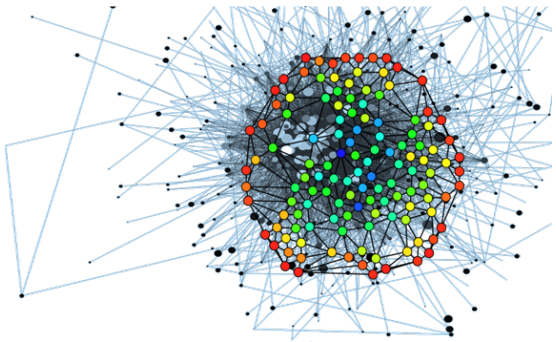
UNIVERSIDAD LOYOLA ANDALUCÍA
Departamento de Métodos Cuantitativos

Seminario: el proceso de ciencia de datos y problemas de clasificación

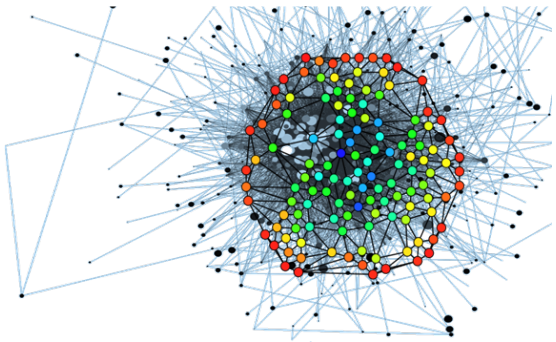
Javier Sánchez-Monedero

Departamento de Métodos Cuantitativos
Universidad Loyola Andalucía
10 de diciembre de 2015

Índice



Índice



Un mundo de datos I

Nuestro mundo gira cada vez más en torno a los datos:

- **Ciencia**: astronomía, genómica, medio-ambiente...
- **Industria y Energía**: redes de sensores, IoT, gestión parques eólicos, previsión de demanda, ciudades inteligentes...
- **Ciencias sociales y humanidades**: libros digitalizados, documentos históricos, datos sociales...

Un mundo de datos II

- **Entretenimiento**: sistemas de recomendación, contenidos digitales, búsquedas multimedia...
- **Medicina**: examen de imágenes médicas, previsión de demanda en hospitales, sistemas expertos...
- **Financias y negocios**: transacciones de mercados automatizadas...

Explosión de datos I

Aunque hace décadas que existen los analistas de datos, también hace décadas que se almacenan datos que no han podido ser procesados hasta hace pocos años:

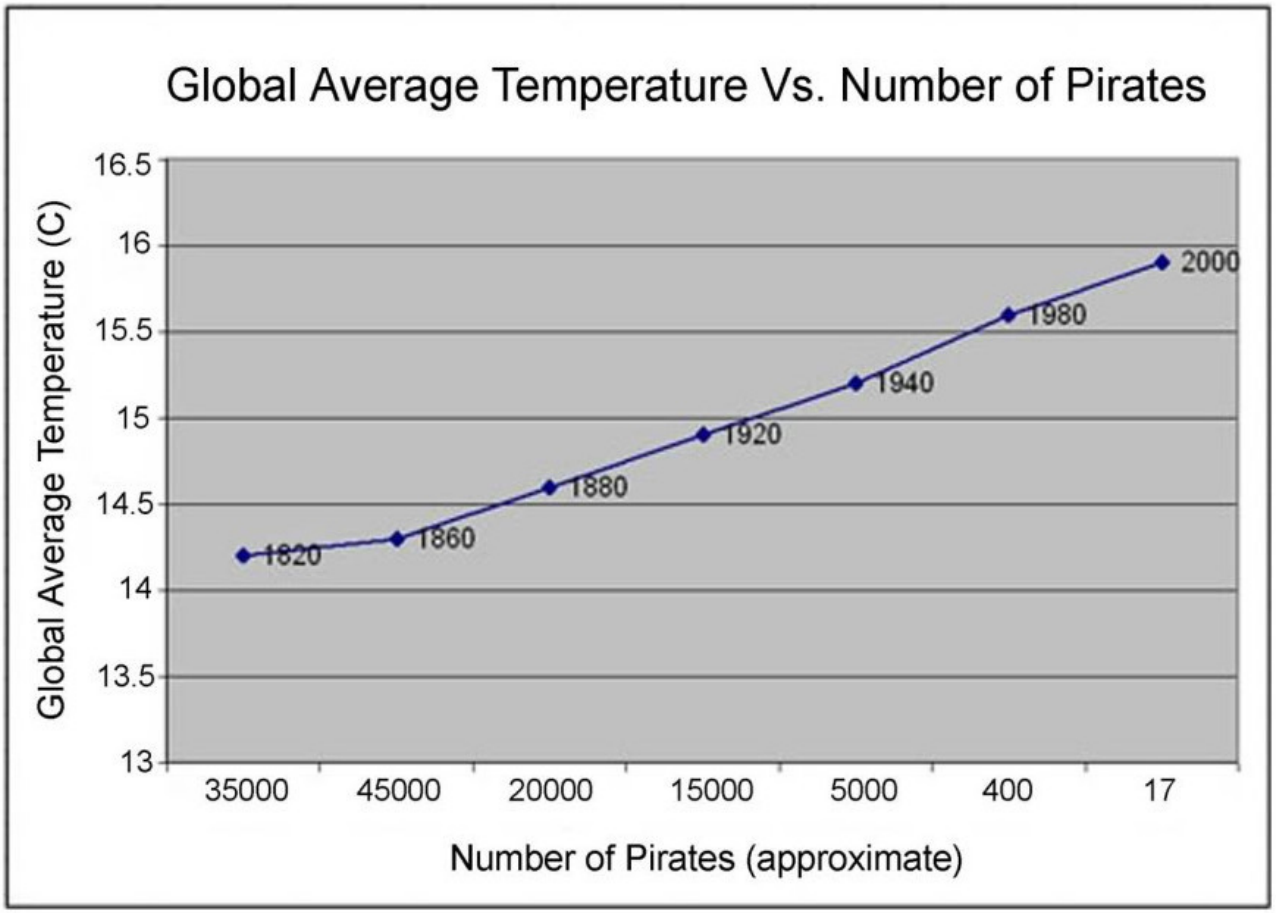
Figura : Fuente Big Band Data

- Tecnologías de bases de datos
- Coste del hardware de almacenamiento
- Aumento del ancho de banda
- Aumento capacidad de procesado
- Software científico

Explosión de datos II

Todo esto nos capacita para pasar de la **información** al **conocimiento**.

Precaución


Global Average Temperature Vs. Number of Pirates
Global Average Temperature (C)
16.5
16
15.5
15
14.5
14
13.5
13
1820
1860
1880
1920
1940
1980
2000
35000
45000
20000
15000
5000
400
17
Number of Pirates (approximate)

Precaución

¡Para frenar el calentamiento global: hagámonos piratas!

Correlación no implica causalidad

Precaución

Stop a la obsesión por los datos

Los datos no lo son todo. También está el conocimiento cualitativo. El análisis de datos puede ayudar en la toma de decisiones, pero siempre **junto al conocimiento de expertos**.
A veces un conocimiento será más adecuado que otros, a veces la mejor solución será una combinación de conocimientos cualitativos y cuantitativos[1].

Ciencia de datos

Estamos en un programa de doctorado de ciencia de datos donde el objetivo es desarrollar y/o aplicar métodos de análisis de datos a problemas reales.

[1] Presentación del proyecto de investigación 'Cherry-picking: los resultados de los procesos participativos'

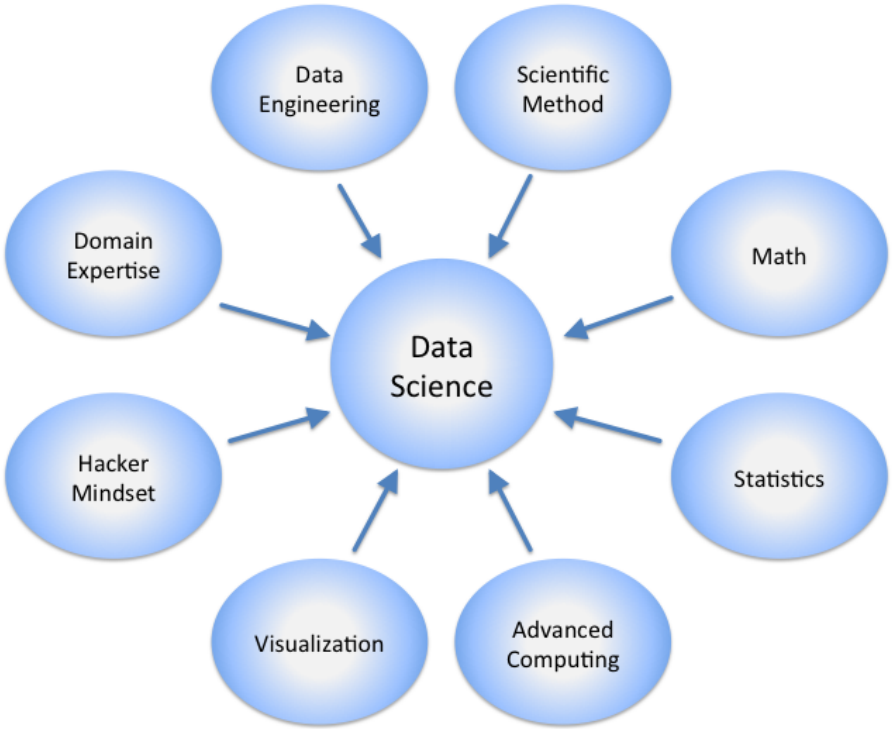
Objetivos de la sesión

- Pequeña introducción a la ciencia de datos.
- Tipos y técnicas de minería de datos.
- Modelos y algoritmos de clasificación de referencia.
- Evaluación de modelos.

Definición de ciencia de datos

Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos

Habilidades del científico de datos



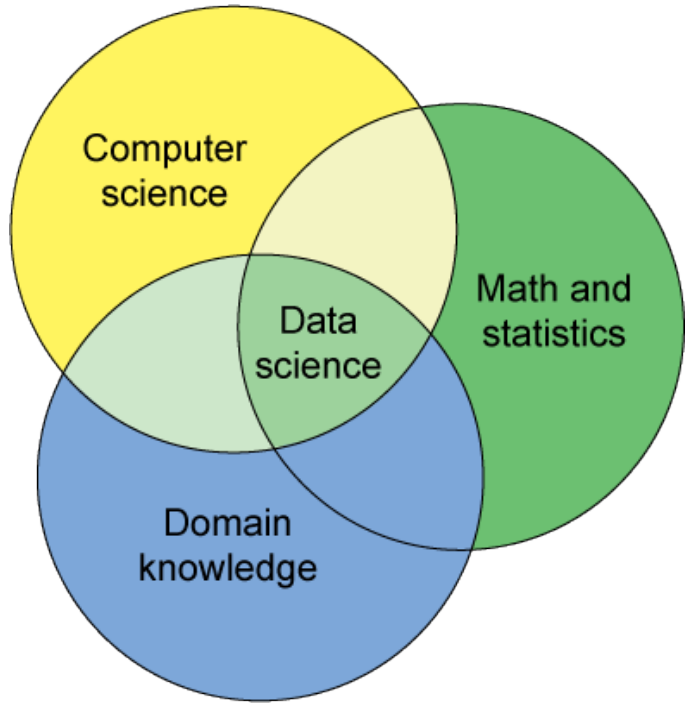


Figura : Fuentes https://www.dreamhost.com/blog/2015/10/22/so-you-want-to-be-part-of-the-data-science-revolution/ y http://www.ibm.com/developerworks/library/os-datascience/

Salidas profesionales

Figura : Anatomy of a Data Scientist

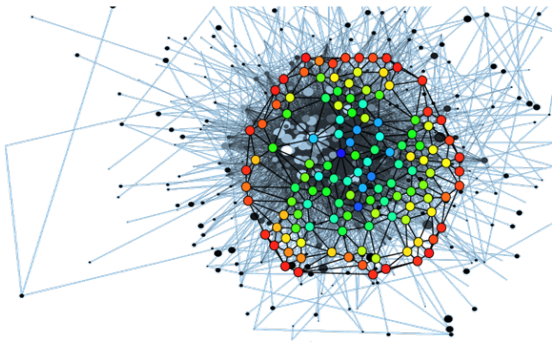
- Científico de datos: habilidades de **programación (distribuida)**, **aprendizaje automático** y **conocimiento específico**.
- Denominado "*the sexiest job of the 21st century*" por el *Harvard Business Review*.
  `https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1`
- Los salarios oscilan entre $110,000$ a $140,000$\$.
- Perfiles exigentes (Machine Learning en Spotify).
- En 2015, 4.4 millones de trabajos se han creado para soportar tareas de *big data*.

¿Qué es un científico de datos?

José Antonio Guerrero: uno de los mejores científicos de datos del mundo (Plataforma Kaggle)

"Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos. De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco."
Fuente, El Confidencial

Índice



Minería de datos y KDD I

Minería de datos

"La **Minería de datos (MD)** es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos." [1]

Aunque *Data Science* y *Big Data* son términos más actuales, desde 1989 se denomina a actividades similares como **KDD** (*Knowledge Discovery from Databases*) o **descubrimiento de conocimiento en bases de datos**.

- El KDD es el **proceso completo de extracción de conocimiento** a partir de bases de datos

Minería de datos y KDD II

- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La **Minería de Datos** es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

Aportación del término ciencia de datos

Tal vez el término "ciencia de datos" añada más actividades, como por ejemplo el énfasis en la visualización de datos, o el trabajar con datos no estructurados (algo bastante común en el área del *big data*).

¿Para qué?

