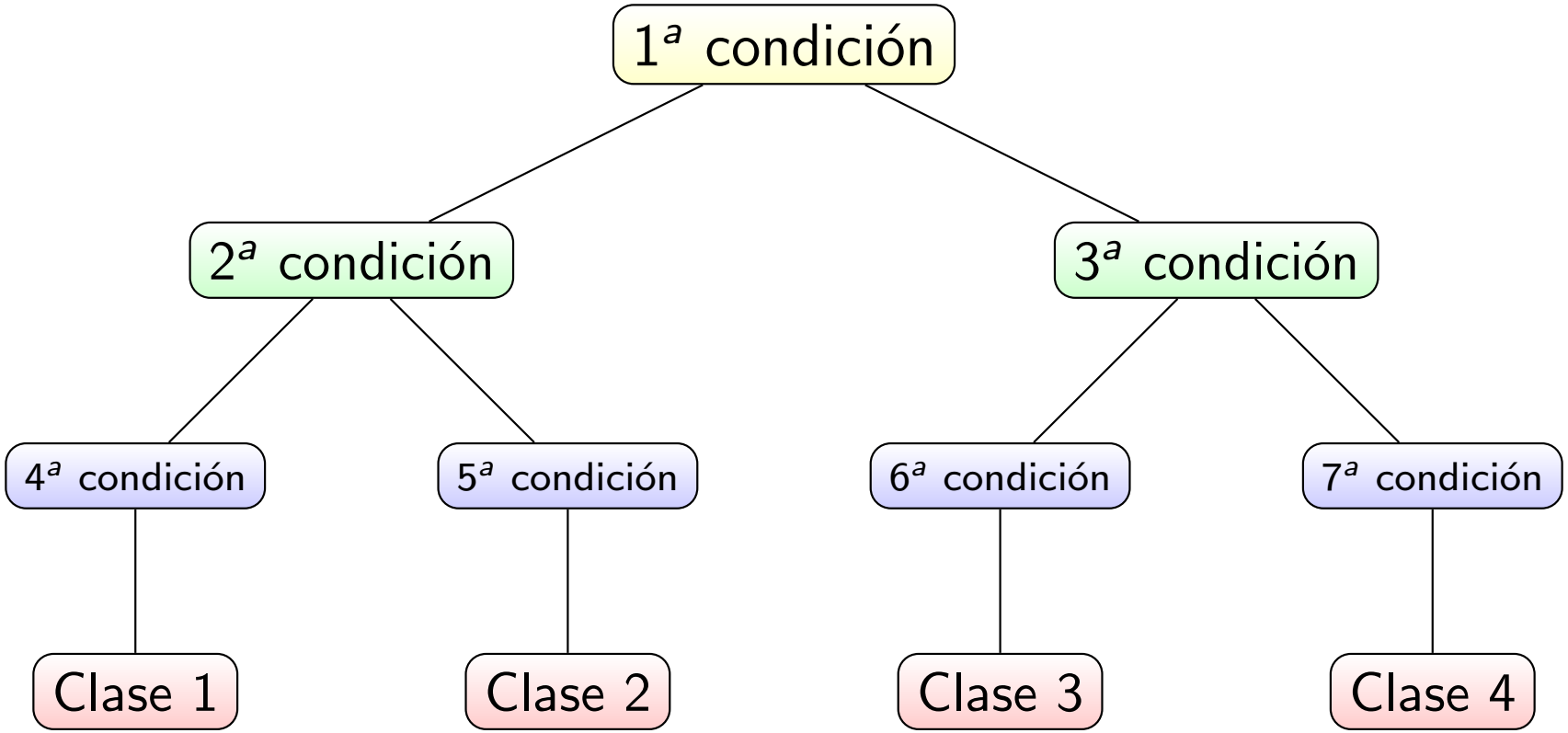
En este curso se verán árboles de inferencia condicional conocidos como:

CART

Classification and regression trees

- La partición recursiva ayuda a explorar la estructura de un conjunto de datos mediante la obtención de reglas de decisión fáciles de visualizar.
- Diferenciaremos:
  - Árboles de classification para datos categoricos
  - Árboles de regresión para datos continuos

CART


1ª condición
2ª condición
3ª condición
4ª condición
5ª condición
6ª condición
7ª condición
Clase 1
Clase 2
Clase 3
Clase 4

