APRENDIZAJE DE MÁQUINAS

JOHN W. BRANCH

Profesor Titular
Departamento de Ciencias de la Computación y de la Decisión
Director del Grupo de I+D en Inteligencia Artificial – GIDIA
jwbranch@unal.edu.co

https://github.com/jwbranch/AprendizajeDeMaquinas
https://www.coursera.org/programs/unal-iuukt

Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

METODOLOGÍA ENSEÑANZA – APRENDIZAJE

El **aprendizaje sincrónico** involucra estudios online a través de una plataforma. Este tipo de aprendizaje sólo ocurre en línea. Al estar en línea, el estudiante se mantiene en contacto con el docente y con sus compañeros. Se llama aprendizaje sincrónico porque la plataforma permite que los estudiantes pregunten al docente o compañeros de manera instantánea a través de herramientas como el chat o el video chat.

Sesiones Remotas vía Google.Meet Sincrónicas y Asincrónicas

El **aprendizaje asincrónico** puede ser llevado a cabo online u offline. El aprendizaje asincrónico implica un trabajo de curso proporcionado a través de la plataforma o el correo electrónico para que el estudiante desarrolle, de acuerdo a las orientaciones del docente, de forma independiente. Un beneficio que tiene el aprendizaje asincrónico es que el estudiante puede ir a su propio ritmo.

Descripción del Curso

El curso introduce los conceptos fundamentales y los métodos más utilizados en el campo del aprendizaje de máquinas enfocados desde las perspectivas de la naturaleza del problema que se requiere resolver, esto es, aprendizaje supervisado orientado a los problemas de clasificación y regresión para aplicaciones de predicción o pronóstico. Aprendizaje no supervisado orientado a tareas de agrupar o etiquetar un conjunto de datos, También se incluyen la aproximación general de técnicas modernas de aprendizaje tales como el aprendizaje por refuerzo y aprendizaje profundo.

Contenido

1. Introducción.
2. Los datos en Aprendizaje de Máquinas.
3. Aprendizaje Supervisado.
4. Aprendizaje NO Supervisado.
5. Aprendizaje por Refuerzo.
6. Aprendizaje con Clases Desbalanceadas y Combinación de Modelos.
7. Aplicaciones y Casos de Éxito.

Bibliografía Recomendada

Osvaldo Simeone (2018), "A Brief Introduction to Machine Learning for Engineers", Foundations and TrendsR in Signal Processing: Vol. 12, No. 3-4, pp 200–431. DOI: 10.1561/2000000102.

Goodfellow, I., Bengio, Y. y Courville, A. (2016) Deep Learning, MIT Press.

Murphy, K. (2012). Machine Learning: A Probabilistic Perspective, MIT Press .

Hastie, T., Tibshirani, R. y Friedman, J. (2011). The Elements of Statistical Learning. Springer. (Available for download on the authors' web-page.)

Szepesvári, C. (2010). Algorithms for Reinforcement Learning. Morgan and Claypool.

Haykin, S. (2008). Neural Networks and Learning Machines. Pearson.

Sutton, R. y Barto, A. (1998). Reinforcement Learning: An Introduction. MIT Press.

Mitchell, T. M. (1997). Machine Learning. 1st. McGraw-Hill Higher Education. (Chapter 1)

EVALUACIÓN

Certificado Coursera *Sesenta años de inteligencia artificial – UNAM (Obligatorio)* *IA para todos – Andrew Ng (Obligatorio)* *Structuring Machine Learning Projects – Andrew Ng (Obligatorio)* Machine Learning - University of Washington-→Curso #1: Machine Learning Foundations: A Case Study Approach (Obligatorio)	**20%** **(Máx. 31 de Dic/2020)**
Informe de Lectura (Individual) Machine Learning Algorithms: A Review Machine Learning aspects and its Applications Towards Different Research Areas	**20%** **(Máx. 21 de Nov/2020)**
Trabajo Final (Debe ser en Grupo 3 ó 5 personas) *Obtener el conjunto de datos (texto o audio o video o imagen) de los siguientes repositorios o cualquier otro disponible:* *http://www.ics.uci.edu/~mlearn/databases/* *https://www.kaggle.com/datasets* *Origen, atributos, clases* *"Scatter plot" de los datos* *Visualización del conjunto en 2D (PCA o MDS)* *Seleccionar un método de Entrenamiento y Evaluar el Desempeño.*	**60%** **(Máx. 21 de Dic/2020)**

APRENDIZAJE DE MÁQUINAS

" Aprendizaje No Supervisado"

JOHN W. BRANCH

Profesor Titular
Departamento de Ciencias de la Computación y de la Decisión
Director del Grupo de I+D en Inteligencia Artificial – GIDIA
jwbranch@unal.edu.co

https://github.com/jwbranch/AprendizajeDeMaquinas
https://www.coursera.org/programs/unal-iuukt

Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

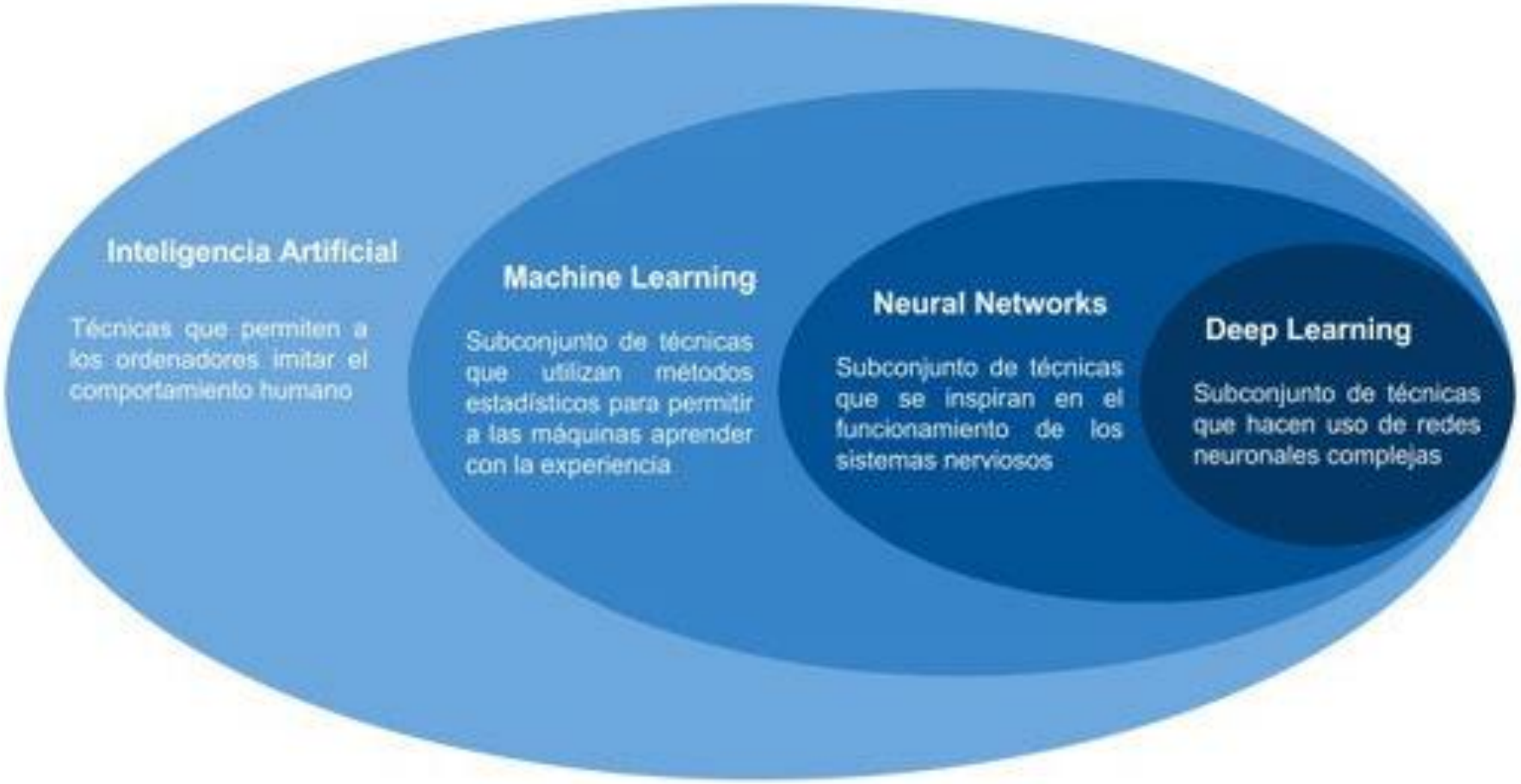
Cultura y Calidad del Dato

Un estudio realizado por ***Harvard Business Review***, menos del 50% de las decisiones estructuradas de las empresas se basan en datos. Este resultado refleja la importancia de implementar una nueva cultura de datos y, sobre todo, llevar los procesos como corresponde.