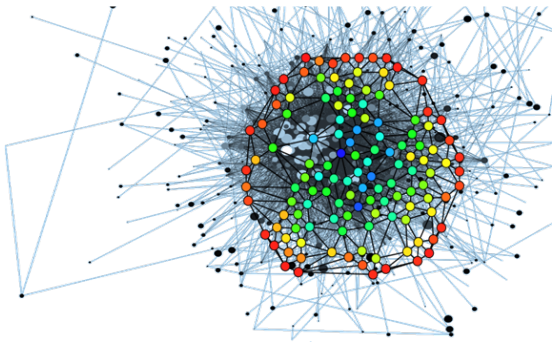
- **Resumir** una gran base de datos
- **Visualizar** datos multi-dimensionales
- **Predecir** valores
- **Explicar** los datos existentes

Orígenes de datos

Las fuentes de datos son muy variadas, a menudo incluso se mezclan, dando lugar a disciplinas como *fusión de información*:

- Bases de datos relacionales
- Bases de datos espaciales y/o temporales: telefonía móvil
- Bases de datos de documentos
- Bases de datos multimedia: imágenes, vídeos, sonidos...
- La *World Wide Web*
- Grandes volúmenes de datos no estructurados (*Big Data*)

Índice



Etapas en el proceso de KDD I

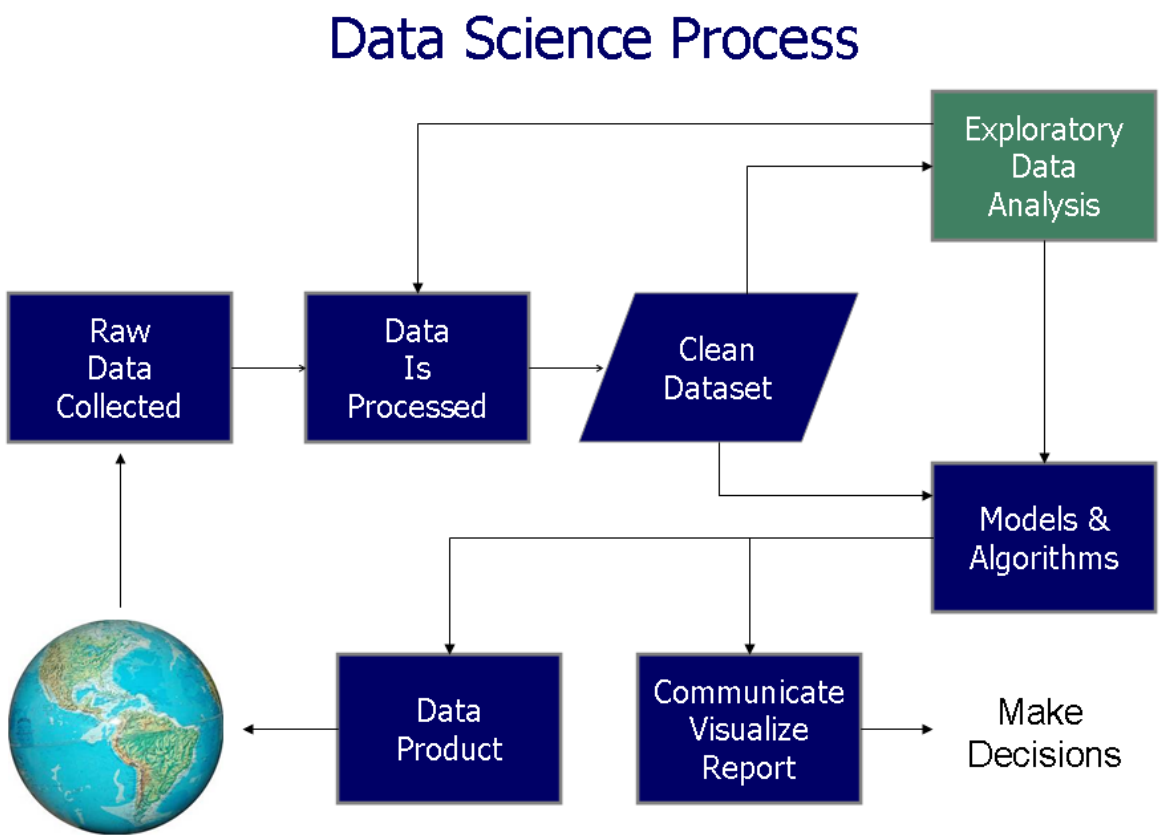


Figura : Fuente https://en.wikipedia.org/wiki/File:Data_visualization_process_v1.png

Etapas en el proceso de KDD II

Según F. Herrera [1]:

1. **Integración y recopilación**: Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori y creación del **almacén de datos** (*Datawarehouse*)
2. **Preprocesamiento**: Selección de datos, limpieza, reducción y transformación
3. **Selección de la técnica** de MD y aplicación de algoritmos concretos de MD
4. **Evaluación**, **interpretación** y **presentación** de los resultados obtenidos
5. **Difusión** y utilización del **nuevo conocimiento**

Etapas en el proceso de KDD III

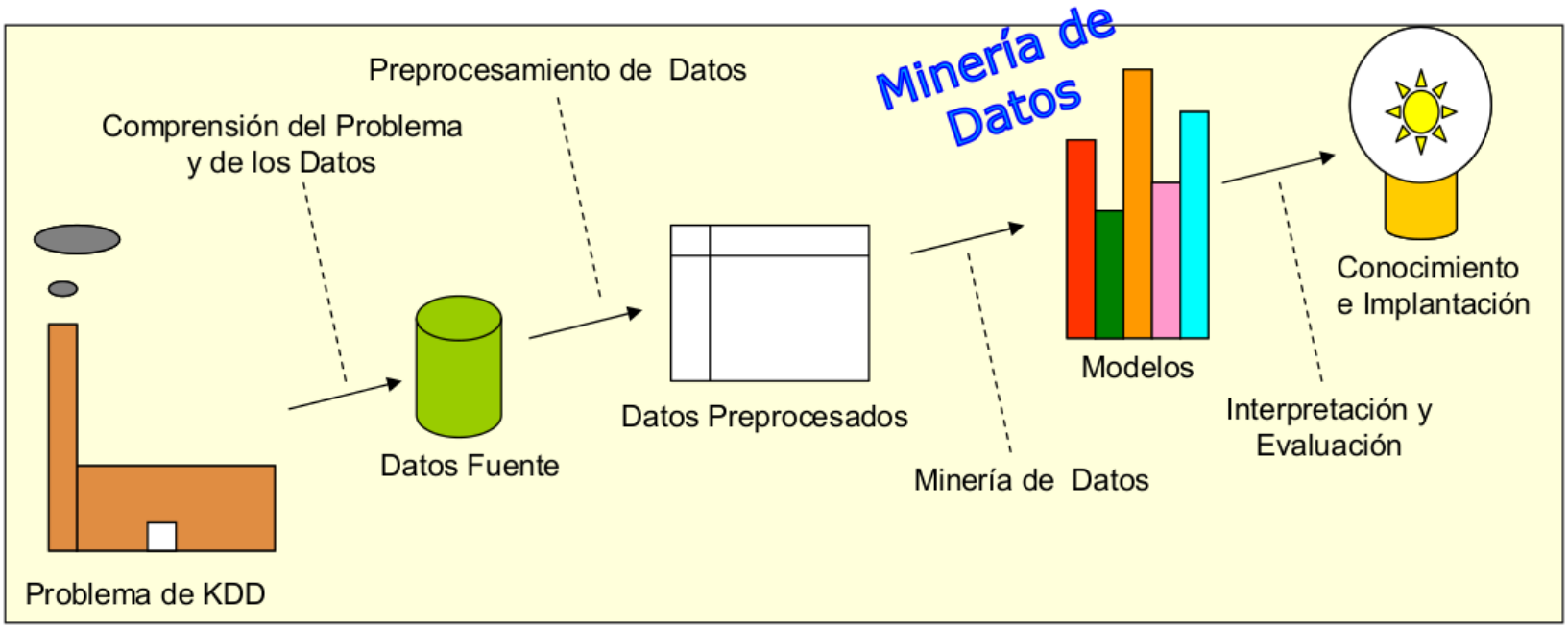
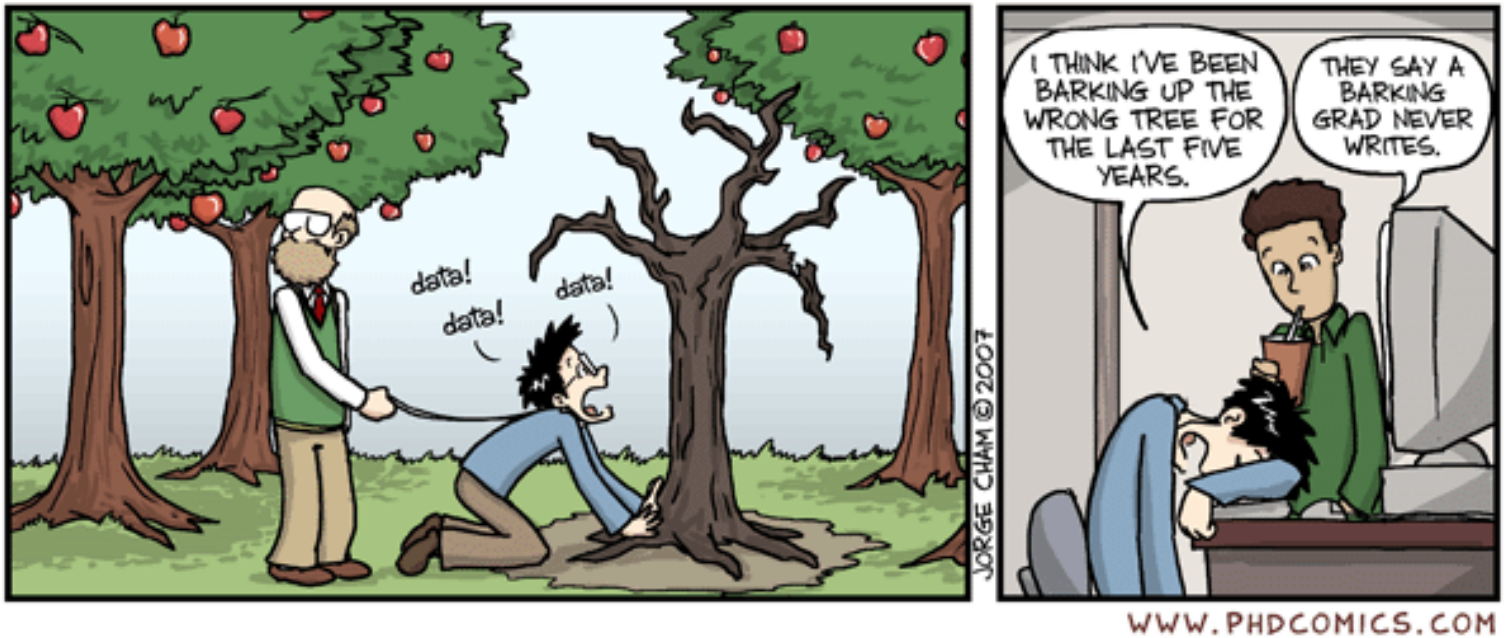


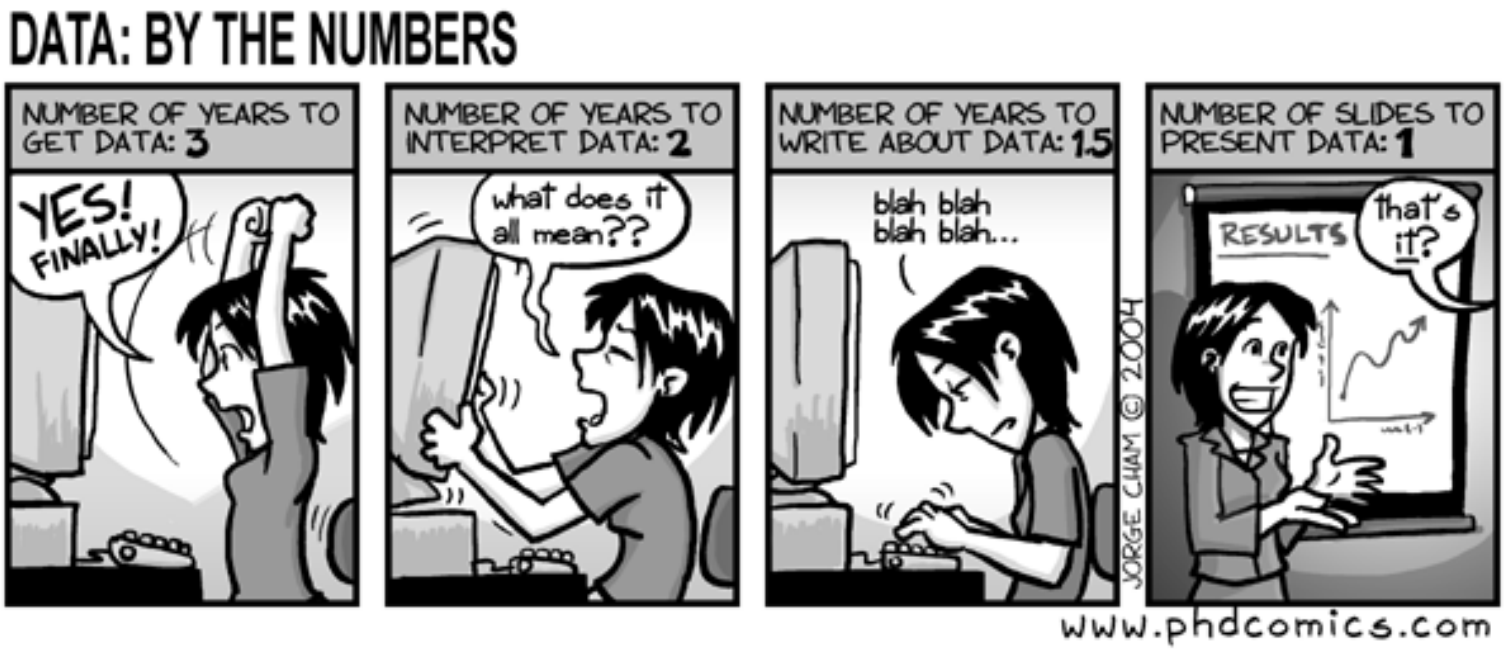
Figura : Etapas en el proceso de KDD, fuente [1]

¿Qué etapa lleva más esfuerzo?

¿Qué etapa lleva más esfuerzo en el proceso de ciencia de datos?



¿Qué etapa lleva más esfuerzo?



¿Qué etapa lleva más esfuerzo?

