



UNIVERSITAT OBERTA DE CATALUNYA (UOC)
MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (*Data Science*)

TRABAJO FINAL DE MÁSTER

ÁREA: MINERÍA DE DATOS Y MACHINE LEARNING

Detección de anomalías en entorno del Internet de las cosas

Autor: Gonzalo Pedro Mellizo-Soto Díaz

Tutor: Carlos Hernández Gañán

Profesor: Jordi Casas Roma

Madrid, 16 de abril de 2019

Copyright



Esta obra está sujeta a una licencia de Reconocimiento - NoComercial - SinObraDerivada

3.0 España de Creative Commons.

FICHA DEL TRABAJO FINAL

Título del trabajo:	Detección de anomalías en el entorno del Internet de las cosas
Nombre del autor:	Gonzalo Pedro Mellizo-Soto Díaz
Nombre del colaborador/a docente:	Carlos Hernández Gañán
Nombre del PRA:	Jordi Casas Roma
Fecha de entrega (mm/aaaa):	06/2019
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Minería de datos y Machine Learning
Idioma del trabajo:	Español
Palabras clave	Machine Learning, IOT, Anomaly Detection

Cita

“Nuestro lema es: más humanos que los humanos”

Eldon Tyrell, *Blade Runner*

Agradecimientos

TO BE DEFINED

Si se considera oportuno, mencionar a las personas, empresas o instituciones que hayan contribuido en la realización de este proyecto.

Abstract

In recent years the amount of connected devices has greatly increased, with an increasing number of applications in the industry each day. These devices can be subject of attacks causing instability or data leaks that can be dangerous both for the users and the enterprises, in order to avoid or confront them, security and early detection are becoming a must in a connected world. The focus is the monitoring and detection of the attacks in Internet of Things devices using state of the art Machine Learning techniques. Models such as SVM, DBScan or Isolation Forests have been used and assembled in order to identify with a better accuracy when an attack is happening. With this assembly, attack detection has increased up to 15 % comparing to traditional methods and individual model usages and times have been considerably reduced. An active use of Machine Learning models has shown a great improvement at anomaly detection by securing the devices and decreasing the reaction times when facing attacks.

Durante los últimos años se encuentra una creciente cantidad de dispositivos conectados entre sí, cada vez con más aplicaciones en la industria. Estos dispositivos pueden ser atacados y provocar inestabilidad o una fuga de datos, por lo tanto la protección y la pronta detección de ataques y/o anomalías es vital en un mundo cada vez más conectado. El objetivo es la monitorización y detección de estos ataques en dispositivos del *Internet of Things* utilizando técnicas del estado del arte de Machine Learning para su detección y poder así responder con una mayor rapidez a los ataques. Para la detección se han utilizado modelos estadísticos, como SVM, DBScan o Isolation Forests, que en su conjunto permitan identificar con mayor precisión cuando se está produciendo un ataque. El conjunto de la clusterización con la clasificación de puntos anómalos muestra una mayor robustez, frente al uso individual de cada uno de los modelos aumentando la detección en hasta un 15 %. Se demuestra cómo el uso de los modelos permite proteger los dispositivos y mejorar la seguridad al disminuir los tiempos de reacción frente a los ataques.

Palabras clave: Machine Learning, IOT, Anomaly Detection

Índice general

Abstract	IX
Índice	XI
Listado de Figuras	XIII
Listado de Tablas	1
1. Introducción	3
1.1. Contexto y justificación del Trabajo	3
1.2. Explicación de la motivación personal	4
1.3. Objetivos del Trabajo	4
1.4. Descripción general del problema	5
1.5. Enfoque y método seguido	5
1.6. Planificación del Trabajo	5
2. Estado del Arte	9
2.1. Métodos tradicionales de detección de anomalías	10

2.2. Machine Learning y la detección anomalías	11
2.2.1. Aprendizaje Supervisado	13
2.2.2. Aprendizaje No Supervisado	22
2.2.3. Aprendizaje Semi-Supervisado	29
 Bibliografía	 29