"...las organizaciones basadas en datos tienen 23 veces más probabilidades de adquirir clientes, seis veces más probabilidades de retener a esos clientes y 19 veces más probabilidades de ser rentables."

McKinsey Global Institute

Aprendizaje vs Inteligencia

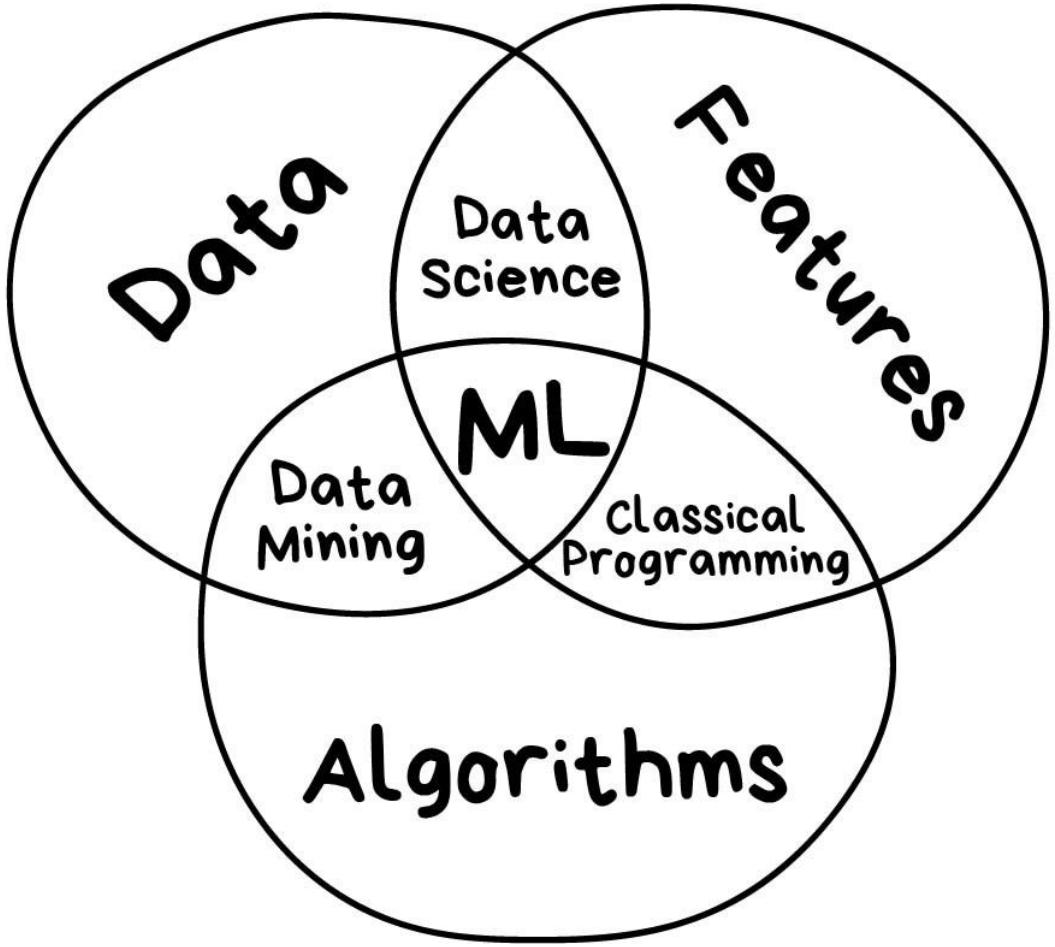


Las máquinas pueden:
Hacer **Pronósticos**
Memorizar
Reproducir
Elegir el mejor artículo

Gidia Grupo de I+D en Inteligencia Artificial

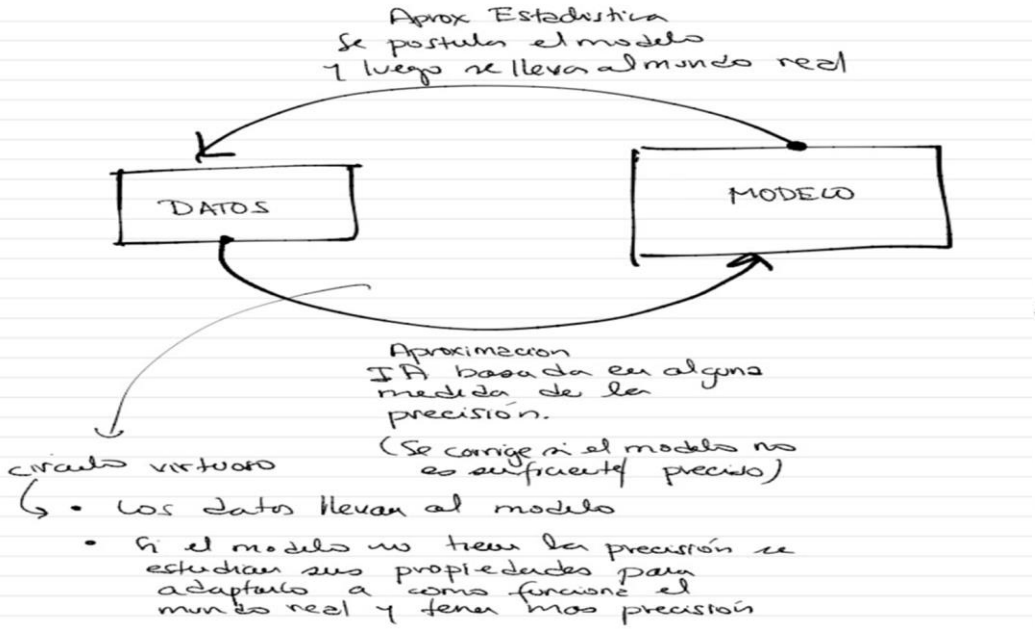
UNIVERSIDAD NACIONAL DE COLOMBIA

Lo tres componentes del aprendizaje de máquinas



Tipos de Aprendizaje de Máquinas

No existe una única forma de resolver un problema en el mundo del aprendizaje de máquinas. Siempre hay varios algoritmos que se ajustan, y la habilidad del científico de datos está en elegir cuál se adapta mejor.



Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

APRENDIZAJE AUTOMÁTICO CLÁSICO

los datos están clasificados en números o categorias

SUPERVISADO

predicen una categoría

CLASIFICACIÓN

clasifica pacientes por tipo (grupos de edad, sexo, otros)
Categorías discretas

predicen un número

REGRESIÓN

dividen pacientes por estatura o IMC
Categorías numéricas

los datos no estan clasificados de ninguna forma

NO SUPERVISADO

se agrupan por similitudes

AGRUPAMIENTO

dividen pacientes por síndromes

se identifican dependencias ocultas

REDUCCIÓN DE DIMENSIONES

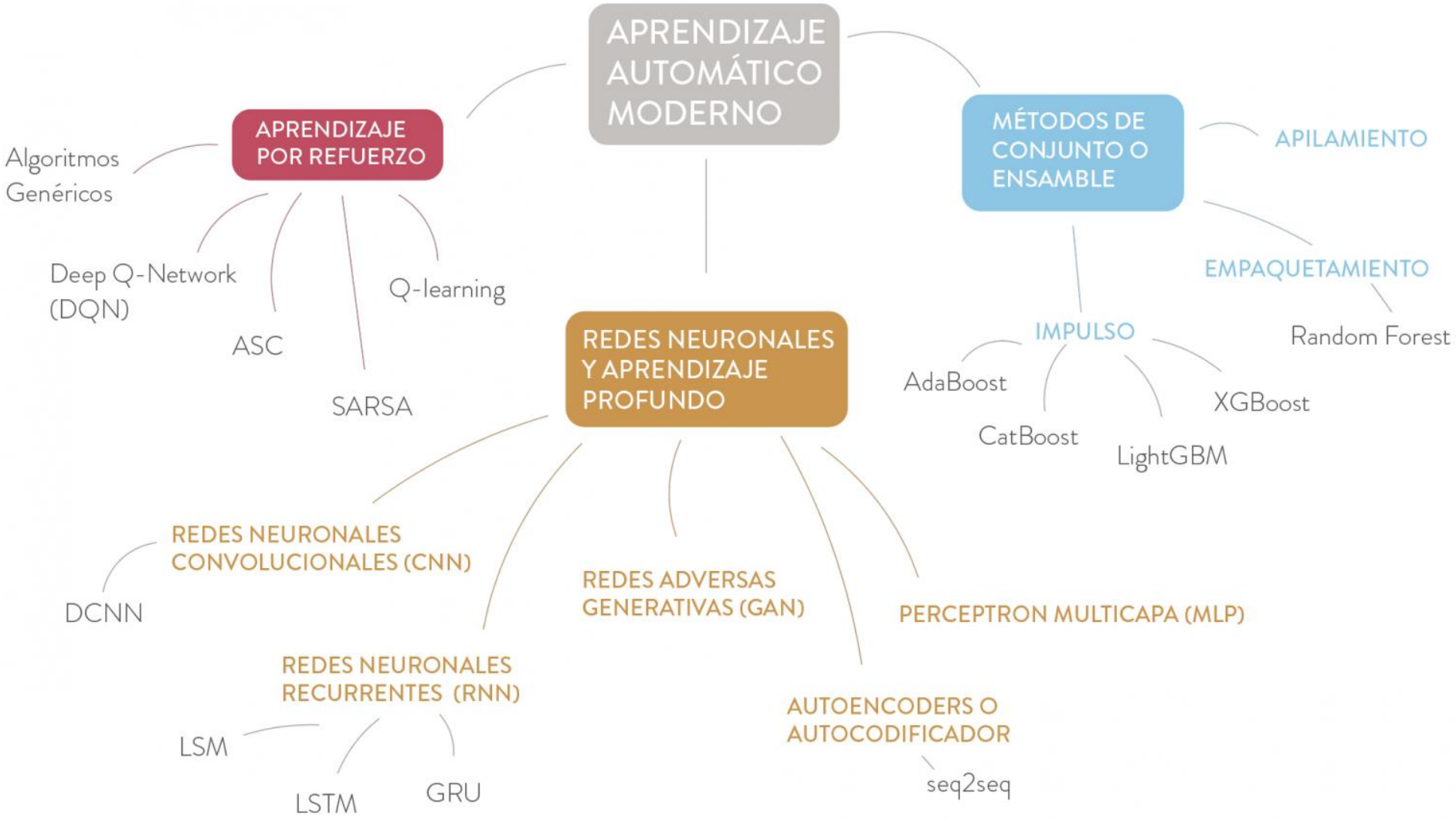
aprovecha las característi-cas para crear mejores modelos