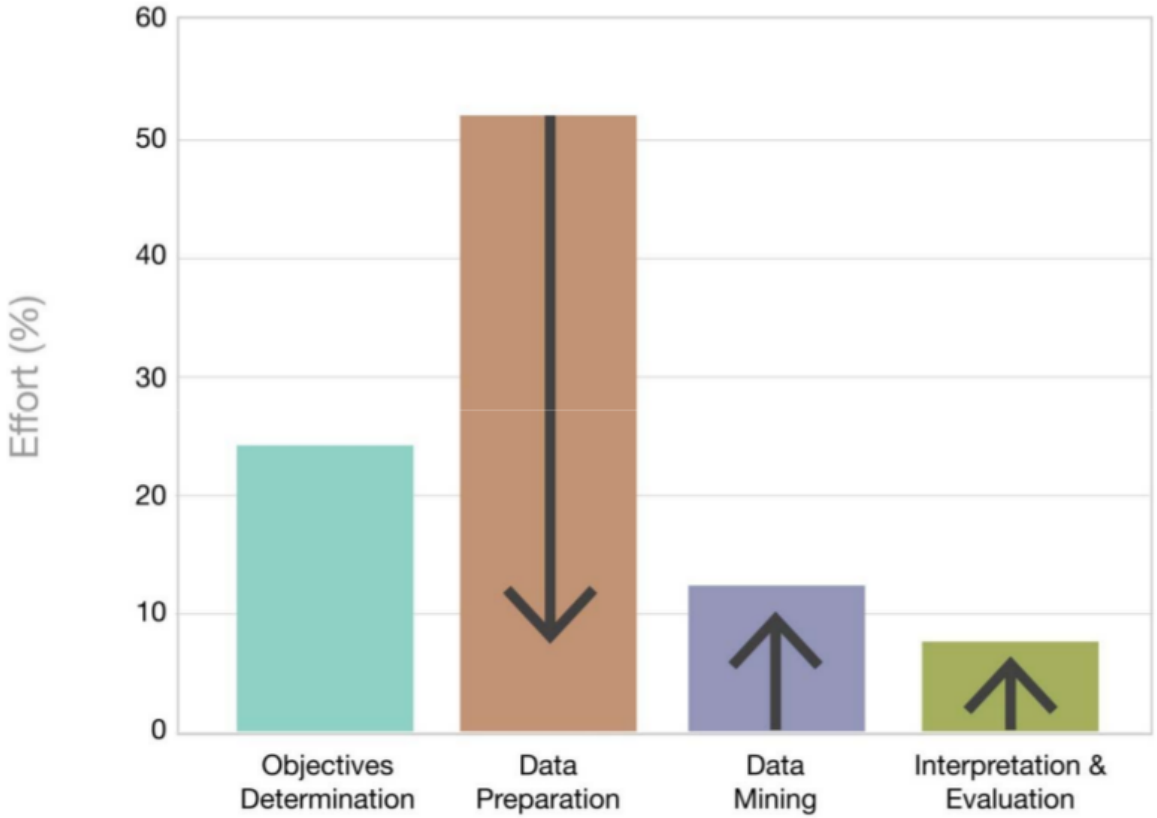


Figura : Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos

Etapas: selección y preprocesado I

A menudo, los resultados van a depender más de la **calidad de los datos en relación al problema** que de la parte de minería de datos en sí:

- Hablaremos de **ruido** en los datos, de **relevancia** de variables... siempre **respecto a un objetivo**. Un variable que consideremos ruido puede ser información útil para otro problema distinto.
- Disponemos de **potentes modelos de aprendizaje automático** no lineales capaces de ajustarse a datos complejos y de alta dimensionalidad. Por ejemplo las redes neuronales son *aproximadores universales*.

Etapas: selección y preprocesado II

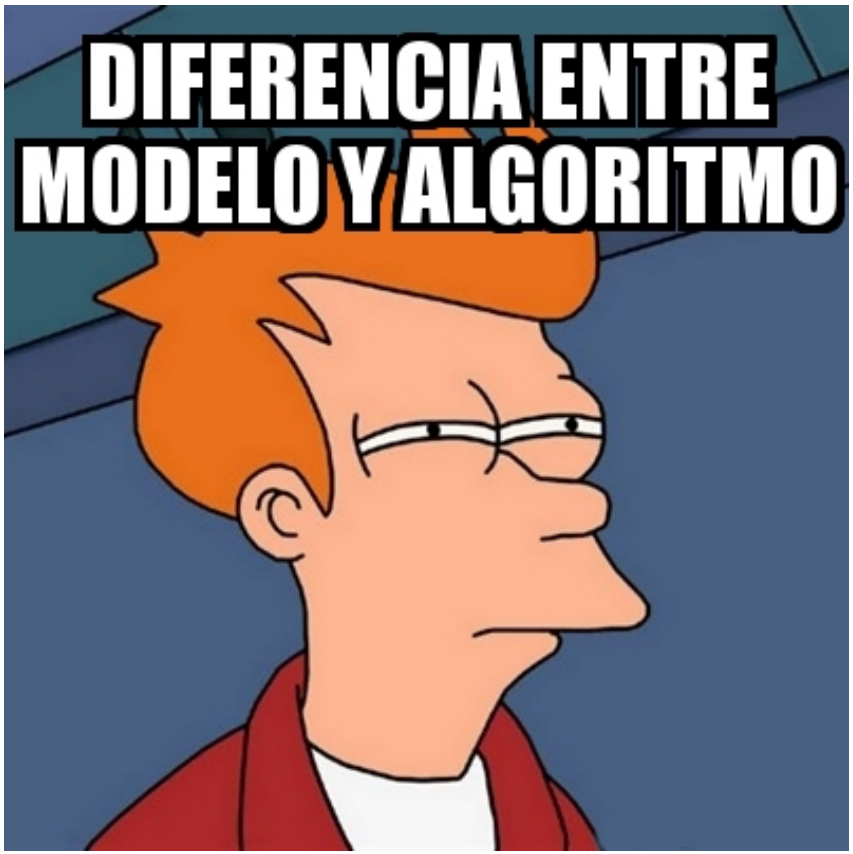
Tareas de esta etapa:

- **Selección** de datos (variables). Ej. Utilizaremos el índice de masa corporal para caracterizar el riesgo de padecer un infarto.
  - **Extracción de características**. Ej. al procesar datos multimedia se extraen características que permitan construir vectores de tamaño fijo necesarios para los modelos.
  - **Selección de características descartables** para reducir el número de variables. Ej. en el campo de análisis de proteínas puede haber cientos de miles de variables de entrada.
- **Limpieza de datos**:
  - Recuperación de valores perdidos (imputación de datos)
  - Tratamiento de valores anómalos (*outliers*)
  - Suavizar ruido
  - Eliminar inconsistencias

Etapas: selección y preprocesado III

- **Transformación o preprocesamiento** de datos (imprescindible para muchas técnicas de MD). Ej. convertir una variable nominal en varias variables binarias, escalado de datos, fecha nacimiento $\rightarrow$ edad...
- **Reducción de dimensionalidad**. No es exactamente lo mismo que la selección de características, ya que se emplean técnicas como PCA para usar combinaciones lineales de variables que reduzcan la dimensión de los datos, pero consideren todas las variables.

Inciso



Diferencia entre modelo y algoritmo

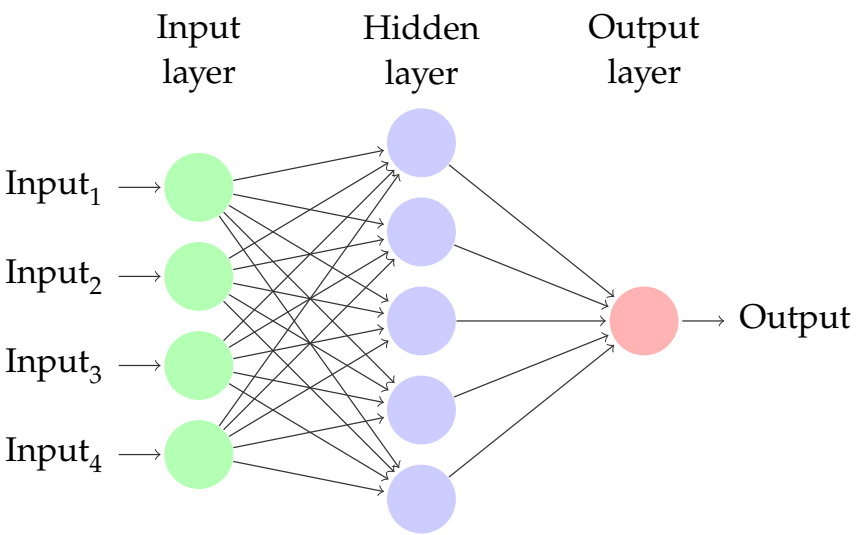


Figura : Modelo de red neuronal

Un **modelo** es, en general, una función o una estructura de datos, que puede representar el conocimiento subyacente en un conjunto de datos. Dependiendo del objetivo del problema, tipo de variables de entrada, tipo de salida esperada, complejidad de los datos, etc. habrá modelos más o menos apropiados.

Diferencia entre modelo y algoritmo

- Un **algoritmo** es una secuencia de pasos con un fin. En informática todos los programas se descomponen en múltiples algoritmos que interaccionan entre si.
- En el contexto de aprendizaje automático, un **algoritmo de aprendizaje** se encarga de que **el modelo aprenda de los datos**, o expresado de otra forma, de que **el modelo se ajuste a los datos**.
- Existen algoritmos correspondientes al campo de la optimización numérica, pero también otros se basan en el aprendizaje estadístico o la inteligencia computacional.

Modelo entrenado

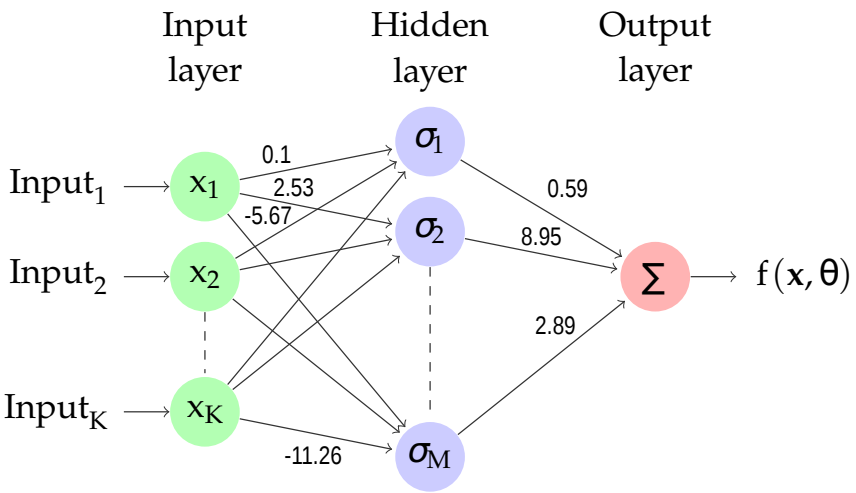


Figura : Modelo de red neuronal entrenado

Ejemplo de red neuronal

En el caso de una red neuronal artificial (RNA), entrenar un modelo consiste en asignar pesos a las conexiones de la red que minimicen una función de error, por ejemplo el error cuadrático medio. El modelo resultante depende de los datos de entrada, los parámetros del modelo (función de base, números de conexiones...) y el algoritmo de aprendizaje.

Etapas KDD: minería de datos I

Objetivo

Producir **nuevo conocimiento**, **construyendo/entrenando** un **modelo** basado en los datos recopilados que sea una descripción de los patrones y relaciones entre los datos con los que se puedan hacer predicciones, entender mejor los datos o explicar situaciones pasadas

Dependiendo del problema que queremos resolver:

- ¿Qué **tipo de conocimiento** buscamos?
  - Predictivo, Descriptivo
- ¿Qué **técnica** es la más adecuada?
  - Clasificación, regresión, agrupación, reglas de asociación, sistema de recomendación...

Etapas KDD: minería de datos II

- ¿Qué **tipo de modelo**?
  - En clasificación: árboles de decisión, redes neuronales, SVM...
- Otros requisitos:
  - ¿Necesitamos expresar grados de certeza, lógica difusa, interpretar el modelo...?
- ¿Qué **algoritmo de aprendizaje** es el más adecuado?
  - Descenso por gradiente, algoritmo evolutivo, algoritmo analítico...

Etapas KDD: evaluación I

Evaluación, **interpretación** y **presentación** de resultados

- Durante el análisis, generaremos varias **hipótesis de modelos**, ¿cuáles son más válidos?
- El teorema "no hay comida gratis" o "nada es gratis" (*No free lunch theorem*, http://www.no-free-lunch.org/) aplicado a la ciencia de datos nos dice que ningún método será el mejor en todos los casos.
- **Criterios** de validación:
  - Precisión
  - Interpretabilidad
  - Novedad de los patrones

Etapas KDD: evaluación II

- **Técnicas de evaluación** / diseños experimentales: como mínimo el conjunto de datos se divide en dos: entrenamiento y test/generalización.
  - Entrenamiento: extraer el conocimiento (ajustar el modelo)
  - Test: validar el modelo (el modelo debe probarse con datos no vistos durante el entrenamiento)
  - Diferentes diseños:
    - Validación simple
    - n-Validación cruzada
    - *Bootstrapping*
  - **Medidas de evaluación**, en función de la tarea:
    - Clasificación: precisión predictiva (%acierto), media geométrica de sensibilidad, curva ROC...
    - Regresión: Error cuadrático medio
    - Agrupamiento: métricas de cohesión y separación entre grupos

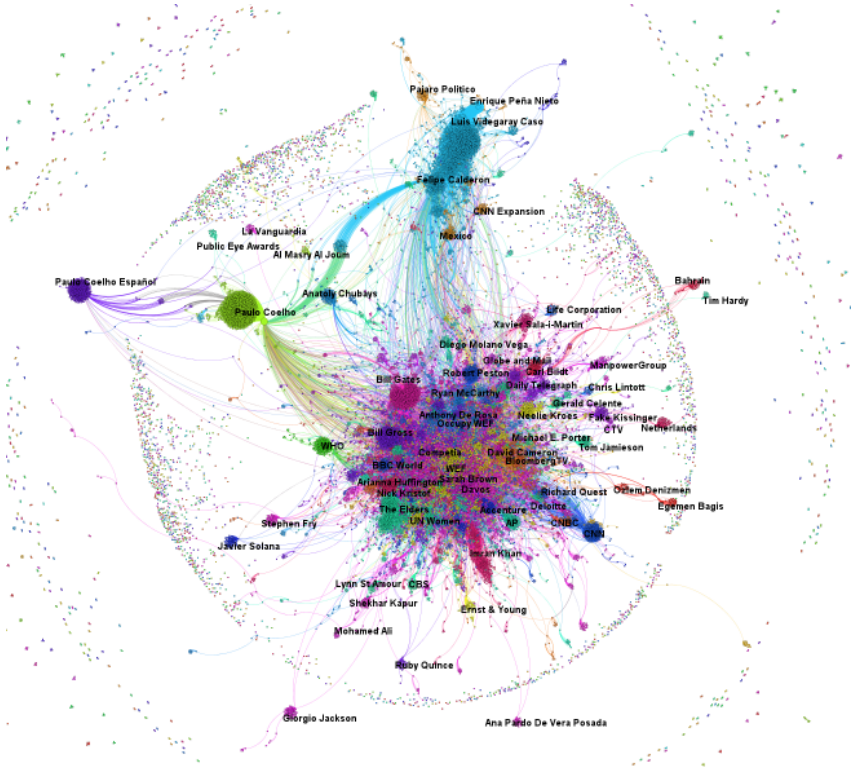
Etapas KDD: evaluación III

- Reglas de asociación: cobertura, confianza...