Modelo probabilístico Naïve Bayes

Teorema de Bayes

Probabilidad *a-posteriori* o probabilidad condicional $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$

Modelo probabilístico Naïve Bayes

Teorema de Bayes

Probabilidad *a-posteriori* o probabilidad condicional $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$

Naïve Bayes o Clasificador bayesiano ingenuo

- Clasificador probabilístico fundamentado en el teorema de Bayes
- Calcula la probabilidad p de la clase C dado un conjunto de características (del inglés *features*) $F_1, ...F_n$

$$p(C|F_1, ...F_n) = \frac{p(F_1, ...F_n|C)p(c)}{p(F_1, ...F_n)}$$

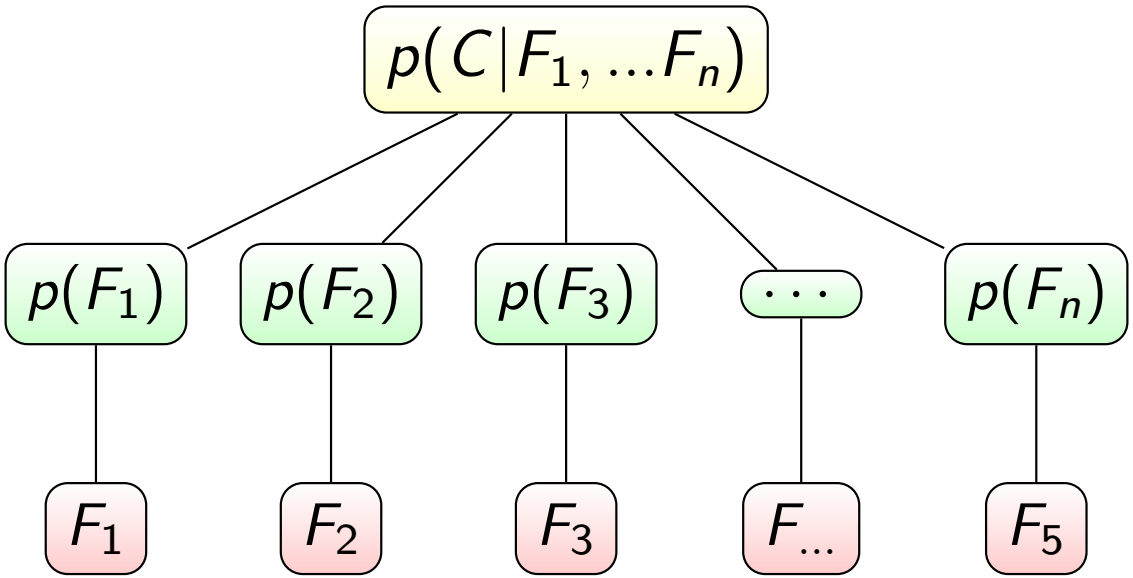
  Siendo:

  - $p(F_1, ...F_n|C)$ la presunción,
  - $p(c)$ la probabilidad total de la clase, y
  - $p(F_1, ...F_n)$ la evidencia.
- El modelo calcula la probabilidad condicional para cada característica por separado, así como las probabilidades *a-priori* que indican la distribución de los datos.

Modelo probabilístico Naïve Bayes

Intuitivamente

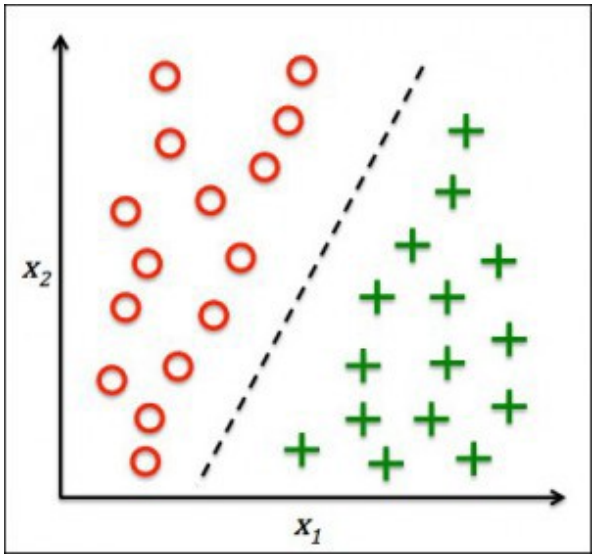
Identificadas una serie de evidencias (las características $F_1, F_2, \ldots, F_n$), calcula la probabilidad p de que se trate de la clase C.