se identifican secuencias

ASOCIACIÓN

agrupa pacientes de modos más convenientes

Tomada de: https://vas3k.com/blog/machine_learning/



Aprendizaje de Máquinas y Predicción

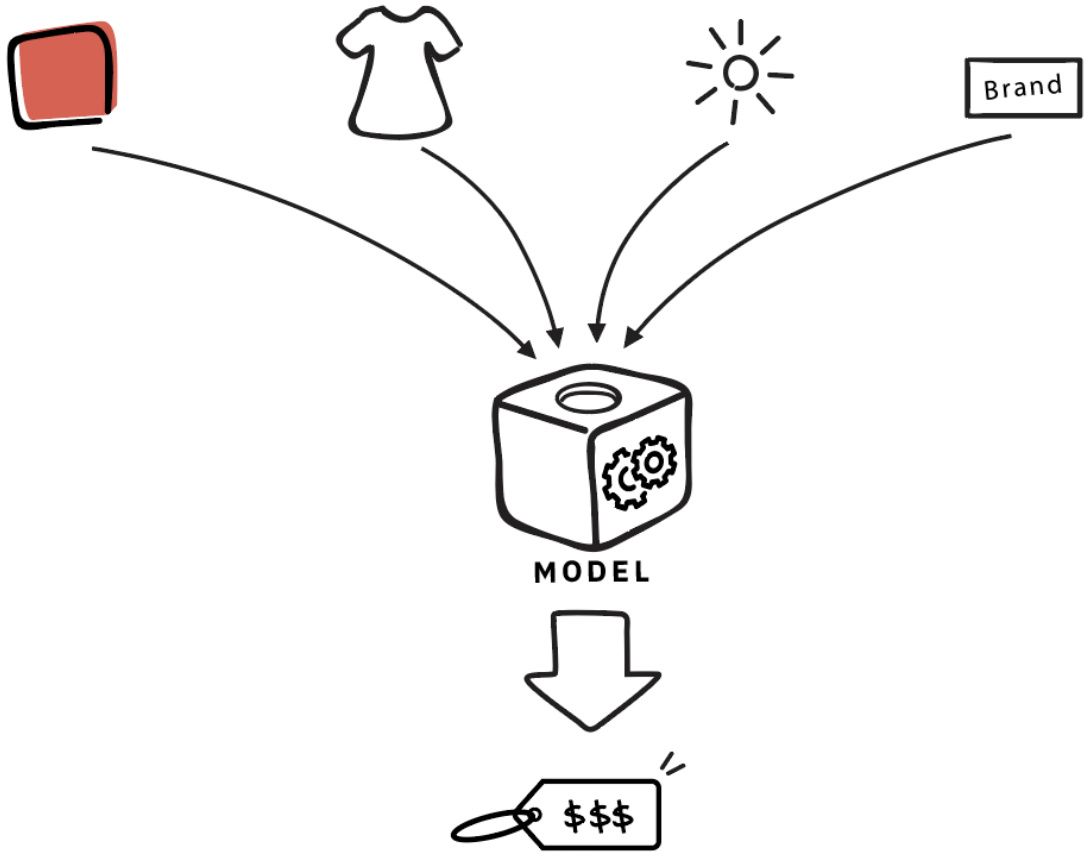
El problema central del aprendizaje de máquinas es la ***predicción***, es decir, aplicar sobre datos nuevos un algoritmo que ha sido entrenado sobre un conjunto de datos históricos.

Aunque suene como predecir el futuro, el término *predicción* generalmente se usa para el procesamiento de datos nuevos *in-situ*. Cuando los datos tienen un componente temporal se utiliza el término *pronóstico*.

En este orden de ideas, cuando se habla de predicción se puede hacer referencia a:

- ***Clasificación*** para obtener una etiqueta o clase conocida.
- ***Regresión*** para obtener un valor numérico.

- ***Agrupamiento*** para descubrir etiquetas o patrones nuevos a partir de datos no etiquetados a partir de medidas de asociación.



Tomada de: https://medium.com/@srnghn

Gidia Grupo de I+D en Inteligencia Artificial

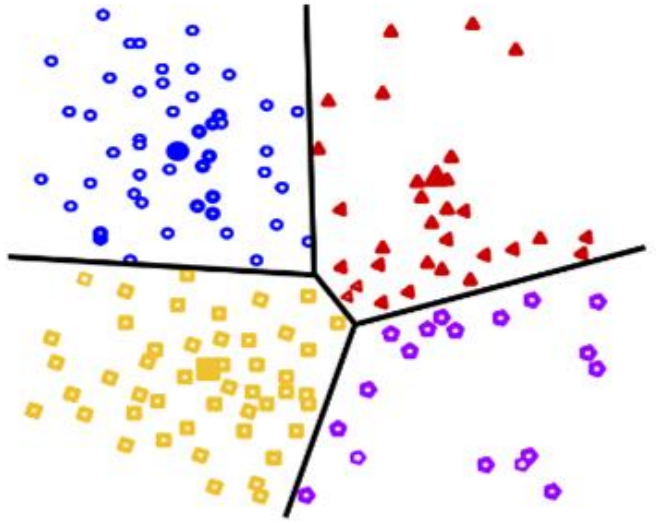
UNIVERSIDAD NACIONAL DE COLOMBIA

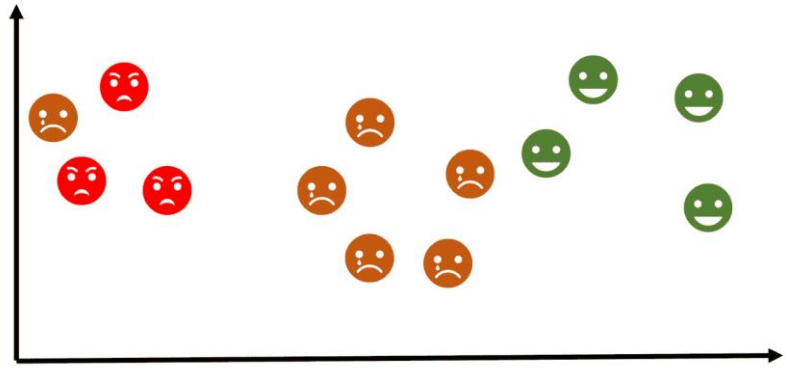
Contenido

1. Introducción al problema de agrupamiento (clustering)
2. Métodos de Agrupamiento
- k-means.
- Agrupación jerárquica.
- Bisecting k-means.
- K-modes.
- T-SNE.
- Autoencoders.
3. Ejemplos.

Aprendizaje No Supervisado:

Se cuenta con un conjunto de datos de entrenamiento, pero no hay una variable específica de salida (se desconocen las clases). En este sentido, el objetivo de los problemas del aprendizaje no supervisado es, agrupar los datos de entrada con base en algún criterio de similitud o disimilitud o determinar la distribución estadística de los datos, conocida como estimación de la densidad.





Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

Aplicaciones

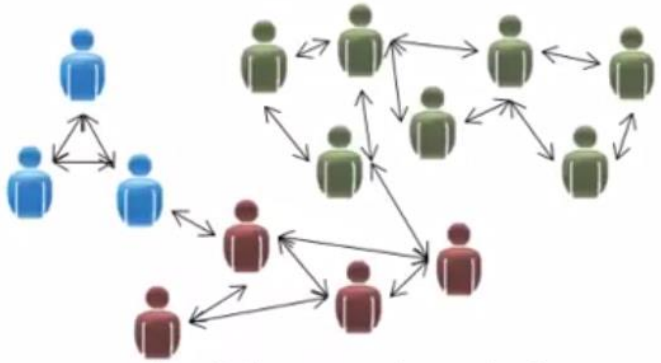
- Biología, para agrupamiento genético y de especies.
- Imágenes médicas, para distinguir entre diferentes tipos de tejido.
- Estudio de mercado, para comprender los diferentes grupos de empresas y clientes en función de algunos atributos.
- Sistemas de recomendación, como darle mejores sugerencias de Amazon

https://towardsdatascience.com/an-easy-introduction-to-unsupervised-learning-with-4-basic-techniques-da7fbf0c3adf

Aprendizaje No Supervisado:

Applications of clustering

Market segmentation



Social network analysis

Organize computing clusters

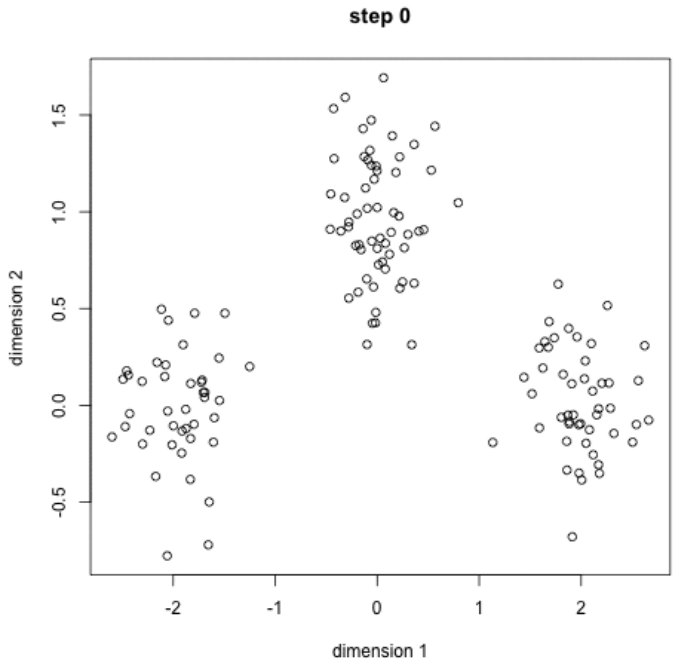
Astronomical data analysis

Andrew Ng

Agrupamiento (Clustering)

El agrupamiento es una técnica que implica la agrupación de puntos de datos. Dado un conjunto de puntos de datos, podemos usar un algoritmo de agrupamiento para clasificar cada punto en un grupo específico.

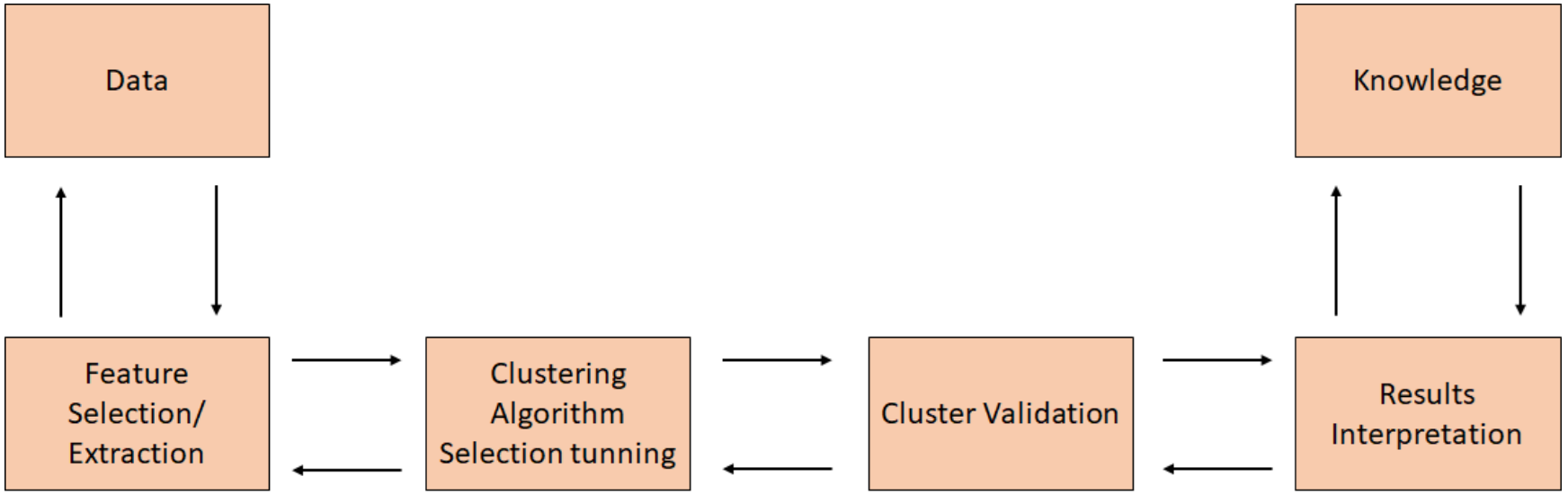
En teoría, los puntos de datos que están en el mismo grupo deberían tener propiedades y / o características similares, mientras que aquellos en diferentes grupos deberían tener propiedades y / o características muy diferentes. La similitud entre puntos generalmente se cuantifica mediante una métrica de distancia basada en algún tipo de conjunto de variables de características.



https://towardsdatascience.com/an-easy-introduction-to-unsupervised-learning-with-4-basic-techniques-da7fbf0c3adf

Proceso de análisis de aprendizaje no supervisado

El proceso general que seguiremos al desarrollar un modelo de aprendizaje no supervisado se puede resumir en el siguiente cuadro:



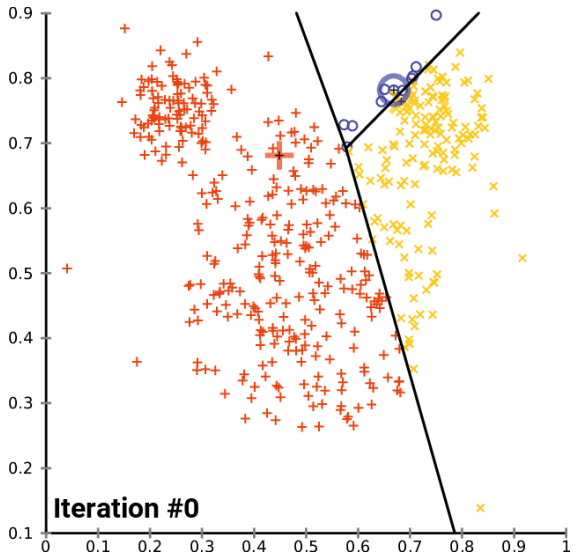
https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e

K-means

El algoritmo de agrupamiento más común y simple que existe es el agrupamiento de K-Means. Este algoritmo implica que le diga cuántos clústeres (o K) posibles hay en el conjunto de datos. Luego, el algoritmo mueve iterativamente los k-centros y selecciona los puntos de datos más cercanos al centroide en el grupo.

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$



https://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51

Algoritmo K-means

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set

   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Donde:

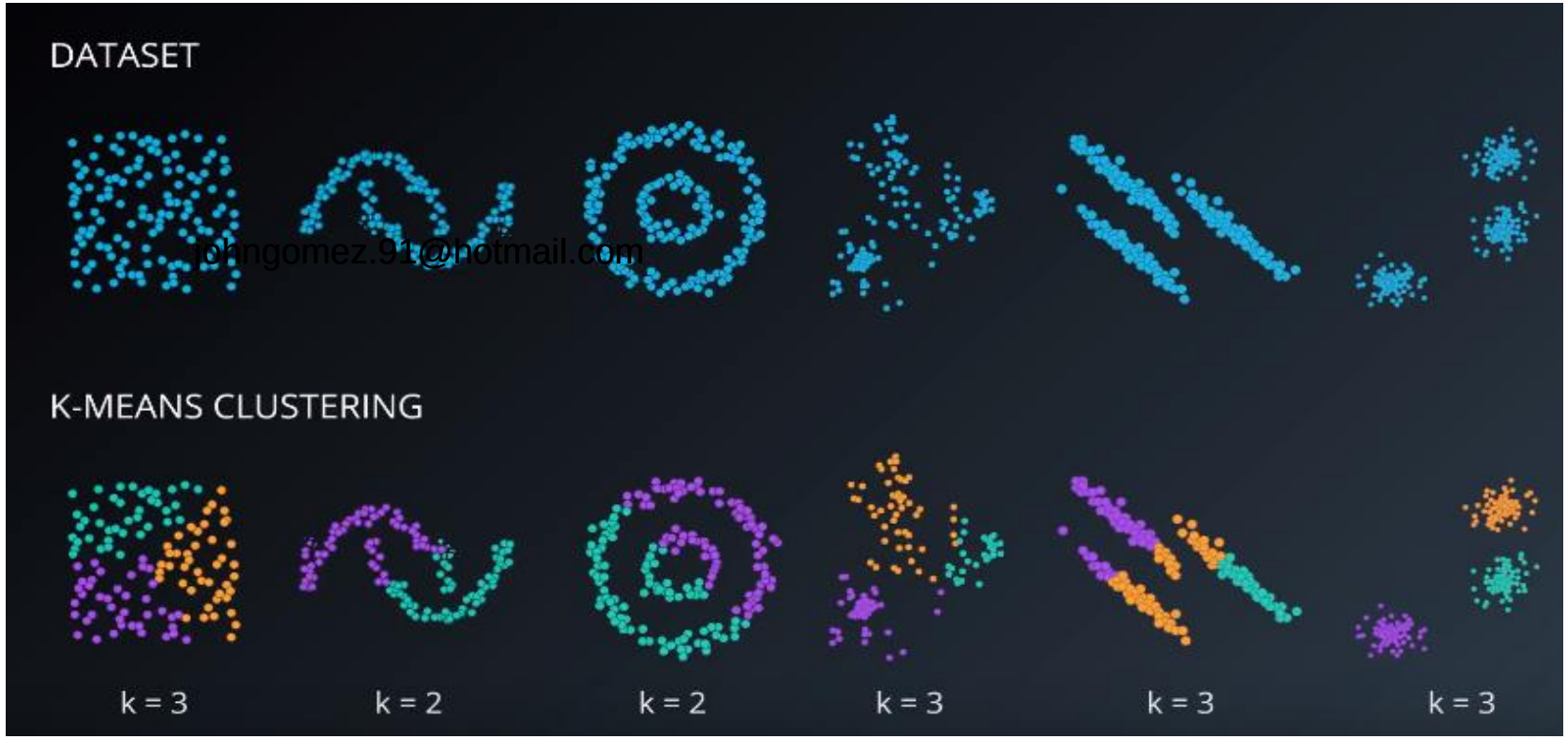
Uk = Cantidad de centroides para inicializar

C(i) distancia para cada punto y los centroides. y se queda con el menor (mas cerca)

$\mu\ (j)$ es el centroide para el grupo j

Ejemplos K-means

La siguiente imagen muestra lo que obtendríamos si utilizamos el agrupamiento de K-means en cada conjunto de datos, incluso si supiéramos de antemano el número exacto de grupos:



Gidia Grupo de I+D en Inteligencia Artificial

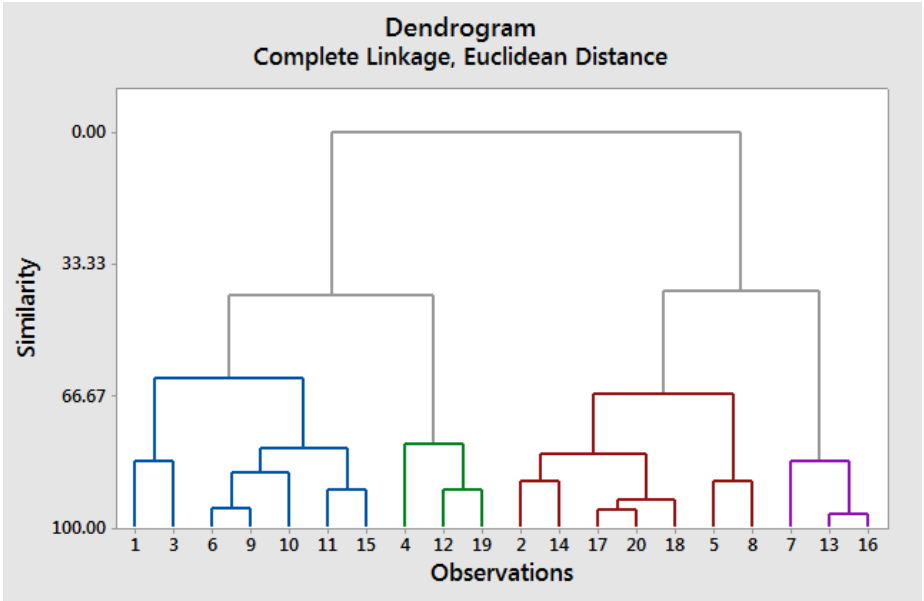
UNIVERSIDAD NACIONAL DE COLOMBIA

Agrupación Jerárquica

https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e

La agrupación jerárquica es una alternativa a los algoritmos de agrupación basados en prototipos. La principal ventaja de la agrupación jerárquica es que **no necesitamos especificar el número de agrupaciones (k)**, la encontrará por sí misma. Además, permite el trazado de dendogramas. Los dendogramas son visualizaciones de una agrupación jerárquica binaria.

Las observaciones que se fusionan en la parte inferior son similares, mientras que las que están en la parte superior son bastante diferentes. Con los dendogramas, las conclusiones se hacen en función de la ubicación del eje vertical en lugar de la horizontal.



Gidia Grupo de I+D en Inteligencia Artificial

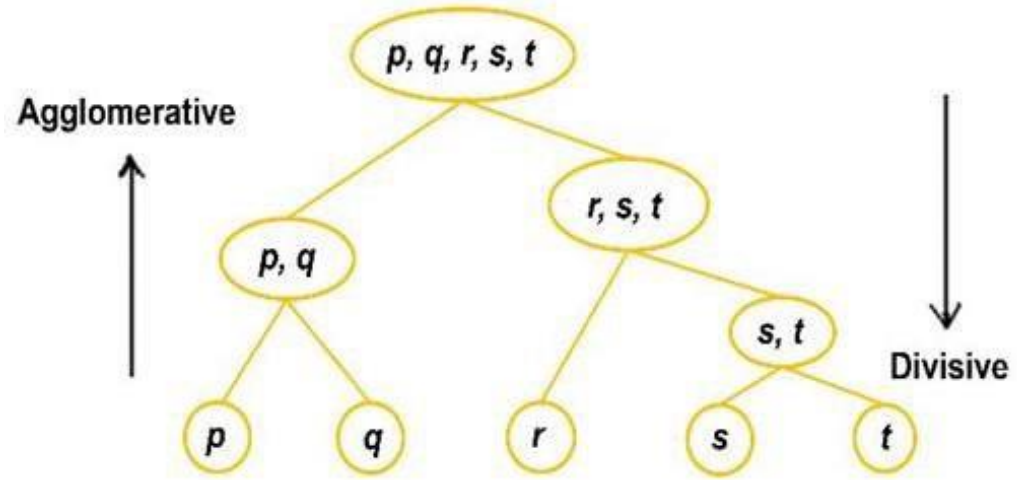
UNIVERSIDAD NACIONAL DE COLOMBIA

Tipos de técnica de agrupación Jerárquica

Esta técnica de agrupamiento se divide en dos tipos:

- **Aglomerativas**: Este es un acercamiento ascendente: cada observación comienza en su propio grupo, y los pares de grupos son mezclados mientras uno sube en la jerarquía.

- **Divisivas**: Este es un acercamiento descendente: todas las observaciones comienzan en un grupo, y se realizan divisiones mientras uno baja en la jerarquía.



Cálculo de similaridad (Jerárquica)

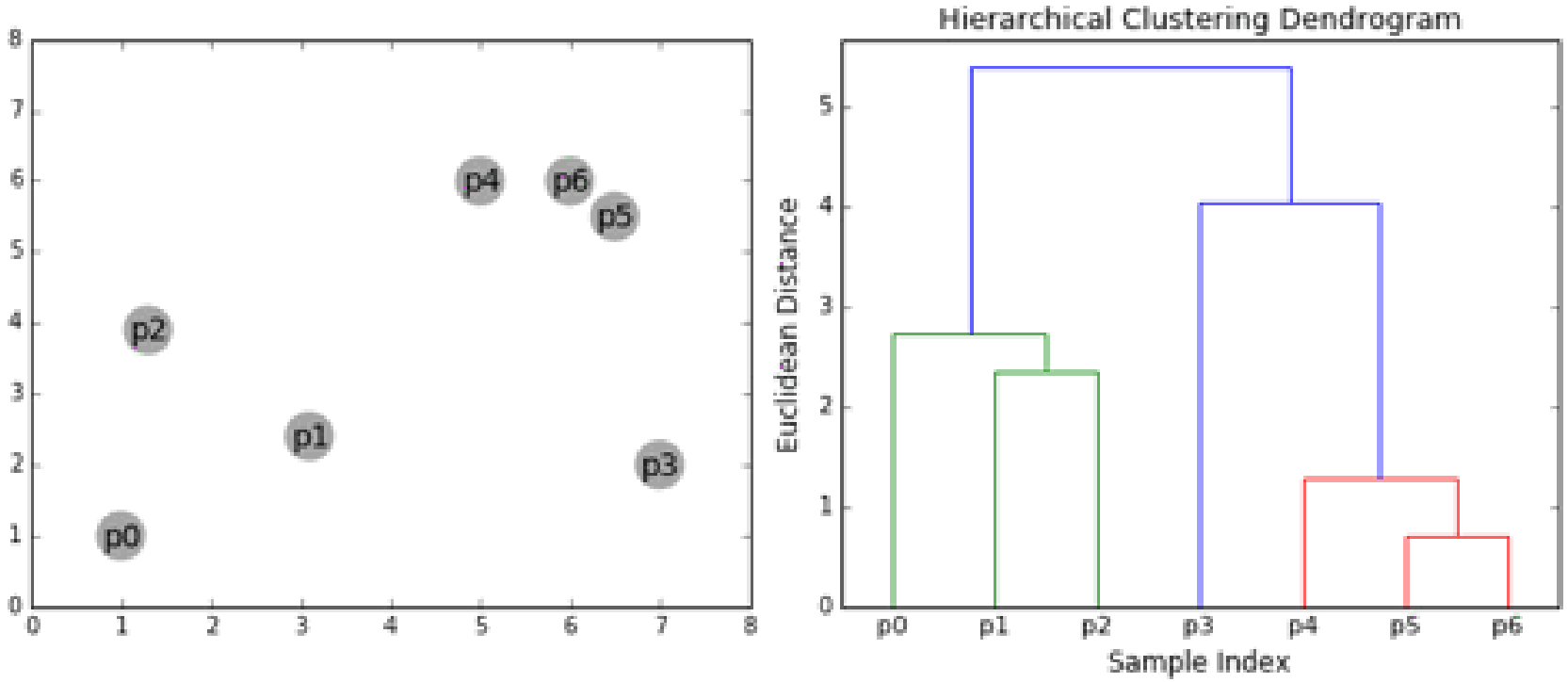
Calcular la similitud entre dos grupos es importante para fusionar o dividir los grupos. Hay ciertos enfoques que se utilizan para calcular la similitud entre dos grupos:

- MIN $\min\{ d(x, y) : x \in \mathcal{A}, y \in \mathcal{B} \}.$
- MAX $\max\{ d(x, y) : x \in \mathcal{A}, y \in \mathcal{B} \}.$
- Promedio grupal $\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y).$
- Distancia entre centroides

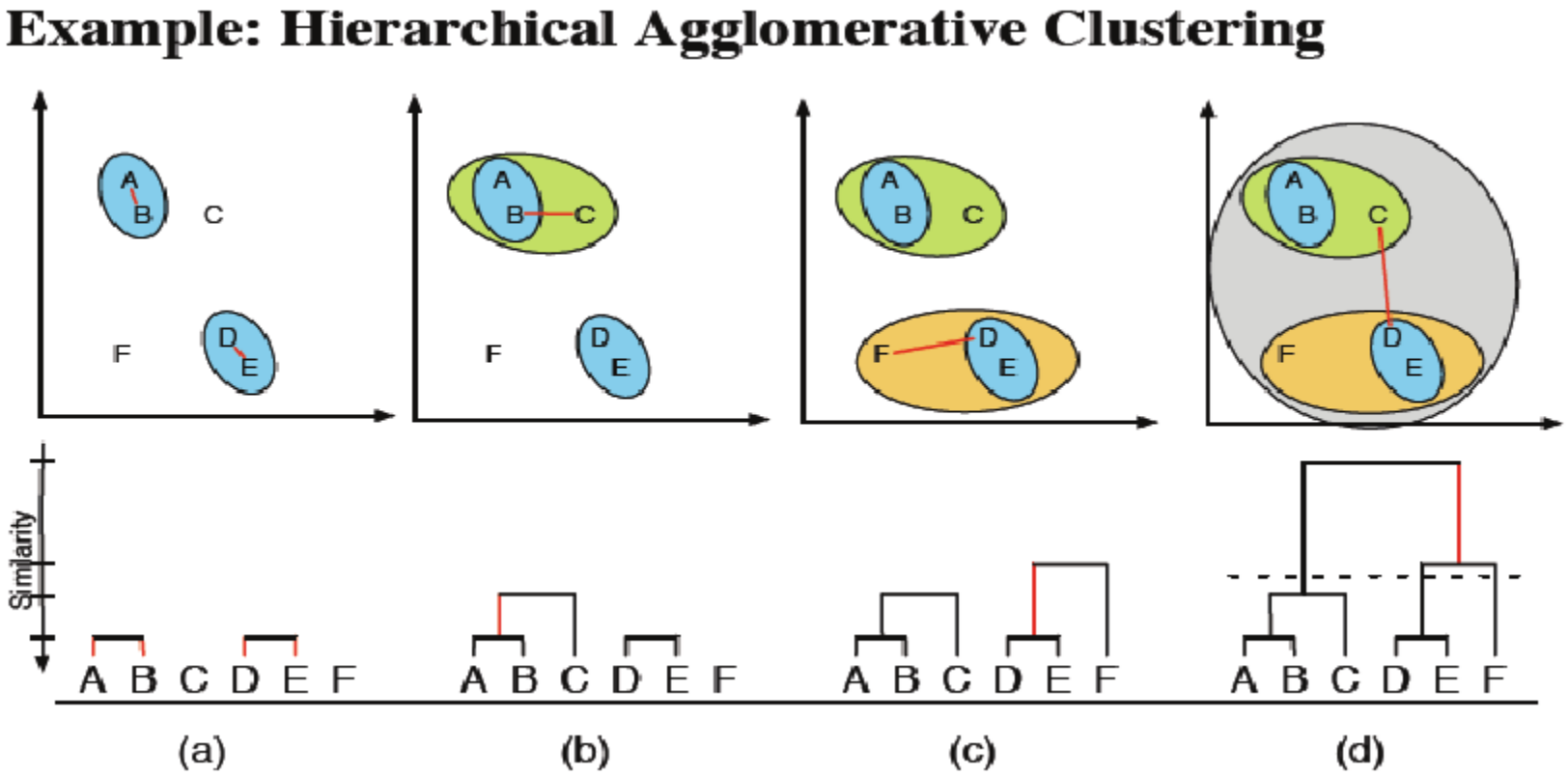
MIN: También conocido como algoritmo de enlace único se puede definir como la similitud de dos grupos C1 y C2 es igual al mínimo de la similitud entre los puntos Pi y Pj, de modo que Pi pertenece a C1 y Pj pertenece a C2.

Sim (C1, C2) = Min Sim (Pi, Pj) tal que Pi ∈ C1 y Pj ∈ C2

Ejemplo de agrupación Jerárquica



Ejemplo de agrupación Jerárquica



https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a

Bisecting K-means

Es una combinación de k-medias y agrupamiento jerárquico. En lugar de dividir los datos en "k" grupos en cada iteración, Bisecting k-means divide un grupo en dos subgrupos en cada paso de bisección (usando k-means) hasta que se obtengan k grupos.

Como Bisecting k-means se basa en k-means, mantiene los méritos de k-means y también tiene algunas ventajas sobre k-means.

Para cada paso de Bisección de k-means, sólo los puntos de datos de un grupo y dos centroides están involucrados en el cálculo y para el algoritmo k-means, el cálculo involucra cada punto de datos y cada K centroides del conjunto de datos. Por lo tanto, el tiempo de cálculo se reduce y el algoritmo de bisecting k-means es más eficiente cuando ‘k’ es grande.

La bisección de k-means produce grupos de tamaños similares, mientras que k-means produce grupos de tamaños muy diferentes.

https://research.ijcaonline.org/volume116/number19/pxc3902799.pdf

Algoritmo bisecting

Paso 1. (Inicialización). Seleccione aleatoriamente un punto, digamos cL ∈ p;
luego calcule el centroide w de M . y calcular cR ∈ p asi cR = w − (cL − w).

Paso 2. Dividir M = [x1, x2,...,xn] en dos sub-grupos ML y MR, de acuerdo con la siguiente regla:

$$\text{rule: } \begin{cases} x_i \in M_L & if \quad ||x_i - c_L|| \leqslant |x_i - c_R|| \\ x_i \in M_R & if \quad ||x_i - c_L|| > |x_i - c_R|| \end{cases}$$

Paso 3. Calcule los centroides de ML y MR, wL y wR.