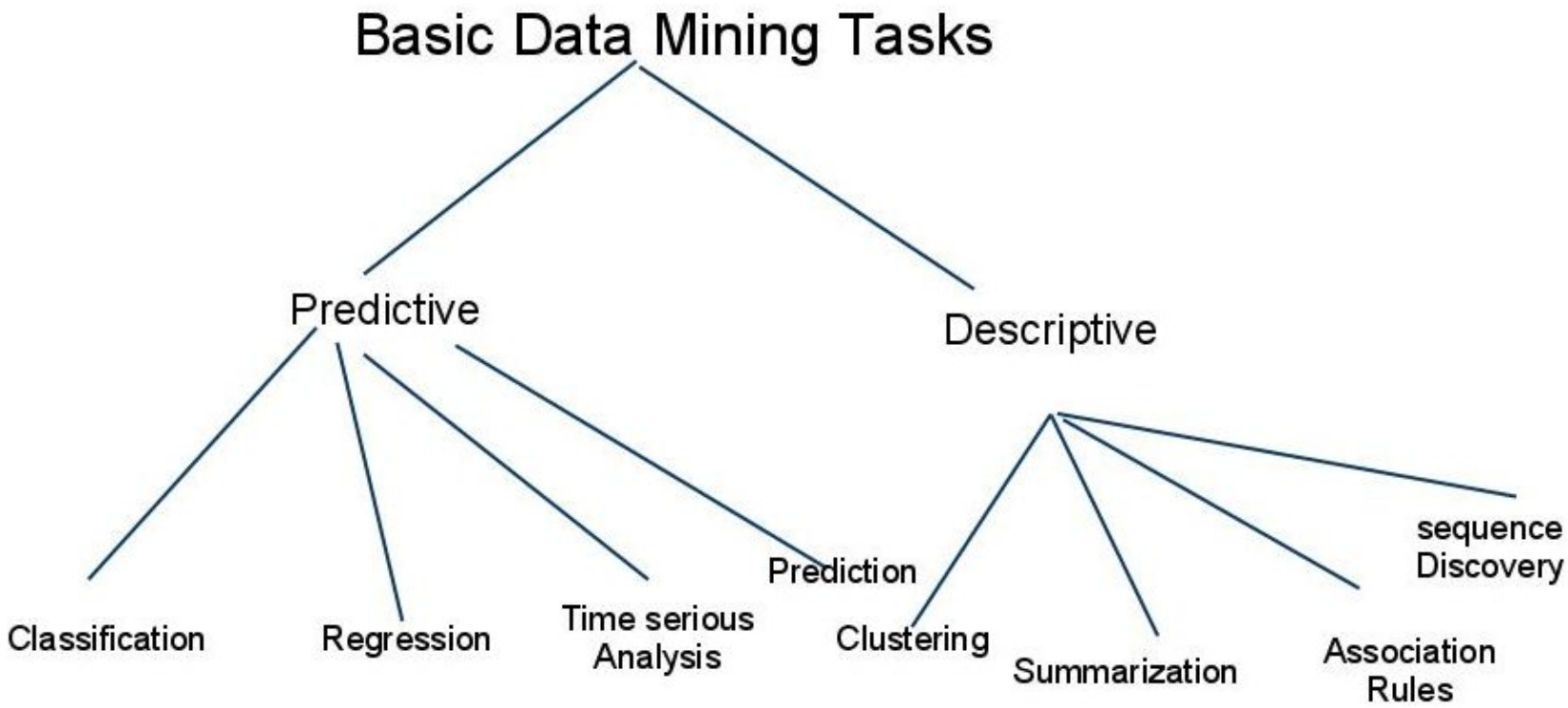
- ¿se puede o no **interpretar y/o visualizar** el modelo? (árboles de decisión vs. SVM/Redes Neuronales)

Etapas KDD: difusión I

- **Contraste** con conocimiento previo.
- **Observación** del sistema.
- **Realimentación** del sistema (aprendizaje *online*)
- **Arte**: el número de artistas implicados en la visualización de datos crece

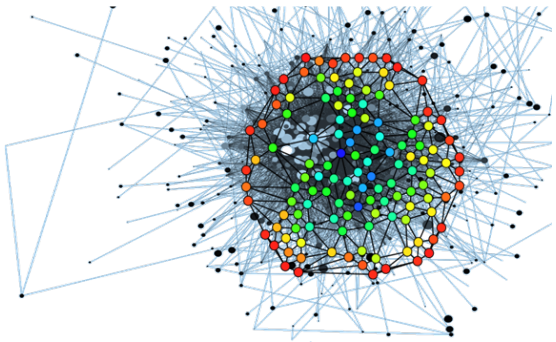


Técnicas de Minería de Datos



Conocer bien las herramientas

Índice



Aprendizaje automático

Machine learning

El *aprendizaje automático* o *aprendizaje máquina* (*machine learning* en inglés) se define como "campo de estudio que proporciona a los ordenadores la capacidad de aprender sin haber sido explícitamente programados".
El aprendizaje automático equivale a "aprender de los datos" con el fin de extraer el conocimiento necesario según diferentes propósitos

Este "**aprender de los datos**" hace que el aprendizaje automático se sitúe entre diferentes ramas que pertenecen a la inteligencia artificial, la estadística y las matemáticas

Aprendizaje automático

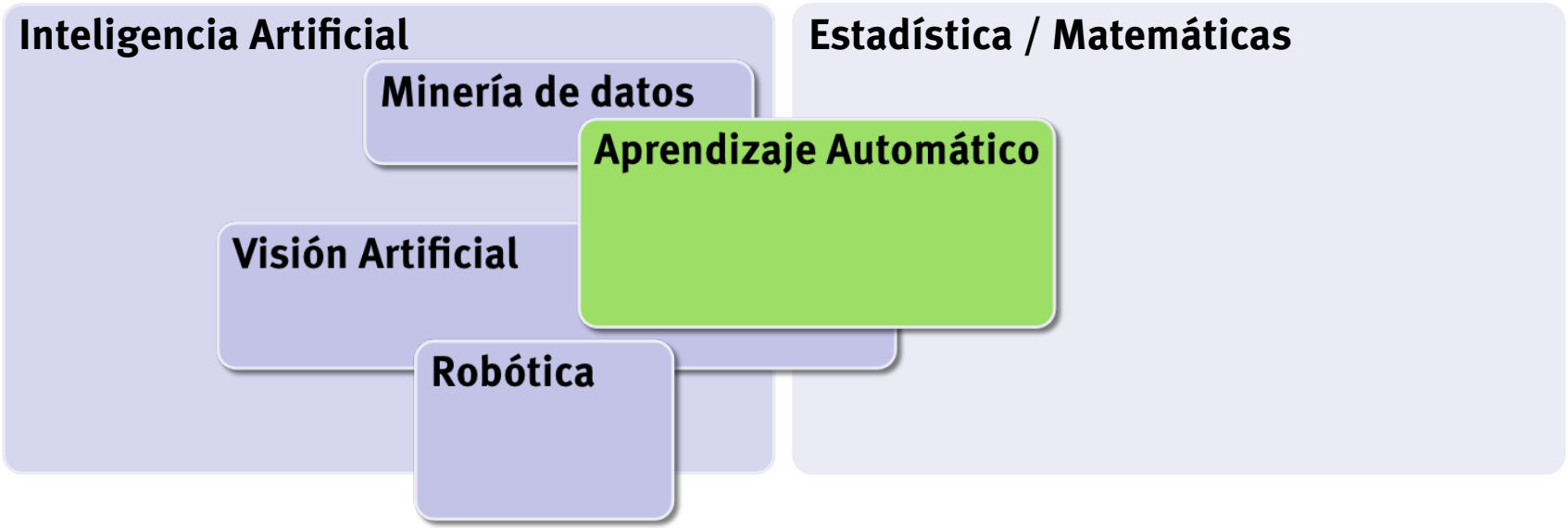
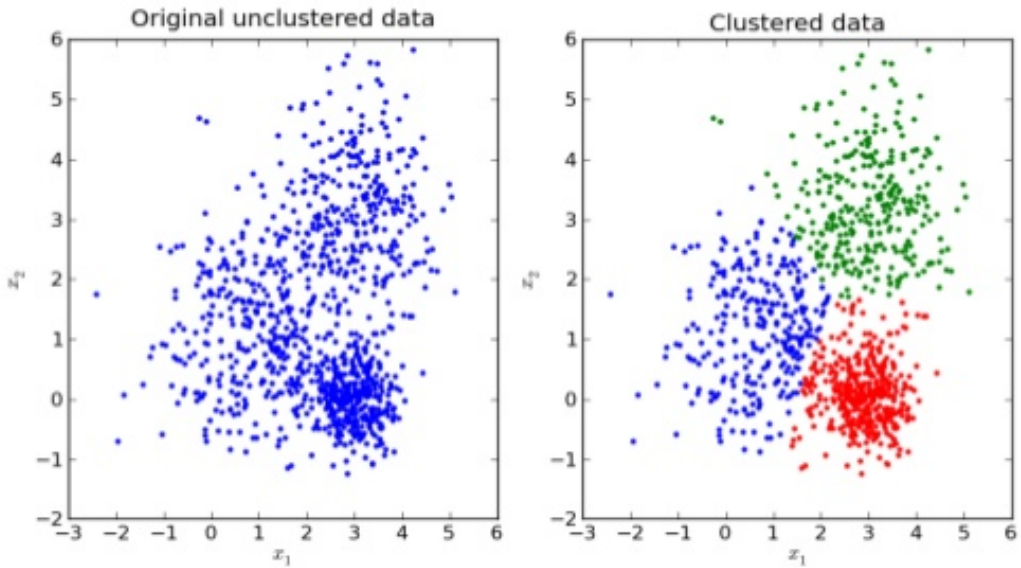


Figura : Aprendizaje automático: dónde encaja y dónde no

Descripción: agrupamiento

K-means Clustering

- partition ***n*** observations into ***k*** clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- http://en.wikipedia.org/wiki/K-means_clustering



http://pypr.sourceforge.net/kmeans.html

Ejemplo de aprendizaje no supervisado con el algoritmo de

Predicción: regresión

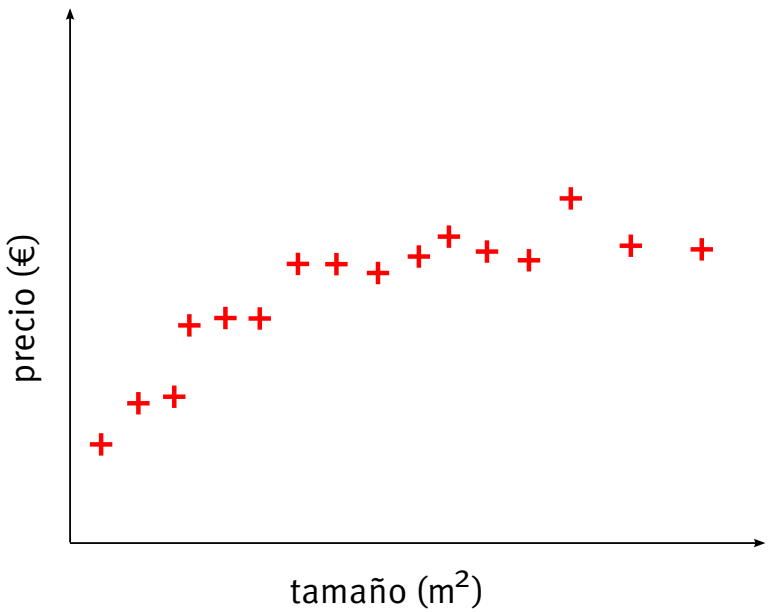
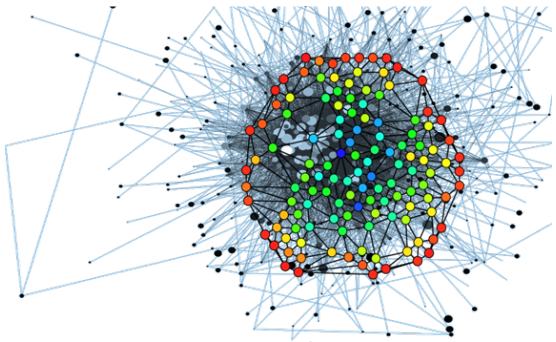


Figura : Ejemplo de problema de aprendizaje supervisado de regresión: Dados estos datos, un amigo tiene una casa de 75 metros cuadrados, ¿por cuánto podría esperar venderla?

Índice



Clasificación binaria

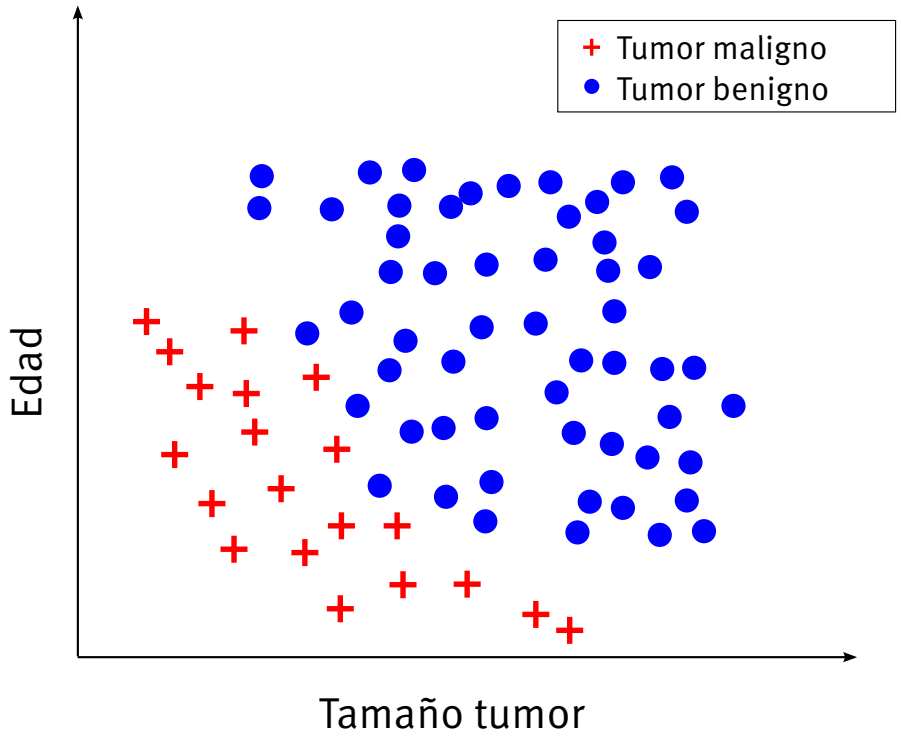


Figura : Ejemplo de problema de **clasificación** ¿Podrías estimar un diagnóstico basado en el tamaño del tumor y la edad del paciente?