Máquinas de soporte vectorial (SVM)

Una SVM trata de construir un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta.

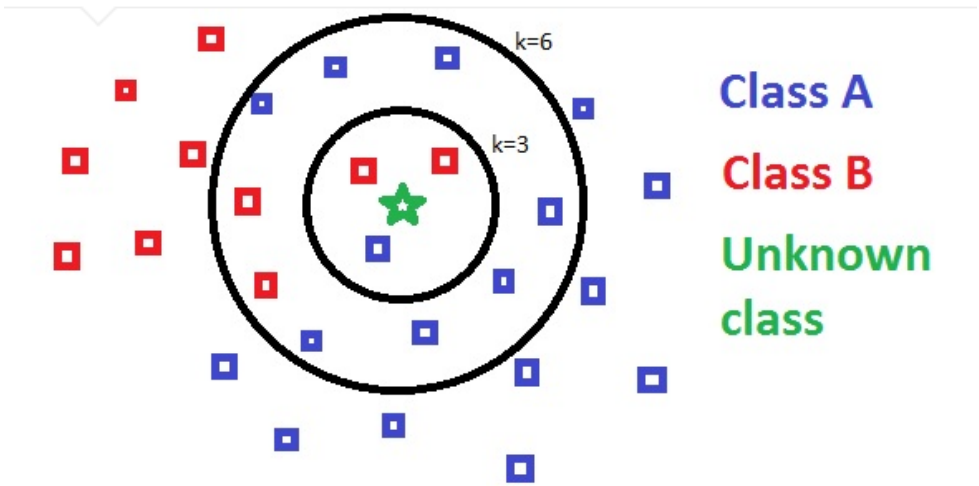
- Los hiperplanos dividen el espacio separando las clases.
- Una buena separación entre las clases permitirá una buena clasificación.
- Cuando se desea clasificar una nueva muestra:
  - se situa la muestra en el espacio vectorial definido en el modelo, y
  - en función del espacio al que pertenezca, quedará clasificada en la clase correspondiente al espacio.



K vecinos más cercanos (*k-nearest neighbors* o k-nn)

Clasificación

- Fase de entrenamiento: Cada elemento del conjunto de entrenamiento es un punto en el espacio que tiene asignado una clase concreta.
- Fase de clasificación: Un nuevo elemento (punto) en el espacio tiene asignada la clase C más frecuente entre los k ejemplos de entrenamiento más cercanos (normalmente usa distancia euclidiana).



Matriz de confusión

		Predicción	
		Positivo	Negativo
Actual	Positivo	TP	FN
	Negativo	FP	TN

Matriz de confusión

		Predicción	
		Positivo	Negativo
Actual	Positivo	TP	FN
	Negativo	FP	TN

A maximizar

Verdaderos positivos (True Posit., TP) Valores correctamente clasificados como positivos.

Verdaderos negativos (True Negat., TN) Valores correctamente clasificados como negativos.

Matriz de confusión

		Predicción	
		Positivo	Negativo
Actual	Positivo	TP	FN
	Negativo	FP	TN

A maximizar

Verdaderos positivos (True Posit., TP) Valores correctamente clasificados como positivos.

Verdaderos negativos (True Negat., TN) Valores correctamente clasificados como negativos.

A minimizar

Falsos positivos (False Posit., FP) Valores mal clasificados como positivos siendo negativos.

Falsos negativos (False Negat., FN) Valores mal clasificados como negativos siendo positivos.