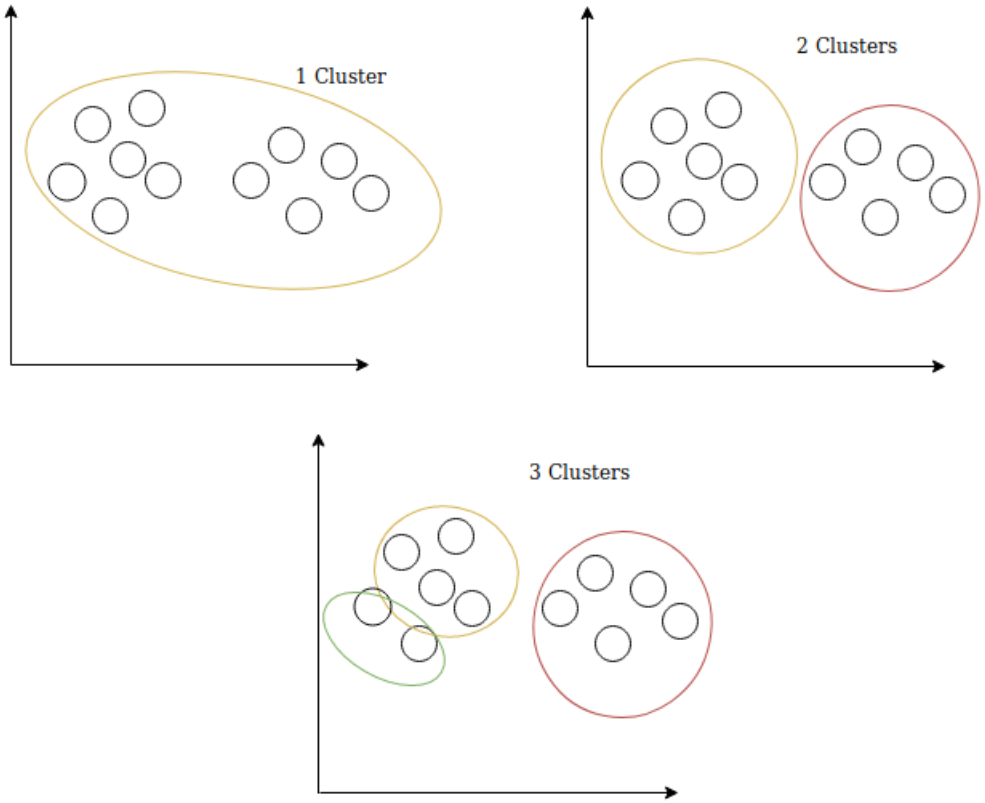
Paso 4. SI wL = cL y wR = cR, parar. De lo contrario, deje cL := wL, cR := wR, y regrese al paso 2.

https://sci-hub.tw/10.3233/IDA-2004-8403

Ejemplo Bisecting K-means

A continuación hay 3 figuras que muestran una definición visual del funcionamiento de este algoritmo.

- Comenzamos con datos completos y tratandolos como un solo gran clúster.
- Dividimos en 2 grupos aplicando k-medias regulares donde k = 2.
- Dado que la condición de detención era tener 3 grupos, digamos más, dividimos uno de los grupos en 2 y finalmente podemos ver 3 grupos separados para los puntos de datos.



https://prakhartechviz.blogspot.com/2019/06/understanding-bisecting-k-means.html

K-modes

La técnica K-Mean no funciona bien en presencia de variables categóricas porque no puede agrupar datos categóricos debido a las diferentes medidas que utiliza.

Mientras que K-Means calcula la distancia euclidiana entre dos puntos, K-Modes intenta minimizar una medida de disimilitud: cuenta el número de "características" que no son las mismas. Usando modas en lugar de medias, los K-Modes pueden manejar eficientemente datos categóricos

Una moda es un vector de elementos que minimiza las diferencias entre el vector en sí y cada objeto de los datos. Tendremos tantos modas como la cantidad de clústeres que necesitemos, ya que actúan como centroides.

https://amva4newphysics.wordpress.com/2016/10/26/into-the-world-of-clustering-algorithms-k-means-k-modes-and-k-prototypes/comment-page-1/

Pasos del algoritmo K-modes

Las modas K asumen que la información del número de grupo probable de datos (es decir, K) es accesible y consta de los siguientes pasos:

1. Genere grupos K seleccionando arbitrariamente objetos de datos y elija K centro de grupo inicial, uno para cada grupo.
2. Asignar objeto de datos al clúster cuyo centro de clúster está cerca hacia ella según la ecuación 3.2.
3. Actualice la base del clúster K en la asignación de objetos de datos más Calcule K últimas modas de cada grupo.
4. Repita los pasos 2 a 3 esperando que no haya cambiado ningún objeto de datos relación de agrupamiento de lo contrario algunos predefinidos adicionales El criterio es cumplir.

Cálculo del algoritmo K-modes

La técnica de k-mode extiende el patrón K-mean a los datos categóricos del grupo mediante la eliminación de la limitación forzada por K-means, como la siguiente modificación:

- Uso de una evaluación de coincidencia simple o distancia de Hamming diferente utilizada para objetos de datos categóricos
- cambiar las medias de agrupación por modas

$$d(x, y) = \sum_{i=1}^{f} \delta(X_j, Y_j)$$

d (x, y) da igual importancia a cada tipo de atributo. Sea Z un conjunto de objetos de datos categóricos descritos por atributos categóricos, A1, A2. . . . . A.m. mientras que lo anterior es utilizado porque la disimilitud determina los datos categóricos objetos, la función de costo se convierte

$$C(Q) = \sum_{i=1}^{n} d(Z_i, Q_i) \qquad (3.2)$$

Donde Zi es el elemento i-ésimo y Qi es el centro del grupo cercano de Zi. La técnica de k-modes minimiza la Función costo definida en la ecuación anterior.

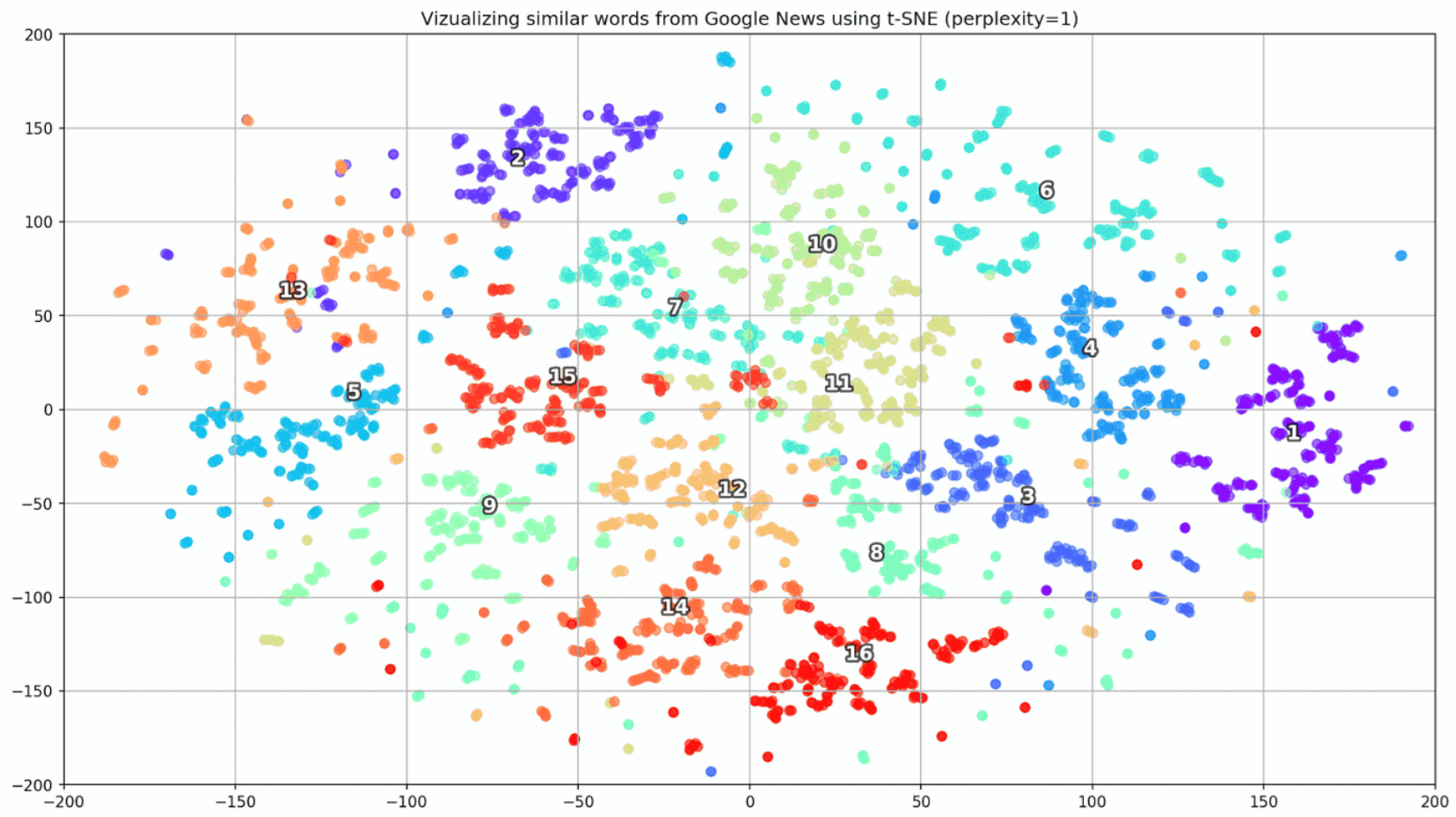
T-SNE

La incrustación de vecinos estocásticos distribuidos en t (t-SNE) **es una técnica no lineal no supervisada que se utiliza principalmente para la exploración de datos y la visualización de datos de alta dimensión.** En términos más simples, t-SNE le da una idea o intuición de cómo se organizan los datos en un espacio de alta dimensión. Fue desarrollado por Laurens van der Maatens y Geoffrey Hinton en 2008.