Clasificación multi-clase

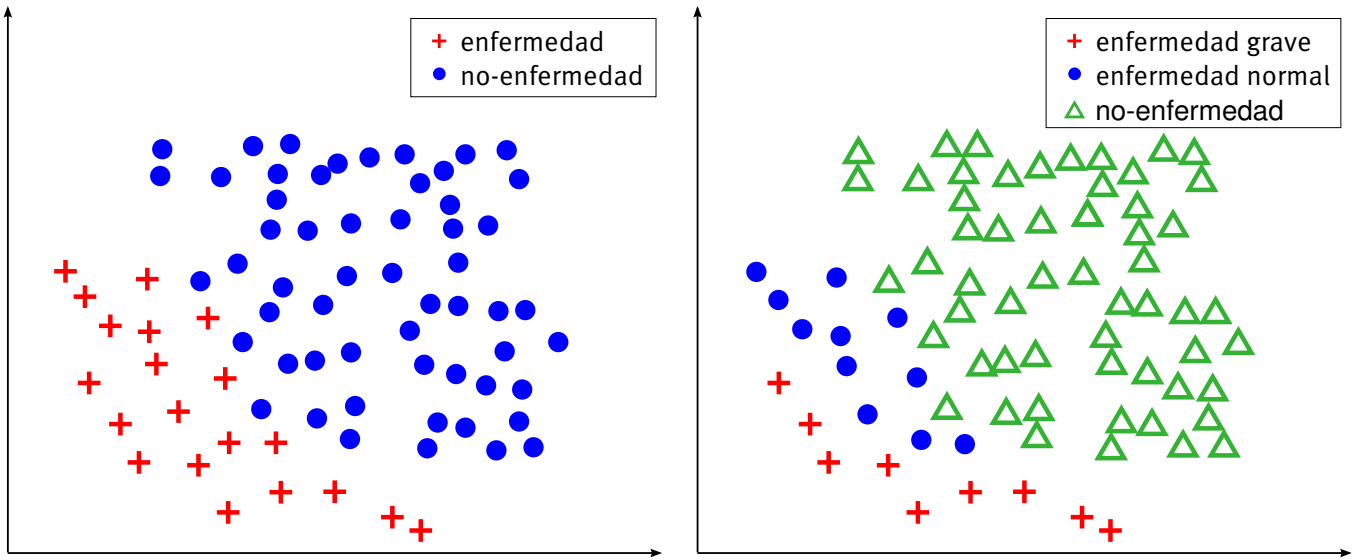


Figura : Un ejemplo de **clasificación binaria** (figura a la izquierda) frente a la **clasificación multi-clase** (figura de la derecha). En el primer caso hay dos estados para un patrón: enfermo o no enfermo. Sin embargo, un experto que se apoye en técnicas de aprendizaje automático puede demandar grados de clasificación más finos, en cuyo caso podría afrontarse el problema como clasificación multi-clase.

Otros tipos de clasificación

Además de la clasificación binaria y multi-clase, existen otros tipos:

- **Clasificación ordinal** (también llamada regresión ordinal).
- Clasificación **multi-etiqueta**.
- Clasificación **semi-supervisada**.

Formulación matemática

- Disponemos de un **espacio de entrada** $\mathcal{X}$ compuesto por patrones etiquetados con $\mathcal{C} = \{C_1, C_2, \ldots, C_Q\}$ donde Q es el número de clases.
- Cada **patrón** se representa un por **vector de características** de dimensión K, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$ y una etiqueta de clase $y \in \mathcal{C}$.
- El objetivo es aprender una función ϕ que relacione los datos del espacio de entrada $\mathcal{X}$ al conjunto finito $\mathcal{C}$.
- El conjunto de patrones de entrenamiento $\mathbf{T}$ está compuesto de N patrones $\mathbf{T} = \{(\mathbf{X}, \mathrm{Y}) = (\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X},\ y_i \in \mathcal{C}\,(i = 1, \ldots, N)\}$, con $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,K})$.

Modelo linear vs no-linear

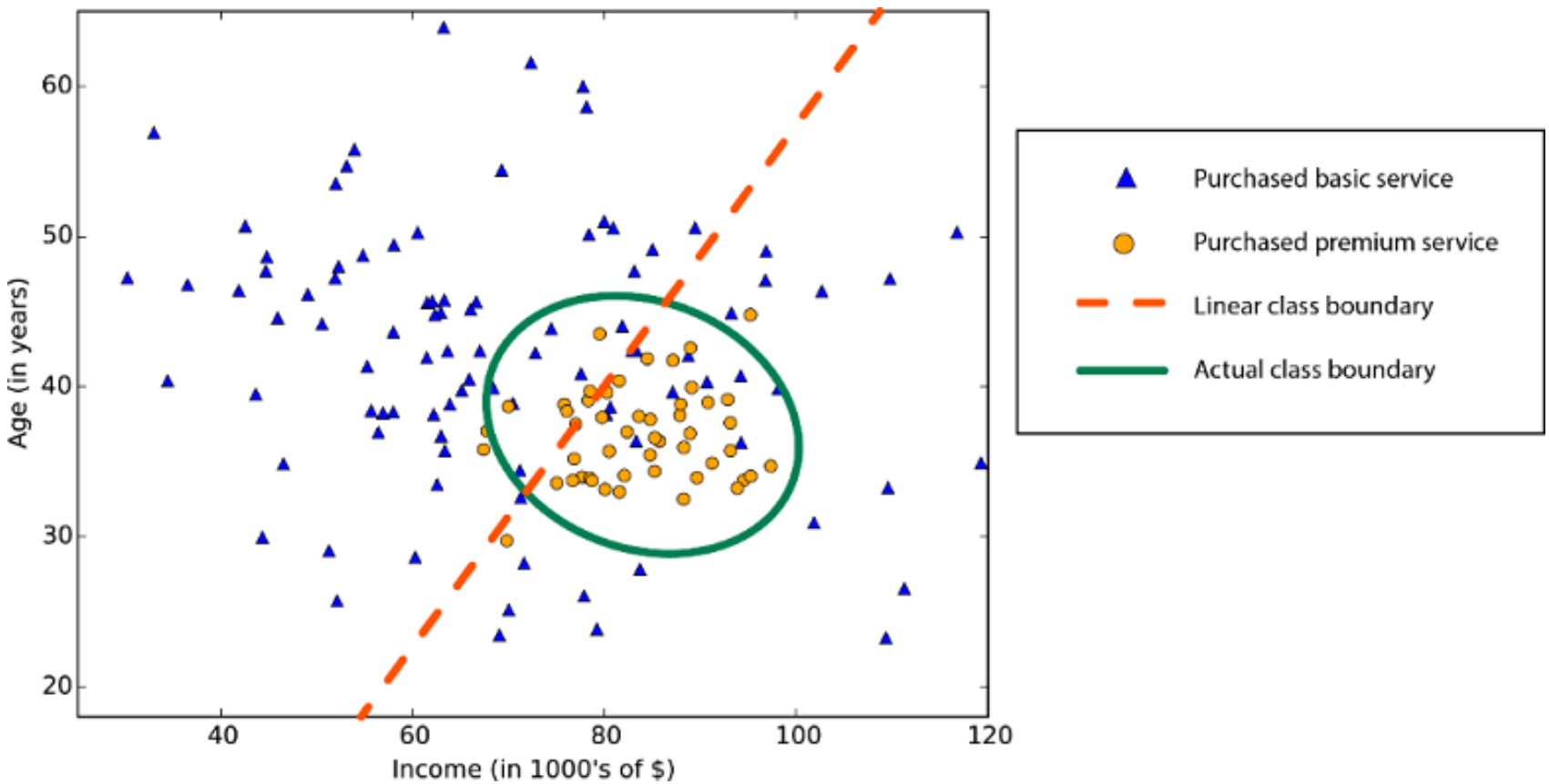


Figura : Fuente How to choose algorithms for Microsoft Azure Machine Learning

Regresión logística

La **regresión logística** es un **método estadístico**, aunque muchos paquetes de aprendizaje automático la incluyen con variantes. En un **modelo lineal** de clasificación binaria

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_K x_{K,i})}} \tag{1}$$

- **Ventajas**: probabilístico, no tiene hiper-parámetros y es interpretable.
- **Desventajas**: modelo lineal.

k-vecinos cercanos

El método **k-vecinos cercanos** o k-NN (K *nearest neighbors*) es uno de los más sencillos. Se basa en estimar la probabilidad de pertenencia de un patrón $\mathbf{x}$ a una clase $F(\mathbf{x}/C_j)$ en base a los k patrones más cercanos que le rodean.

- **Ventajas**: probabilístico, no lineal, tiene variantes que mejoran el rendimiento y robustez a patrones anómalos.
- **Desventajas**: altamente dependiente de k, no interpretable.

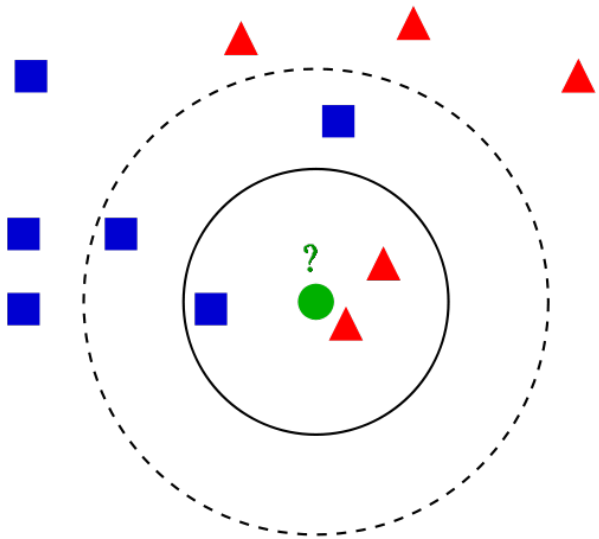
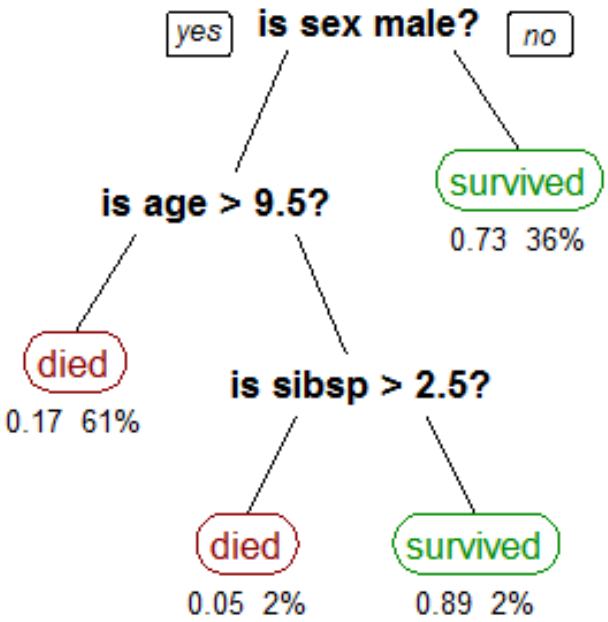


Figura : Fuente https://commons.wikimedia.org/wiki/File:KnnClassification.svg

Árboles de decisión I

El algoritmo de aprendizaje construye un **árbol de decisión** donde las hojas son las categorías y las ramas representan valores de características o conjuntos de características. C4.5 es la implementación más extendida. Ej. probabilidad de sobrevivir al accidente del Titanic.



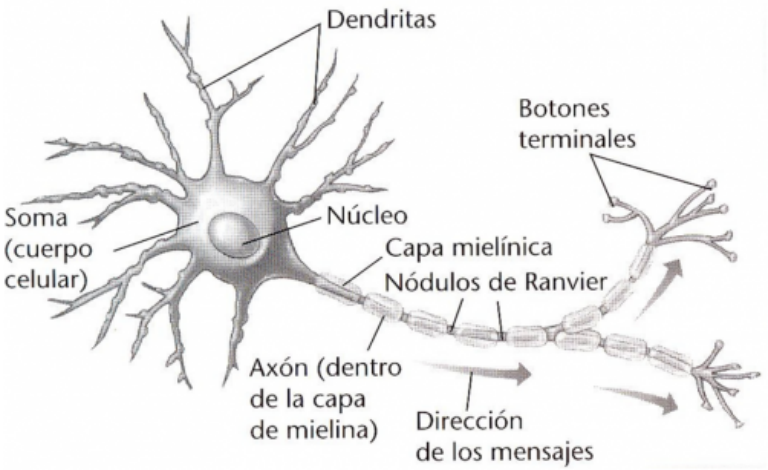
Fuente
https://commons.wikimedia.org/wiki/File:CART_tree_titanic_survivors.png

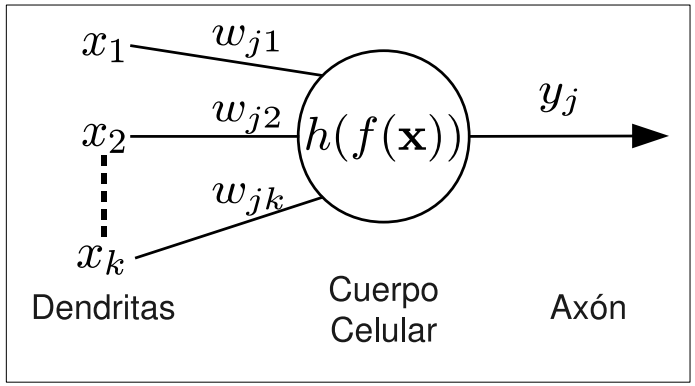
Árboles de decisión II

- **Ventajas**: probabilístico, no lineal, interpretable, no sensible a hiper-parámetros
- **Desventajas**: puede sobreentrenar, pueden ser muy complejos y perder Interpretabilidad.