Medidas de precisión

Precisión Proporción de identificaciones positivas correctas

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

Exhaustividad Proporción de positivos reales

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

Exactitud Proporción de predicciones correctas

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

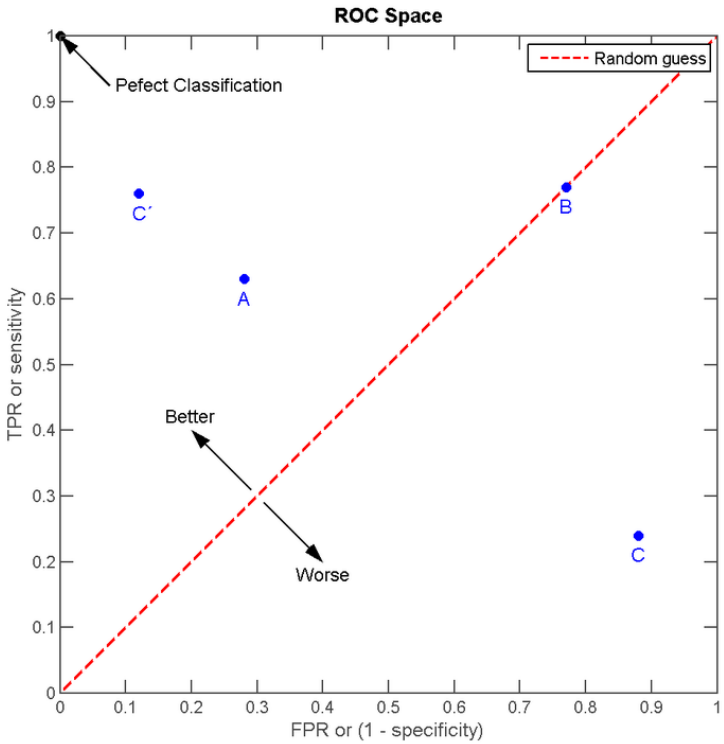
F1

$$F1_{Score} = 2 \times \frac{(Recall \times Precision)}{(Recall+Precision)}$$

Medidas de precisión

Curva de ROC (*Receiver Operating Characteristic*)

Gráfica que representa cómo varía el ratio de verdaderos positivos ($TPR = \frac{TP}{P}$) frente al ratio de falsos positivos ($FPR = \frac{FP}{N}$).



Area bajo la curva (*Area Under the Curve, AUC*)

Proporciona una medida de lo bien que un algoritmo de clasificación permite distinguir entre las clases.

Domain generation algorithm

DGA

Generan nombres de dominio que pueden utilizarse para acceder a servidores de comando y control (C&C) o a proveedores de cualquier otro tipo de *malware*.

Domain generation algorithm

DGA

Generan nombres de dominio que pueden utilizarse para acceder a servidores de comando y control (C&C) o a proveedores de cualquier otro tipo de *malware*.