Se basa en **una técnica de reducción de dimensionalidad no lineal.** La idea básica de t-SNE es **reducir el espacio dimensional manteniendo la distancia relativa por pares entre puntos.** En otras palabras, el algoritmo asigna datos multidimensionales a dos o más dimensiones, donde los puntos que inicialmente estaban lejos uno del otro también siguen lejos después de modelar, y los puntos cercanos también siguen cercanos.

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

Ejemplo algoritmo T-SNE



Autoencoders

Es un tipo de red neuronal artificial utilizada para aprender a copiar su entrada en su salida. Tiene una capa interna (*oculta*) que describe un *código* utilizado para representar la entrada, y está constituido por dos partes principales: un **codificador** que asigna la entrada al código y un **decodificador** que asigna el código a una reconstrucción de la entrada original.

El objetivo de un autoencoder **es aprender una representación (codificación) para un conjunto de datos, típicamente para la reducción de dimensionalidad , entrenando a la red para ignorar el "ruido" de la señal.** Junto con el lado de reducción, se aprende un lado de reconstrucción, donde el codificador automático intenta generar a partir de la codificación reducida una representación lo más cercana posible a su entrada original, de ahí su nombre.

Autoencoders

La forma en que podemos identificar el fraude es mediante la detección de una anomalía. En el aprendizaje no supervisado, **se puede detectar una anomalía con los autoencoders.**

Autoencoder traduce los datos originales en una representación aprendida, en base a esto podemos ejecutar una función y calcular qué tan lejos está la representación aprendida de los datos originales.

Los datos fraudulentos se reconstruyen con una tasa de error más alta, esto ayuda a identificar anomalías. **Los codificadores automáticos son adecuados para el aprendizaje no supervisado: no se requieren datos etiquetados para el entrenamiento.**

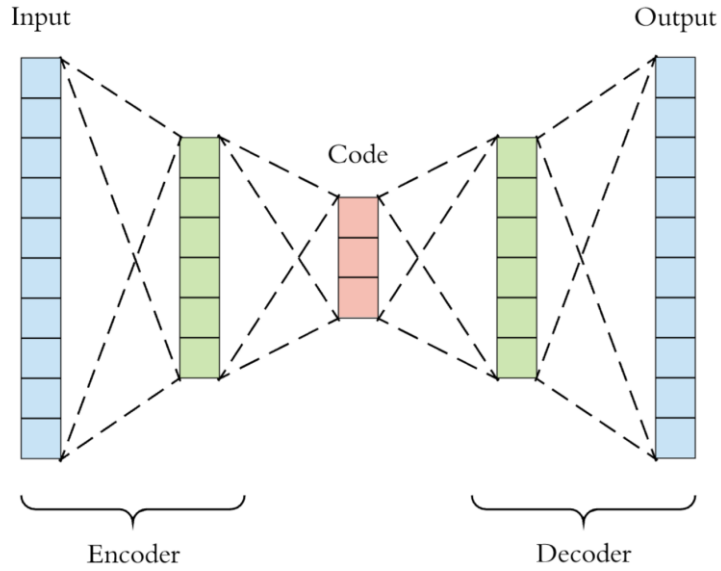
Algoritmo de Autoencoders

https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368

La matemática detrás de las redes es bastante fácil de entender. Esencialmente, dividimos la red en dos segmentos, **el codificador y el decodificador.**

$$\phi : \mathcal{X} \to \mathcal{F}$$
$$\psi : \mathcal{F} \to \mathcal{X}$$
$$\phi, \psi = \underset{\phi,\psi}{\arg\min} \|X - (\psi \circ \phi)X\|^2$$

La función del codificador, denotada por ϕ, asigna los datos originales X a un espacio latente F, que está presente en el cuello de botella. La función de decodificador, denotada por ψ, asigna el espacio latente F en el cuello de botella a la salida. La salida, en este caso, es la misma que la función de entrada. Por lo tanto, básicamente estamos tratando de recrear la imagen original después de una compresión no lineal generalizada.

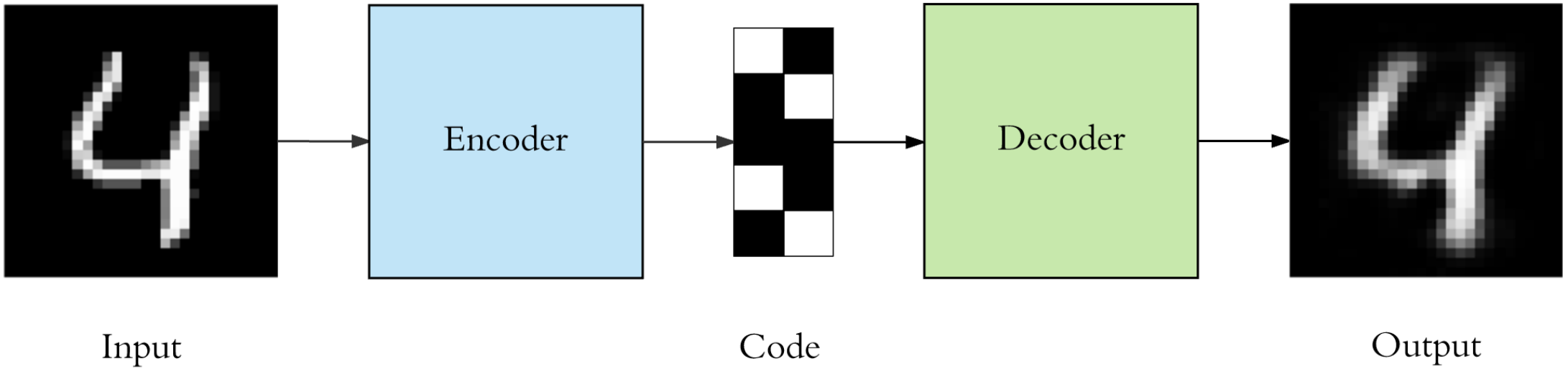


Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

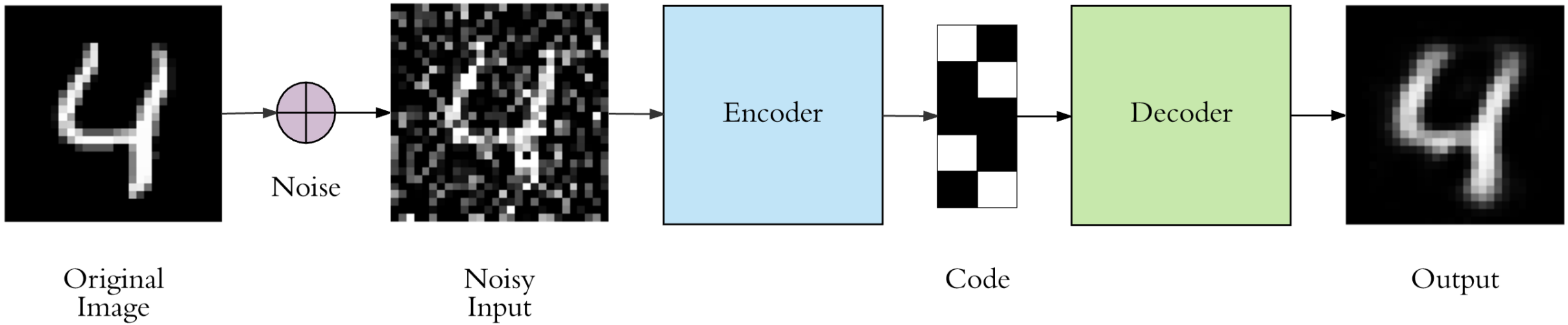
Autoencoders (ANN)

Estos tipos de arquitecturas se encargan de codificar y decodificar una entrada. En aplicación, el encoder es posible utilizarlo como un reductor de dimensionalidad de la entrada dada.



Autoencoders (ANN) - Aplicación

Es posible utilizar estos tipos de arquitecturas para realizar limpieza de ruido en imágenes. Tales como “reducción de ruido de sal y pimienta” o “arreglo de características de imágenes deterioradas”.



Aplicaciones del algoritmo de autoencoder

- Reducción de dimensionalidad
- Recuperación de información
- Detección de anomalías
- Procesamiento de imágenes
- Análisis de imágenes
- Traducción automática de lenguajes
- Predicción de popularidad en redes social

Preguntas

Motivación

- **OBSERVE EL VIDEO Y RESPONDA A LAS SIGUIENTES PREGUNTAS:**

- ¿Cuántos datos se requieren para entrenar un sistema de visión artificial?

- ¿Es posible decir que los computadores ya sobrepasaron la capacidad humana?

- ¿Qué problemas evidencian los sistemas de visión artificial, y en general de los sistemas de Reconocimiento de Patrones?

https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=es

Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

UNIVERSIDAD
NACIONAL
DE COLOMBIA