Concepto de Red Neuronal Artificial

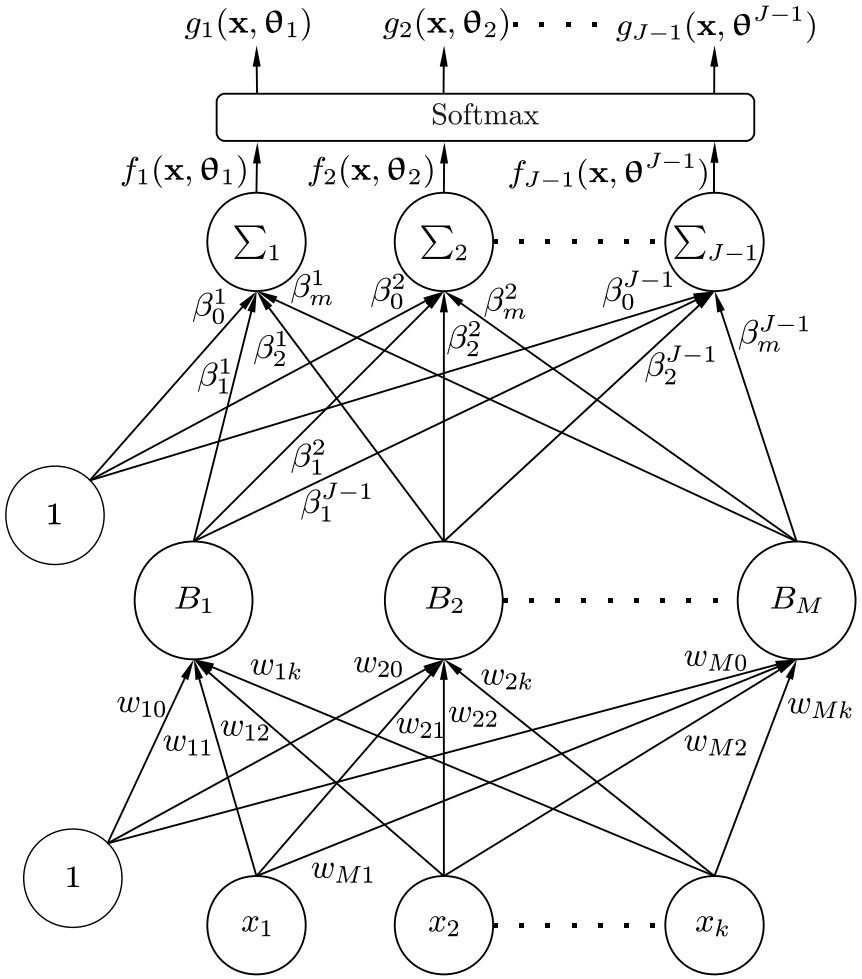
- Técnica de modelado fundamentada en la **emulación de los sistemas nerviosos biológicos**.





- Combina una gran cantidad de elementos simples de procesado (**neuronas**), altamente interconectados y agrupados en capas.
- Una red neuronal es una relación funcional matemática entre unas variables de entrada y unas variables de salida.

Modelo de RNA para clasificación



Para J clases,

$$f_j(\mathbf{x}, \theta_j) = \beta_0^j + \sum_{i=1}^{M} \beta_i^j B_i(\mathbf{x}, \mathbf{w}_i)$$

$$\text{para } 1 \leq j \leq J-1$$

$$g_j(\mathbf{x}, \theta_j) = \frac{e^{f_j(\mathbf{x}, \theta_j)}}{\sum_{i=1}^{J} e^{f_i(\mathbf{x}, \theta_i)}}$$

Características de las RNA I

- Las RNA son **aproximadores universales**, es decir, pueden aprender cualquier función matemática aumentando su capacidad, que depende de la complejidad (número de capas y número de conexiones).
- Hay **muchos tipos de arquitecturas** de RNA, aquí sólo hemos mostrado las redes *single layer feedforward*.
- Otros tipos de RNA: redes recurrentes, redes convolucionales, etc.
- Aunque nunca han perdido uso, recientemente vuelven a la actualidad científica sobre todo en el campo del **aprendizaje profundo** (*deep learning*). Por ejemplo, el motor de búsqueda de imágenes de Google funciona con RNA para aprender conceptos etiquetados en imágenes[2].

Características de las RNA II

- **Ventajas**: probabilístico, no lineal, aproximador universal, versatilidad (reconocimiento del habla, textos...)
- **Desventajas**: a menudo sobreentrenan, el coste de entrenar RNA es alto con los métodos más populares, no son interpretables

[2] Inceptionism: Going Deeper into Neural Networks

Máquinas de vector soporte

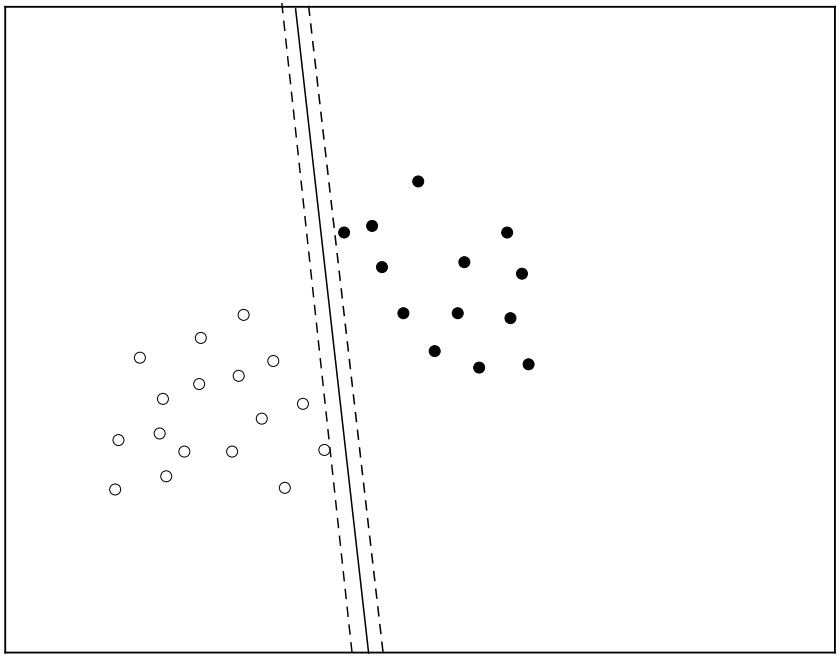
La idea básica de las **máquinas de vector soporte** (SVM, *support vector machines*) es simple: encontrar el **hiper-plano que define la máxima separación entre patrones de dos clases diferentes**:

$$f(\mathbf{x}) = \hat{y} = \operatorname{sgn}\left(\langle \mathbf{w} \cdot \phi(\mathbf{x}) \rangle + b\right),$$

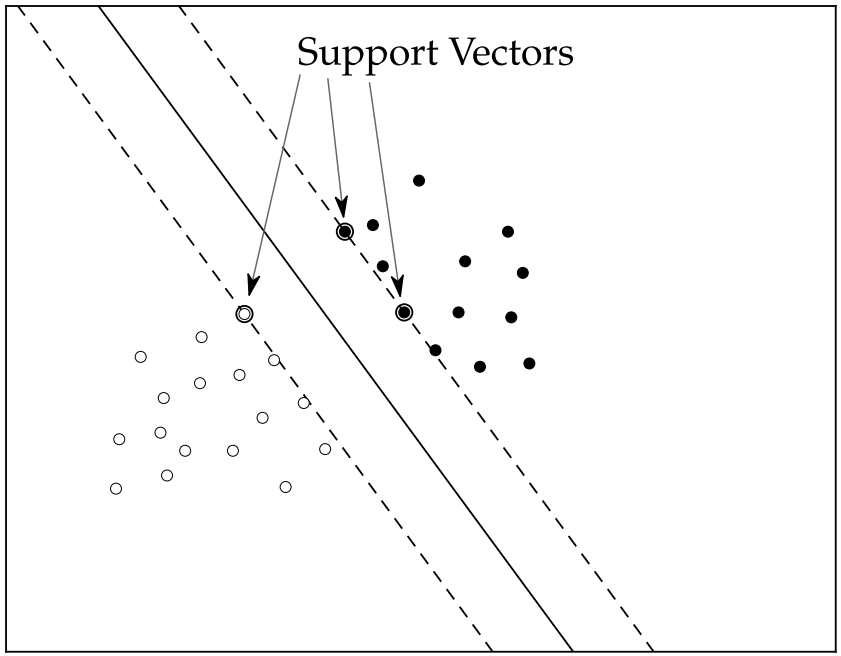
donde $\hat{y} = +1$ si $\mathbf{x}$ corresponde a la clase positiva y $\hat{y} = -1$ en otro caso.

El modelo correspondiente para el caso no lineal y de margen blando es más complejo, y comprender el proceso de cálculo del hiper-plano óptimo requiere de habilidades matemáticas en optimización.

Máximo hiper-plano separador

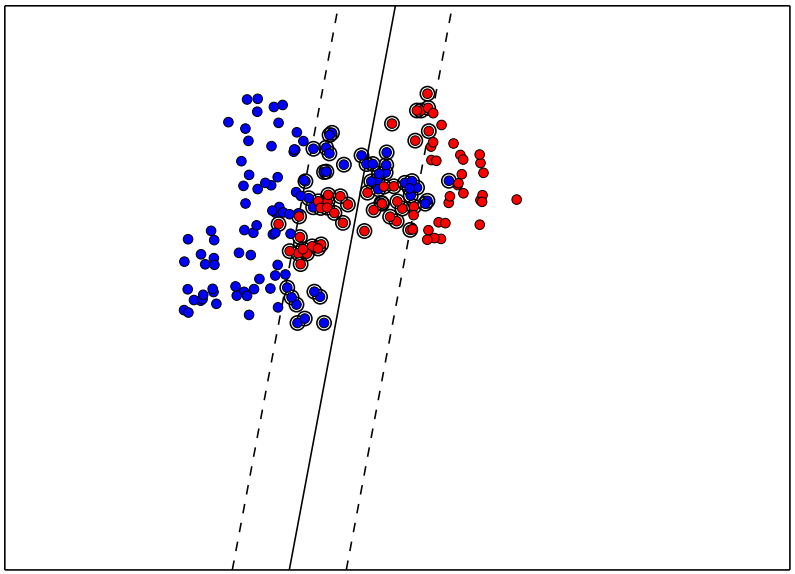


Small margin

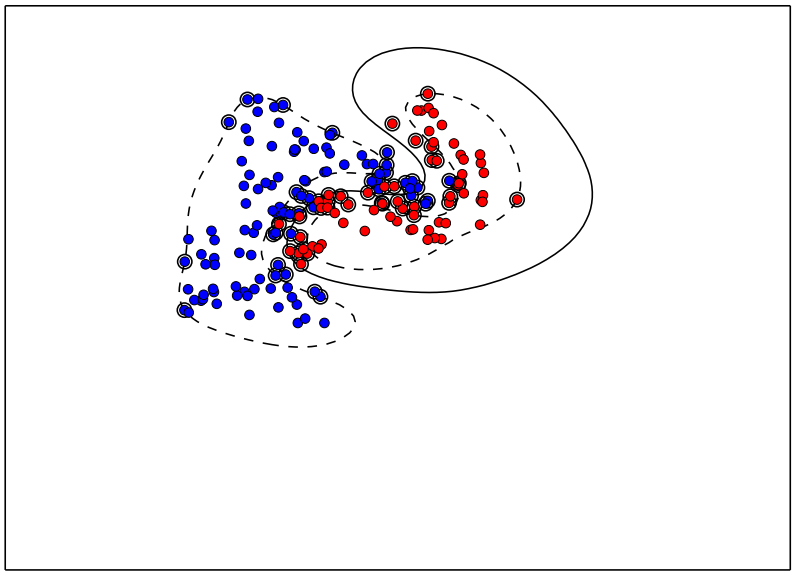


Large margin

Ejemplo de SVM lineal y no lineal



Linear: $u^T v$ RBF: $\exp(-\gamma|u-v|^2)$ Poly: $(\gamma\, u^T v + r)^d$



Linear: $u^T v$ RBF: $\exp(-\gamma|u-v|^2)$ Poly: $(\gamma\, u^T v + r)^d$

Características de los SVM I

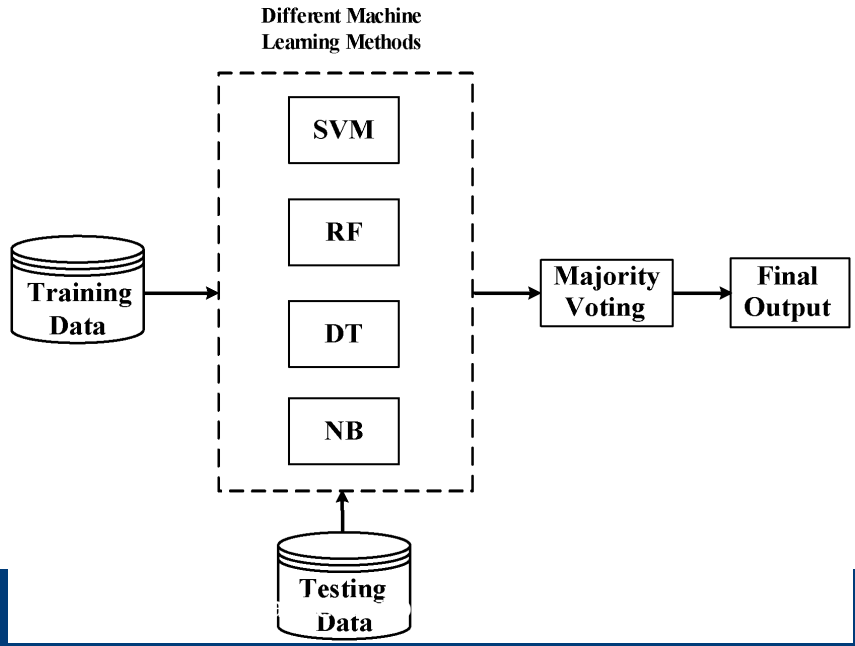
- Para **pocos patrones** ($< 2{,}000 - 10{,}000$ dependiendo de la dimensionalidad) los modelos no lineales son muy potentes, previo ajuste de los hiper-parámetros de coste y ancho del kernel gausiano.
- Para bases de datos mayores, el modelo lineal es bastante competitivo (implementación liblinear).
- **Ventajas**: no lineal, más robusto al problema de la alta dimensionalidad, *sparsity,* etc.
- **Desventajas**: a menudo sobre-entrenan si no se ajustan bien los hiper-parámetros, no son interpretables, ineficientes para entrenar grandes volúmenes[3]

[3] Si bien hay propuestas para la resolución distribuida del problema de optimización

Ensembles de clasificadores I

Ensemble

En aprendizaje automático definimos un ensemble como un **conjunto de modelos** diferentes entre si que unen sus predicciones con el propósito de hacer una predicción más robusta. Se basan en la premisa de que cuantos más elementos decidan más robusta será la decisión.



Ensembles de clasificadores II

Resumen de los ensembles:

- Se considera que un ensemble aporta más que un modelo único en base a la **diversidad de modelos**. La diversidad se consigue:
  - Utilizando diferentes modelos (árboles, RNA, SVM...)
  - Favoreciendo la selección de modelos diferentes en el entrenamiento
  - Proporcionando distintos subconjuntos de entrenamiento a los modelos
  - ...
- **Esquema de decisión**. Por ejemplo: votación por mayoría, votación al modelo más fiable (máxima probabilidad de pertenencia a clase)...

Ensembles de clasificadores III

Rendimiento

Tanto el **tiempo de entrenamiento** como el **tiempo de test** (lo que tarda el modelo de ensemble en evaluar un patrón) son obviamente mayores que en modelos únicos.
Dependerá de los requisitos de nuestro problema elegir este tipo de técnicas. Ejemplos de escenarios:

- En la asignación de un trasplante de hígado consideramos que el emplear unos milisegundos más en decidir es irrelevante a cambio de hacer una operación con éxito.
- Un vehículo inteligente debe frenar cuanto antes al detectar a un peatón en la trayectoria.
- En la bolsa, en un segundo se realizan 3 millones de operaciones de compra/venta de valores.