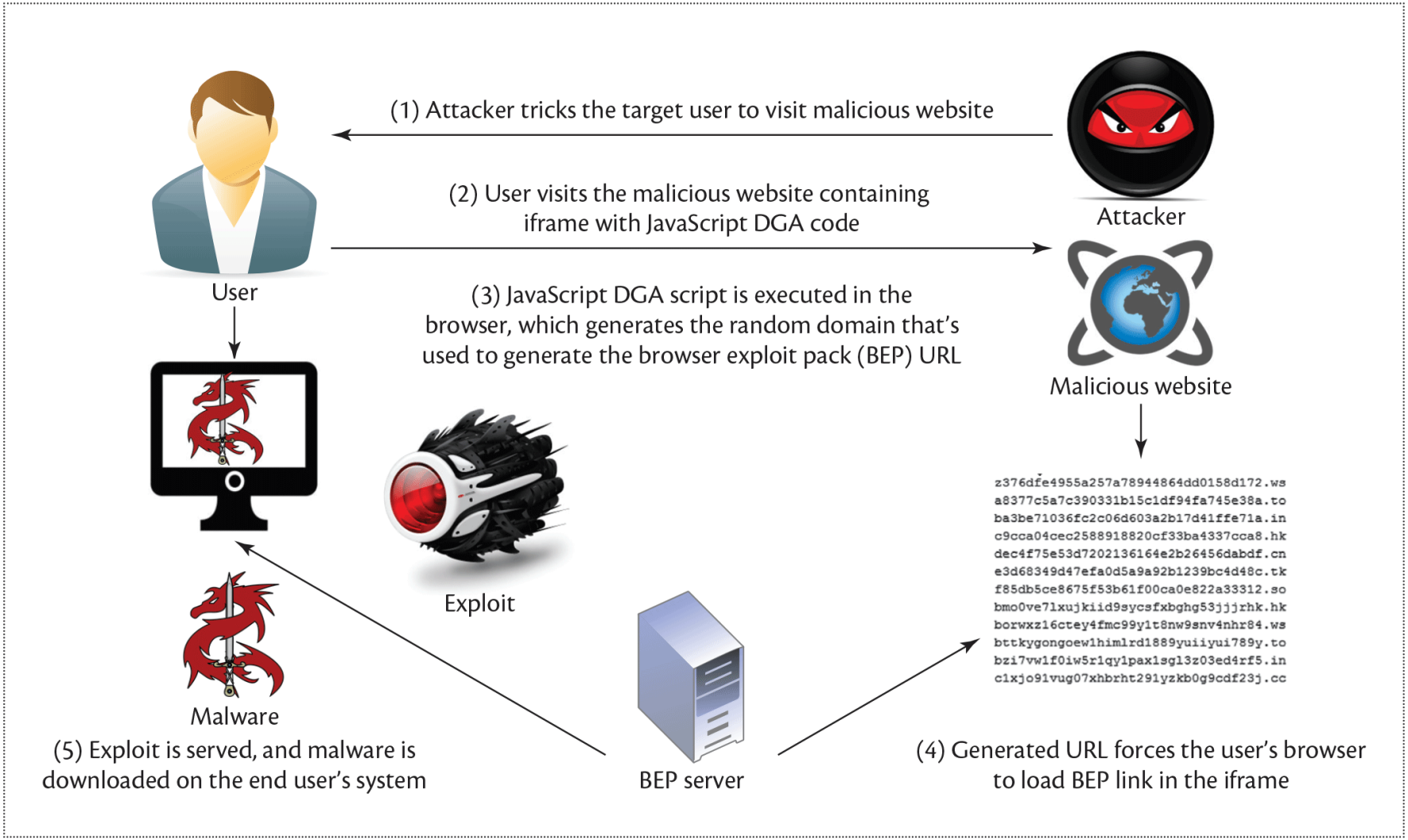
Ejemplo en Python (de wikipedia!)

```
def generate_domain(year, month, day):
  """Generates a domain name for the given date."""
  domain = ""

  for i in range(16):
    year = ((year^8 * year) >> 11)^((year & 0xFFFFFFF0) << 17)
    month = ((month^4 * month) >> 25)^16 * (month & 0xFFFFFFF8)
    day = ((day^(day << 13)) >> 19)^((day & 0xFFFFFFFE) << 12)
    domain += chr(((year^month^day) % 25) + 97)

return domain + ".com"
```

Domain generation algorithm



Sood, A.K., & Zeadally, S. (2016). A Taxonomy of Domain-Generation Algorithms. *IEEE Security & Privacy*, 14, 46-53.

Domain generation algorithm

Abordemos la identificación de DGA mediante aprendizaje automático:

Domain generation algorithm

Abordemos la identificación de DGA mediante aprendizaje automático:

1. Analizaremos y calcularemos algunas características de los nombres de dominios.

Domain generation algorithm

Abordemos la identificación de DGA mediante aprendizaje automático:

1. Analizaremos y calcularemos algunas características de los nombres de dominios.
2. Entrenaremos algoritmos de aprendizaje automático para que *"aprendan"* a distinguir un dominio legítimo y uno generado.

Domain generation algorithm

Abordemos la identificación de DGA mediante aprendizaje automático:

1. Analizaremos y calcularemos algunas características de los nombres de dominios.
2. Entrenaremos algoritmos de aprendizaje automático para que *"aprendan"* a distinguir un dominio legítimo y uno generado.
3. Validaremos los algoritmos para determinar el mejor.

¿De dónde sacamos los datos?

Direcciones legítimas de Alexa

- La compañía: `https://www.alexa.com`
- El dataset: `http://s3.amazonaws.com/alexa-static/top-1m.csv.zip`

Direcciones DGA de Bambenek Consulting

- La compañía: `http://www.bambenekconsulting.com/`
- El dataset: `http://osint.bambenekconsulting.com/feeds/dga-feed.txt`

N-gramas

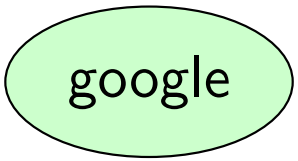
¿Cuáles son las secuencias "legítimas"?

N-gramas: Agrupaciones de 'N' elementos de una secuencia

N-gramas

¿Cuáles son las secuencias "legítimas"?