Extensión de clasificación binaria

- Algunos de los métodos presentados tienen **naturaleza multiclase** (ej. RNA y árboles de decisión).
- Otros modelos como los SVM o la regresión logística **sólo pueden aplicarse a la clasificación binaria**.
- Sin embargo, **utilizando ensambles de modelos binarios se pueden implementar modelos multi-clase**. Las dos opciones más conocidas son **una-vs-una** y **una-vs-todas**.

Esquema una-vs-todas

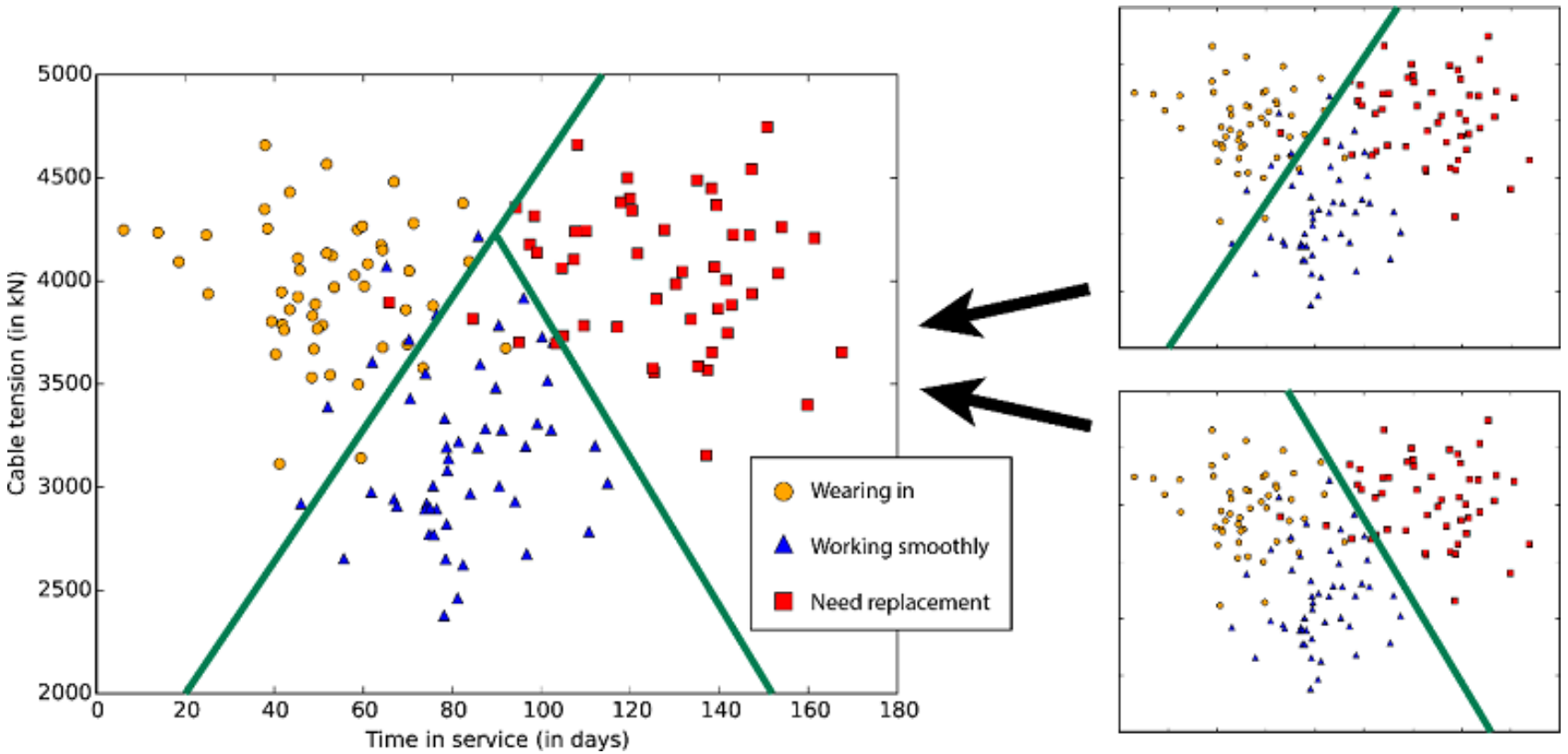
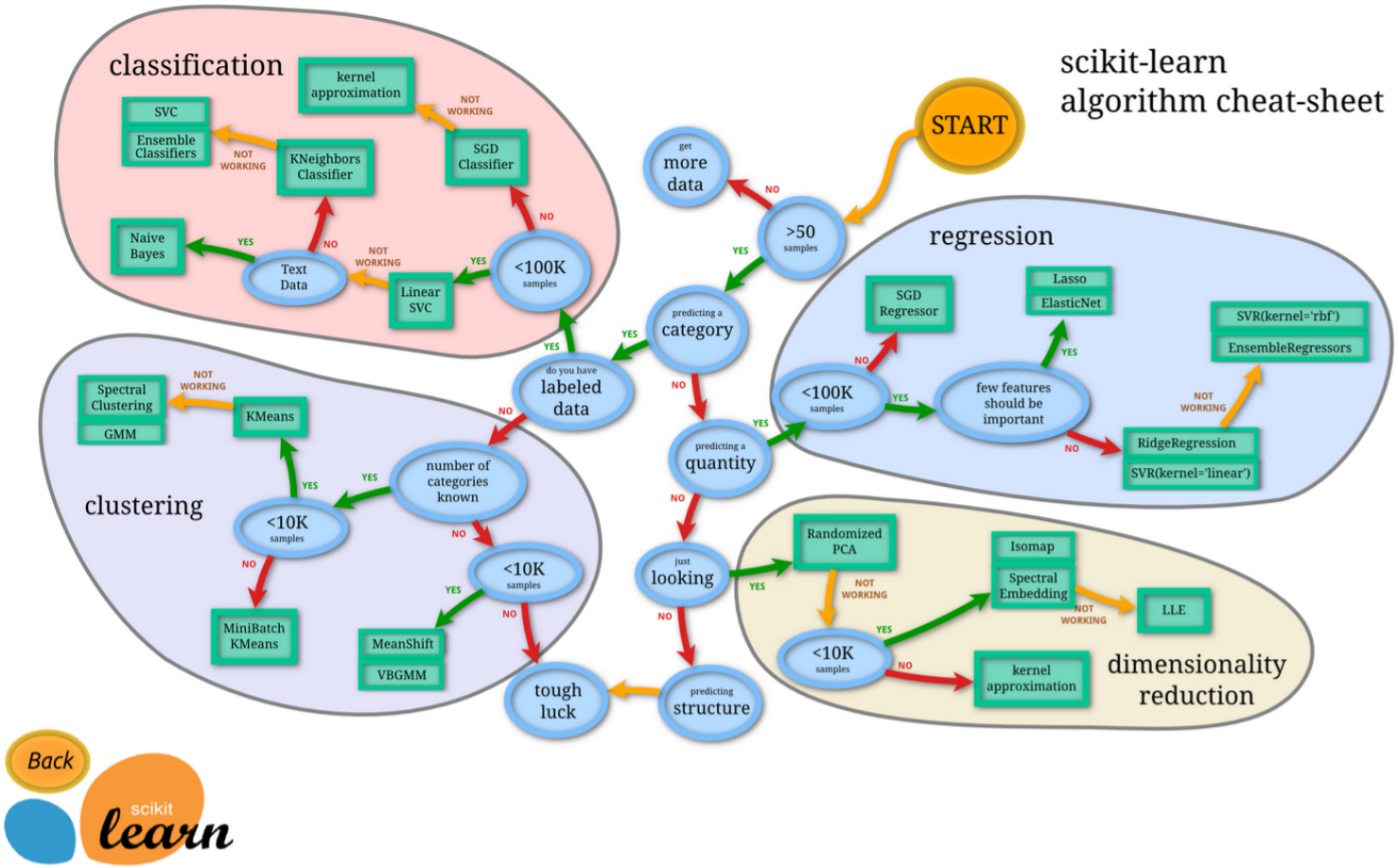


Figura : Esquema una-vs-todas (*one-vs-all*) fuente How to choose algorithms for Microsoft Azure Machine Learning

Visión general de modelos de scikit-learn



scikit-learn.org: Choosing the right estimator

Evaluación del rendimiento de modelos

- Hay un amplio **abanico de métricas** de rendimiento de modelos de clasificación.
- Las métricas de rendimiento son importantes:
  - Para la **evaluación y comparación** de modelos.
  - Para **guiar** los algoritmos de aprendizaje y de optimización de hiper-parámetros, esto es muy relevante para modelos como RNA o SVM.
- De nuevo, elegir la métrica de evaluación **depende** completamente **del problema** a resolver.

Motivación para la diversidad de métricas

Dependiendo del problema, las métricas pueden darnos una mala visión del rendimiento del modelo.

Problema de medicina

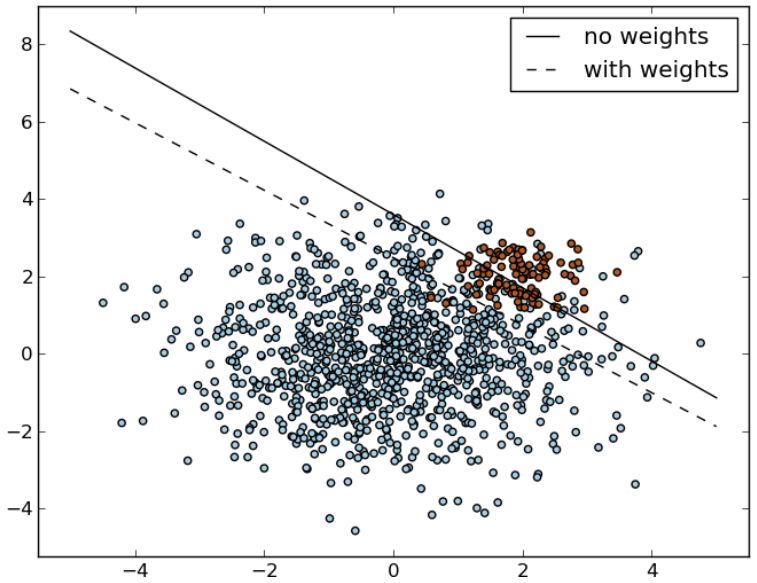
Tenemos un modelo que diagnostica con un 99 % de precisión a todos los pacientes que acuden a hacer un test de cáncer.

Distribución de los patrones

El modelo clasifica a todos los pacientes como sanos, sean como sean. El 99 % de los pacientes no presentaba cáncer y los clasificó bien, el 1 % tenía cáncer y los clasificó como sanos.

Clasificación desbalanceada

- **Clasificación desbalanceada** se refiere a conjuntos de datos donde el número de patrones perteneciente a una clase varia notablemente.
- Los clasificadores tienen a **ignorar a las clases minoritarias**
  - Habitualmente, estas son las **clases más interesantes** (ej. detección de una enfermedad)
  - Es un campo muy activo desde hace años en los entornos binarios y más recientemente en entornos multi-clase



El problema del desbalanceo depende del ruido y solapamiento entre clases

Matriz de confusión

Cuadro : Matriz de contingencia o **matriz de confusión** que refleja la relación entre falsos positivos, verdaderos positivos, verdaderos negativos, y falsos negativos.

	actual class (observation)	
predicted class (expectation)	TP (true positive) Correct result	FP (false positive) Unexpected result
	FN (false negative)	TN (true negative)
	Missing result	Correct absence of result

Métricas clasificación binaria I

Métricas cuando el interés reside en la clase minoritaria (ejemplo diagnóstico de enfermedad):
La *Precisión* es el porcentaje de predicciones correctas:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

Recall (también *S* o ratio de verdaderos positivos):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

F-measure se define como la media armónica de *Precision* y *Recall*:

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{4}$$

Métricas clasificación binaria II

Métricas cuando el interés reside en ambas clases:
La *Sensibilidad* se define como:

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

La *Especificidad* (también valor de predicción positiva, o *True Negative Rate* (TNR)) mide la proporción de patronces negativos correctamente clasificados como tales (ej. la cantidad de personas que no tienen una enfermedad que se clasifican como que no la padecen):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{6}$$

Métricas clasificación binaria III

Finalmente, la **media geométrica** de la Sensibilidad y Especificidad es muy usada tanto en el caso binario como en el multi-clase, representa el balance entre las dos:

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}. \tag{7}$$

Análisis ROC

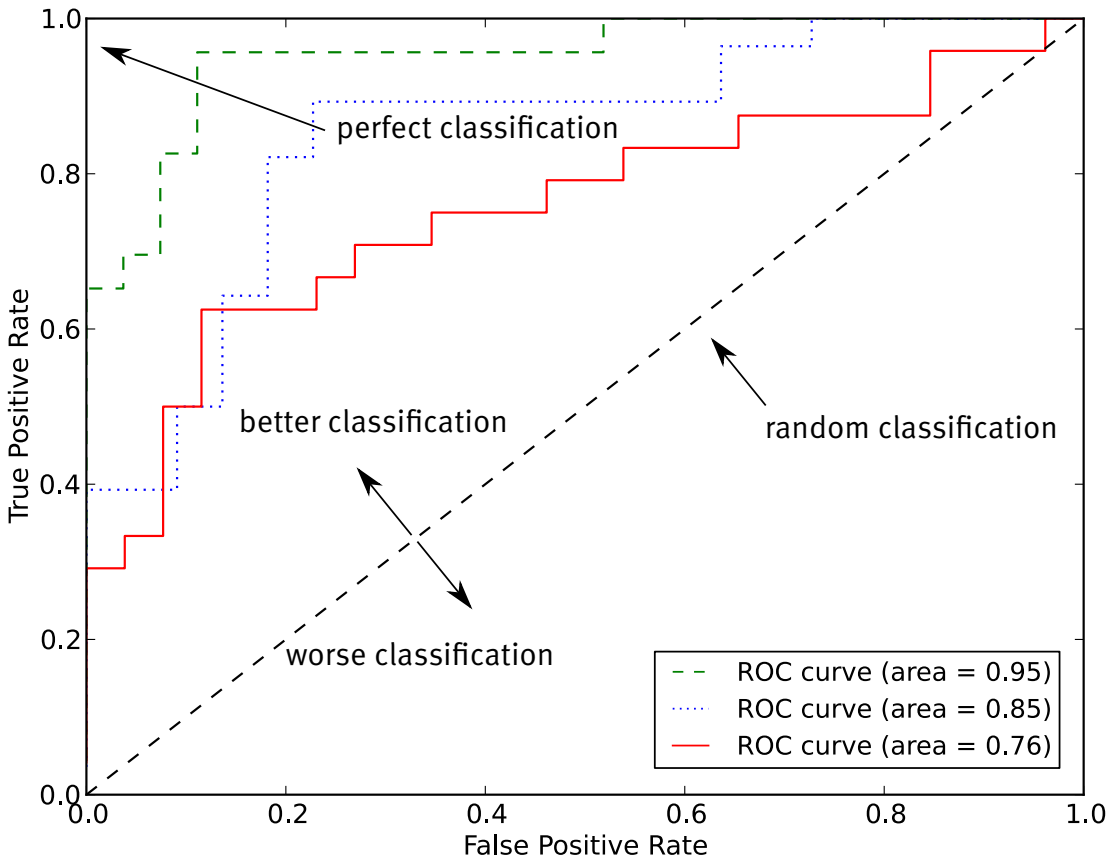
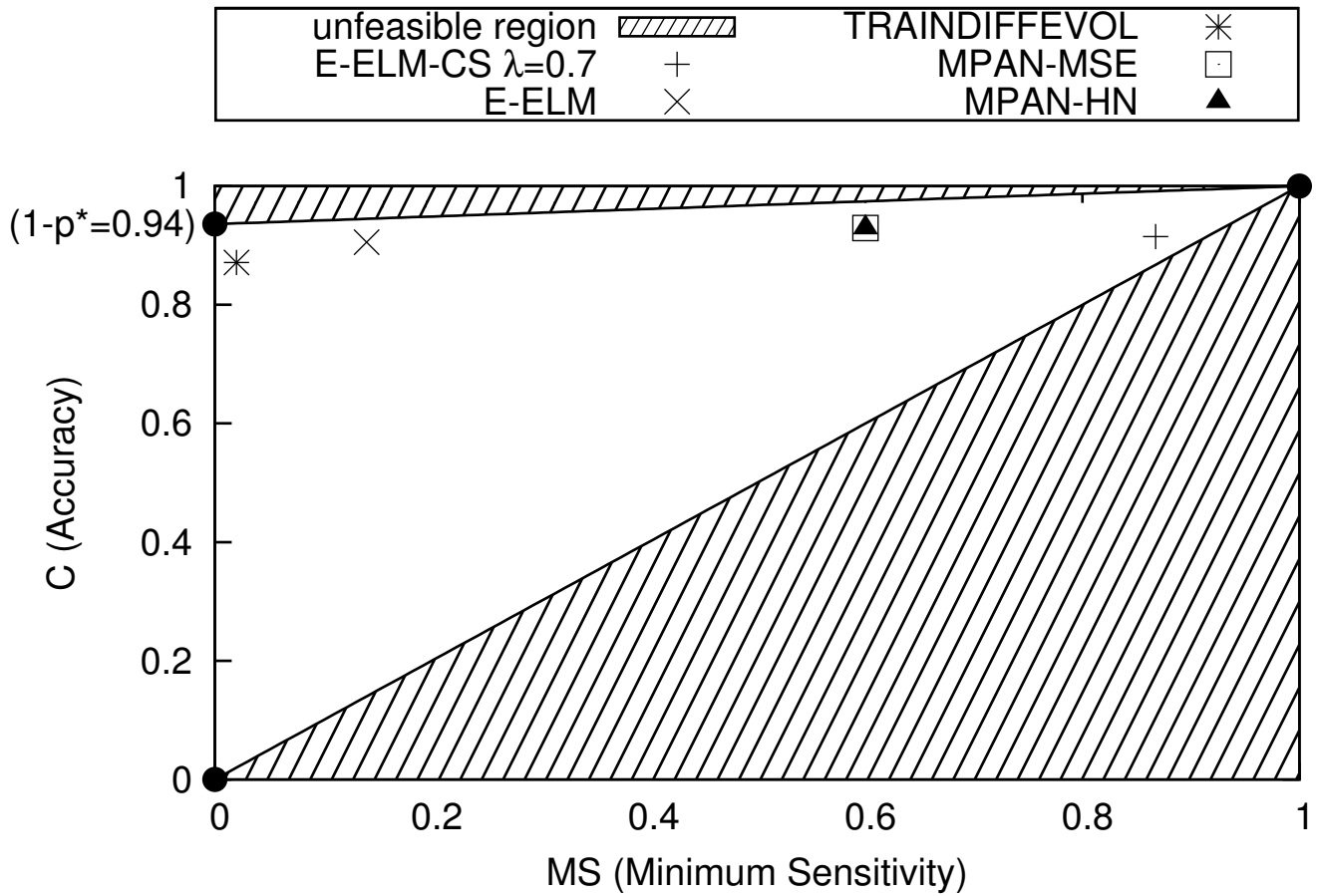


Figura : Comparación de la curva ROC (*Receiver operating characteristic*) de tres clasificadores.

Efecto de métricas en algoritmo evolutivo

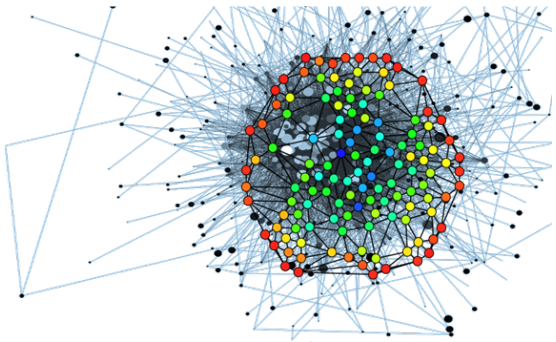


Más información en [2, 3, 4]

Conclusiones sobre evaluación

- La evaluación de modelos debe hacerse con una o más métricas que **visualicen distintos aspectos del rendimiento** de una clasificador.
- Un mismo modelo y algoritmo de aprendizaje, guiado por distintas métricas pueden construir modelos muy diferentes.

Índice



Melanoma thickness classification

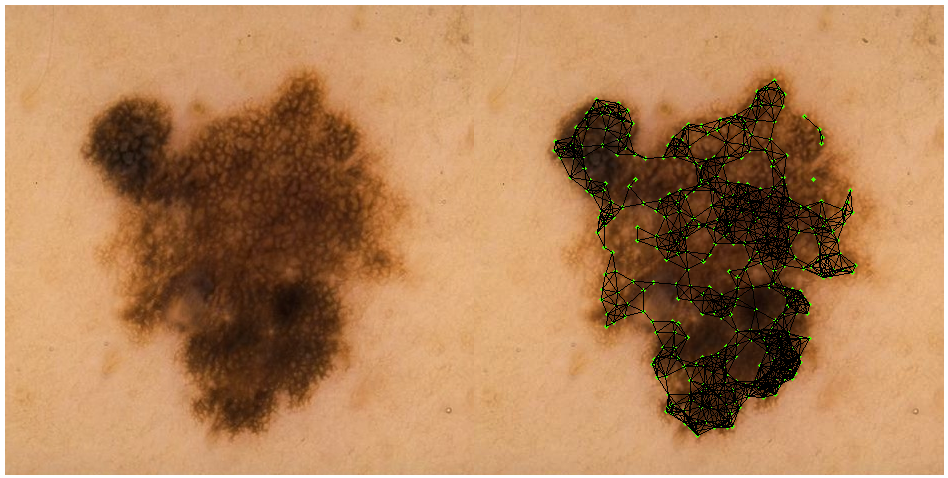
- **Melanoma** is a malignancy of melanocytes, the cells that produce the pigment melanin that colors the skin.
- Melanoma is **the most dangerous skin cancer**.
- **Thickness/Depth** of the melanoma is the most important factor associated with **patients survival**.
- Measured with a **biopsy** of the suspected lesion:
  - Invasive method
  - Inaccurate if the biopsy is not performed in the deepest region.

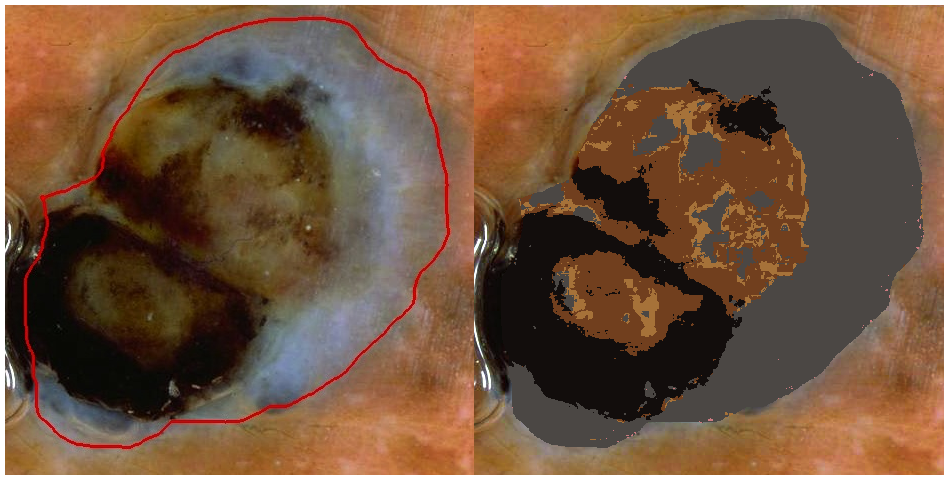
Goal

¿Can we provide a non-invasive alternative or a complement to the biopsy?

- We propose a **computational image analysis system** from dermoscopic images.
- We **extract features** based on clinical findings.
- We use machine learning to build **classification models** which perform finer classification than previous systems.

Features extraction





Classification results

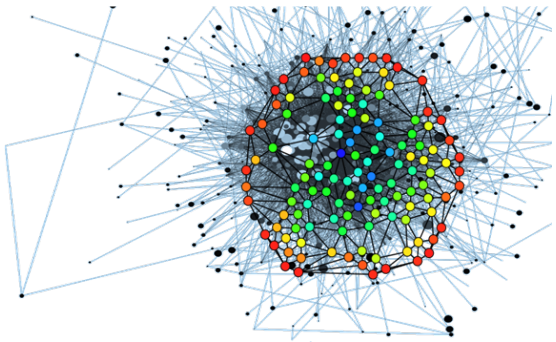
Binary problem			
Method	*Acc*	*MS*	*WAcc*
LIPU	**0.776**	*0.602*	**0.268**
LR	0.752	0.530	0.304
PUNNs	0.720	0.494	0.337
KDL	0.712	**0.663**	0.300
SVC	*0.764*	0.518	*0.298*

Three classes problem			
Method	*Acc*	*MS*	*AMAE*
LIPU	**0.684**	0.185	0.656
LR	0.632	0.069	0.813
PUNNs	0.648	0.148	0.759
KDLOR	0.644	**0.552**	**0.446**
SVC	*0.664*	0.259	0.675
REDSVM	0.624	*0.345*	0.583
SVORIM	0.636	*0.345*	*0.579*

Publication

More information at: A. Sáez, J. Sánchez-Monedero, P. A. Gutiérrez,C. Hervás-Martínez, *Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images*, IEEE Transactions on Medical Imaging, Accepted, 2015.
http://dx.doi.org/10.1109/TMI.2015.2506270

Índice



Conclusiones ciencia de datos

- Introducción a la ciencia de datos, minería de datos y aprendizaje automático.
- Vista general de técnicas de minería de datos.
- Principales modelos de clasificación, aunque faltan modelos, como los grafos probabilísticos.
- Hemos identificado algunas ventajas e inconvenientes de modelos de clasificación.
- Hemos destacado la importancia de una buena selección de métricas de evaluación
- Destacamos la importancia de ajustar propiamente los parámetros de los modelos y algoritmos.

Conclusiones clasificación I

Hemos identificado una serie de objetivos contrapuestos que se cumplen por lo general:

- Simplicidad del modelo vs precisión
- Interpretabilidad del modelo vs precisión
- Escalabilidad del modelo vs precisión
- Velocidad del modelo vs precisión

La **eficiencia** puede condicionar la aplicabilidad del algoritmo:

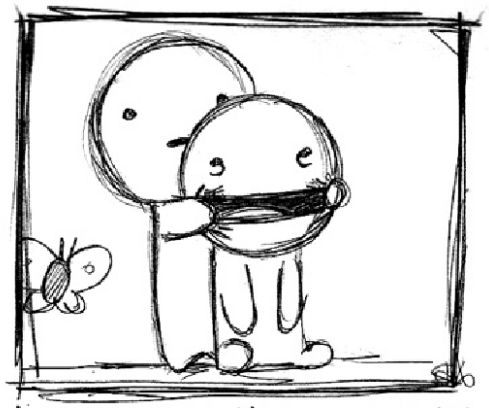
- En 2009 el algoritmo ganador del premio de 1.000.000 $ Netfix, no se implementó nunca debido al coste computacional.

¿Por qué aprender a programar?

- Preprocesamiento relacionado con la naturaleza del problema.
- Conversión de ficheros.
- Creación de nuevos modelos.
- Proporcionar métricas de validación no disponibles en el software utilizado.
- Automatizar experimentos y generación de informes.

¿Preguntas?

¡Gracias!



References I

Francisco Herrera.

Introducción a la ciencia de datos, minería de datos y big data. curso de aproximación práctica a la ciencia de datos y big data: heramientas knime, r, hadoop y mahout, 2014.
URL: http://sci2s.ugr.es/otherPostGraduateCourses/CienciaDatosBigData.

F. J. Martínez-Estudillo, P. A. Gutiérrez, C. Hervás-Martínez, and J. C. Fernández.

Evolutionary learning by a sensitivity-accuracy approach for multi-class problems.
In *Proceedings of the 2008 IEEE Congress on Evolutionary Computation (CEC'08)*, pages 1581–1588, Hong Kong, China, 2008. IEEE Press.

Javier Sánchez-Monedero.

Retos en clasificación ordinal: redes neuronales artificiales y métodos basados en proyecciones. Challenges in ordinal classification: artificial neural net- works and projection-based methods.
PhD thesis, Universidad de Granada, September 2013.
URL: http://www.uco.es/ayrna/publications/thesis/ThesisDissertationJSM.pdf.

J. Sánchez-Monedero, P. A. Gutiérrez, F. Fernández-Navarro, and C. Hervás-Martínez.

Weighting efficient accuracy and minimum sensitivity for evolving multi-class classifiers.
Neural Processing Letters, 34(2):101–116, 2011.

Christopher M. Bishop.

Pattern Recognition and Machine Learning.
Springer, 1st ed. 2006. corr. 2nd printing edition, August 2007.

References II

Christopher M. Bishop.

Neural Networks for Pattern Recognition.
Oxford University Press, USA, 1 edition, January 1996.

M. Tim Jones.

Data science and open source, 2013.
URL: http://www.ibm.com/developerworks/library/os-datascience/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay.
Scikit-learn: Machine learning in Python.
Journal of Machine Learning Research, 12:2825–2830, 2011.

Wikipedia.

Machine learning — wikipedia, the free encyclopedia, 2015.
[Online; accessed 11-December-2015].
URL:
https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=694758013.