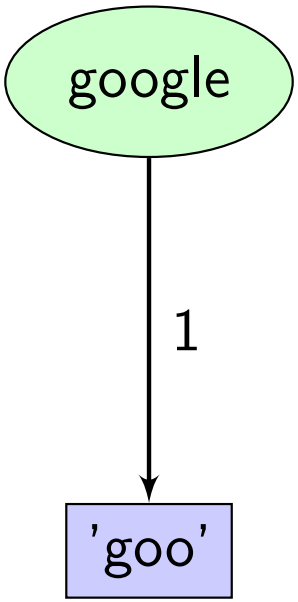
N-gramas: Agrupaciones de 'N' elementos de una secuencia



N-gramas

¿Cuáles son las secuencias "legítimas"?

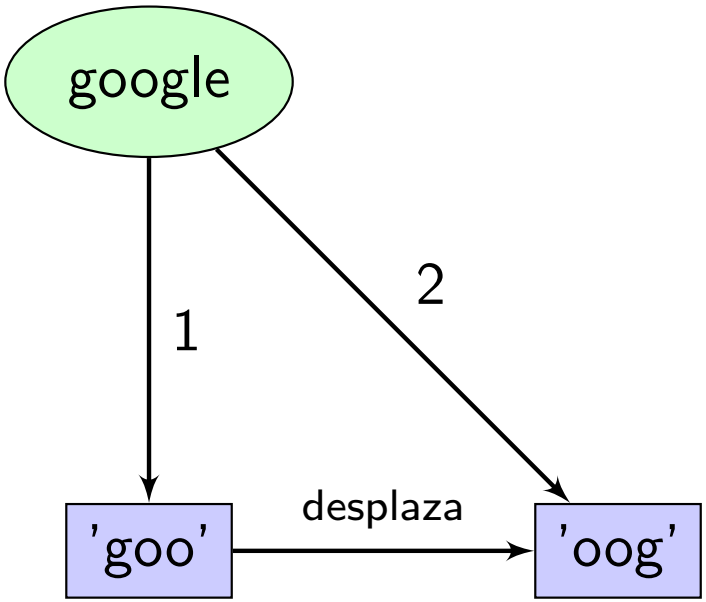
N-gramas: Agrupaciones de 'N' elementos de una secuencia



N-gramas

¿Cuáles son las secuencias "legítimas"?

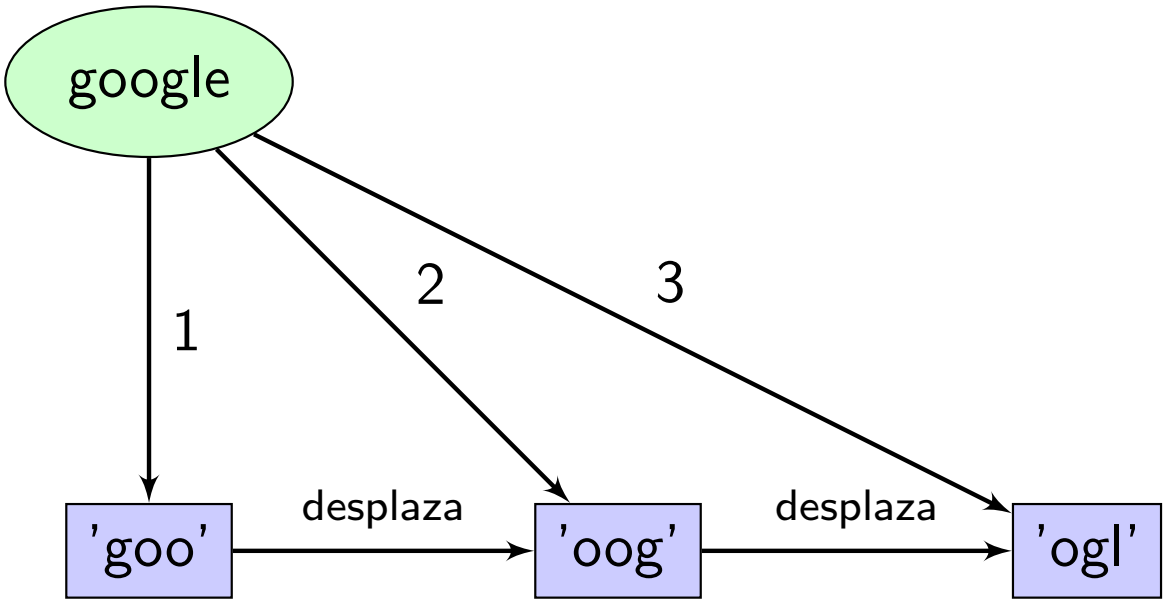
N-gramas: Agrupaciones de 'N' elementos de una secuencia



N-gramas

¿Cuáles son las secuencias "legítimas"?

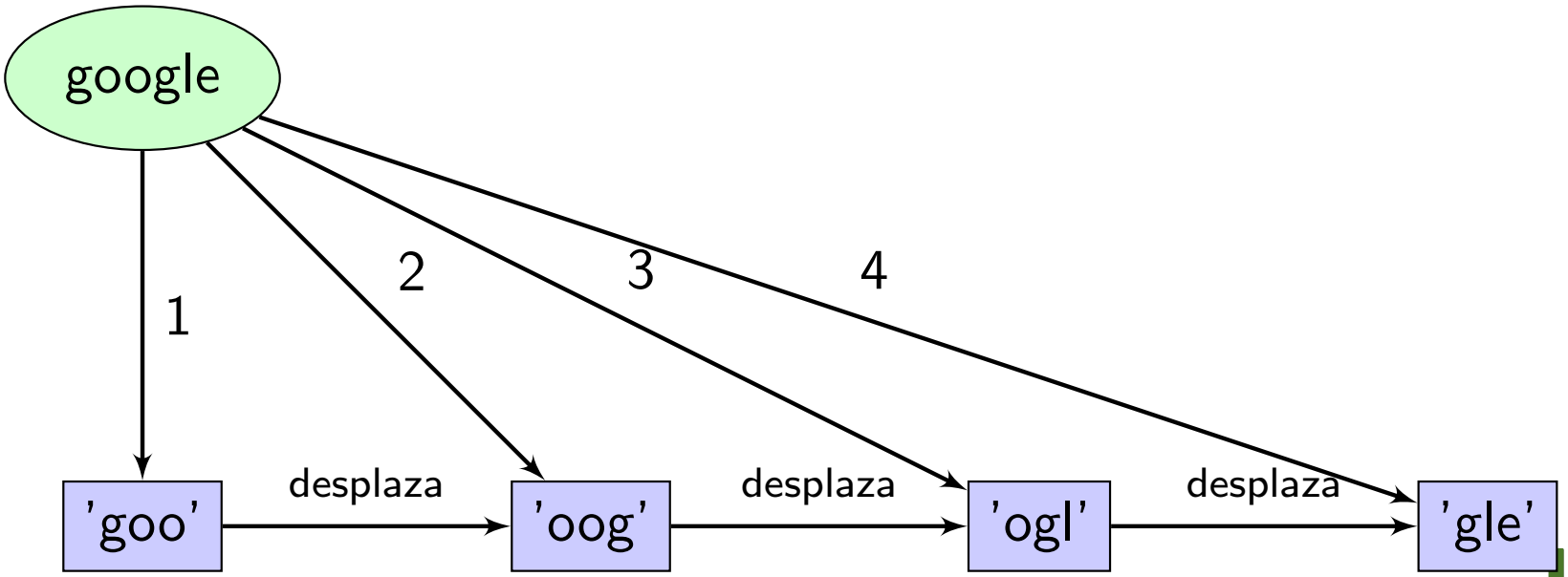
N-gramas: Agrupaciones de 'N' elementos de una secuencia



N-gramas

¿Cuáles son las secuencias "legítimas"?

N-gramas: Agrupaciones de 'N' elementos de una secuencia



Desorden

Entropía de Shannon (teoría de la información)

Da una medida del desorden de las combinaciones.

Desorden

Entropía de Shannon (teoría de la información)

Da una medida del desorden de las combinaciones.

¿Cómo se calcula?

Sea S una cadena de información y X un símbolo de la cadena:

$$Entropy\ H(X) = -\sum_{X \in S} \left(p(X) \log_2 p(X) \right)$$

$$p(X) = \frac{count(X)}{len(S)}$$

Contenidos

Reglas de asociación

- Usadas para descubrir hechos (asociaciones) que se dan dentro de un conjunto de datos.

$$Antecedentes \rightarrow Consecuentes$$

$$\{Variable_i = X, Variable_j = Y\} \rightarrow \{Variable_k = Z\}$$

- Los datos deben ser categóticos, nada de datos numéricos.

Reglas de asociación: Indicadores de interes

Soporte frecuencia en la que antecedentes aparecen en el dataset.

Confianza frecuencia en la que antecedentes y consecuentes se encuentran en el dataset, es decir, la frecuencia en al que la regla se ha evaluado como cierta.

Lift relación (ratio) del soporte observado respecto del soporte esperado si antecedente y consecuente fueran independientes.

- $lift = 1$ cantidad de entradas acorde a lo esperado bajo condiciones de independencia.
- $lift > 1$ cantidad de entradas superior a lo esperado bajo condiciones de independencia. Algo hace que las entradas se encuentren en el dataset más veces de lo normal.
- $lift < 1$ cantidad de entradas inferior a lo esperado bajo condiciones de independencia. Algo hace que las entradas no se encuentren en el dataset más veces de lo normal.

Reglas de asociación: Ejemplo

- Emplearemos el dataset del "titanic".
- Instalaremos el *add-on* llamado "Associate" (menú $Options \rightarrow Add-ons\ldots$)

K-means

K-means

- Objetivo: particionar un conjunto de n observaciones en k grupos.
- Cada observación pertenece al clúster cuyo centroide es más cercano.
- Los centroides de cada clúster son la media de los valores de cada variable.

- Quizá uno de los más populares.
- Todos los valores deben ser numéricos.
- Requiere especificar el número de clústers (agrupaciones).

Jerárquico aglomerativo

Agrupamiento jerárquico

Método de análisis de clústers que busca construir una jerarquía de dichos clústers.

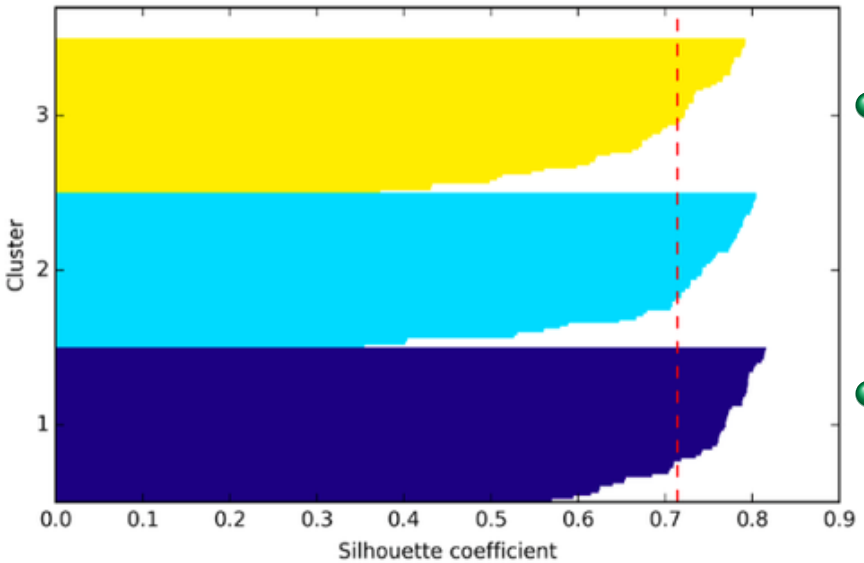
Estratégia aglomerativa

- Aproximación de abajo-arriba (*bottom-up*) donde cada observación comienza siendo su propio clúster, y pares de clústers se unen según nos movemos hacia arriba en la jerarquía.
- Para decidir qué clústers combinar se emplea una medida de distancia entre conjutos de observaciones.
  Ej. Euclídea, Euclídea cuadrado, Manhatan, Máximo, Mahalanobis, Coseno, Jaccard, ...

Midiendo la calidad de los clústers mediante la silueta

Silueta (*Silhouette*)

- Método de interpretación y validación de la consistencia de los datos dentro de los clústers.
- Gráfica que muestra la distancia promedio entre las instancias de los datos dentro del clúster y las instancias con el clúster más cercano.



- Para cada instancia de datos:
  - Si silueta cerca de 1 $\rightarrow$ instancia cerca del centro del clúster.
  - Si silueta cerca de 0 $\rightarrow$ instancia en el borde entre dos clústers.
- La calidad de la agrupación se evalua como el valor promedio de las siluetas de las instancias de los datos que contiene.