Índice de figuras

1.1. Planificación de tareas	7
2.1. Relación entre Inteligencia Artificial, Machine Learning y Deep Learning	12
2.2. Regresión Lineal	14
2.3. Red Neuronal Artificial	15
2.4. Arquitectura de un Perceptron	15
2.5. Función de activación: Función Escalón	16
2.6. Backpropagation	17
2.7. Hiperplanos SVM	20
2.8. Hiperplanos para distintos valores de C	21
2.9. Efecto de aplicación de <i>Kernel</i> radial en SVM	21
2.10. Datos de dos variables sin PCA	23
2.11. PCA aplicado a dos variables	23
2.12. Ejemplo de agrupamiento	25
2.13. Arquitectura de un Autoencoder	27

2.14. Ejemplo de detección de anomalías con Autoencoders	28
--	----

Índice de cuadros

Capítulo 1

Introducción

1.1. Contexto y justificación del Trabajo

Cada vez se encuentran más dispositivos conectados entre sí, no solo en la industria, sino también en los hogares, esta conexión entre dispositivos los hace vulnerables a ataques informáticos que pueden afectar en gran medida a los usuarios, no solo pueden provocar un mal funcionamiento de los mismos, sino que también puede provocar la fuga de datos de distinta sensibilidad. La previsión en los futuros años es estar cada vez más conectados y una pronta detección de los ataques puede ayudar a evitar los problemas derivados de los mismos, mediante una pronta reacción o haciendo frente a la previsión de un ataque.

Actualmente se utilizan distintas medidas de seguridad como control de acceso físico al dispositivo, encriptación de datos, firewalls, securización de red, etc... Sin embargo, en muchos casos la detección de anomalías aplica un umbral estacionario, provocando que en el caso de un ataque ya sea demasiado tarde para reaccionar o no se sea capaz de identificar patrones extraños previos al ataque. Con la aplicación de nuevas técnicas se pretende mejorar los tiempos de reacción e incluso predecir cuándo puede suceder un ataque.

1.2. Explicación de la motivación personal

La razón principal de la elección del proyecto es la posibilidad de aprender y utilizar técnicas de Machine Learning en un sector desconocido que permita diversificar conocimientos. Todo se encuentra cada vez más conectado e investigar cómo clasificar cuando se está produciendo un ataque permite profundizar en conocimientos de seguridad y aplicarlos en un problema real.

La aplicación de técnicas en un problema real ayuda a comprender mejor el uso de las herramientas y el porqué y cuándo se deben de utilizar. De este modo, se añade un nuevo conocimiento que aportar al ámbito profesional y puede utilizarse también en un entorno privado.

1.3. Objetivos del Trabajo

Los objetivos del trabajo son los siguientes:

- Adquisición de conocimiento del sector y de los dispositivos IOT.
- Lectura y comprensión de las técnicas del estado del arte en detección de anomalías en dispositivos conectados.
- Obtención de datos reales de conexiones a dispositivos IOT, en caso de no ser posible, generación de datos sintéticos.
- Prueba de las técnicas encontradas y evaluación en el problema actual.
- Investigación de algoritmos tradicionales de Machine Learning y su utilidad.
- Desarrollo de la solución utilizando los algoritmos más aptos.
- Evaluación de la detección de patrones y ataques de las técnicas utilizadas.
- Comparación de los resultados obtenidos con el estado del arte.

1.4. Descripción general del problema

La cantidad de dispositivos informáticos existentes que pueden ser víctimas de un ataque es masiva, por lo que securizar correctamente los dispositivos es fundamental. A pesar del uso de distintos protocolos de seguridad, se generan nuevos tipos de ataque todos los días, por lo tanto la detección es una necesidad a la hora de evitar los problemas derivados.

La pérdida de control de los dispositivos o la fuga de información de los mismos puede provocar pérdidas millonarias a las empresas o provocar una gran inseguridad a los usuarios de productos IOT, así como causar posibles daños a infraestructuras o personas. Por otro lado, muchos de los dispositivos pueden no tener la capacidad de computación necesaria para incluir una capa de seguridad robusta, que puede compensarse con una detección temprana.

1.5. Enfoque y método seguido

El enfoque pasa por conocer el problema y dar respuesta a preguntas específicas, mediante la aplicación de los conocimientos obtenidos y la evaluación de la propia aplicación de los mismos.

El método seguido durante el desarrollo se basa en el marco de trabajo *Agile* pensando en el desarrollo de la solución como un producto. De este modo se podrá iterar sobre una arquitectura definida y enfocarse en el desarrollo del producto en su totalidad, frente a un modelo tradicional de cascada donde cada fase de desarrollo recae en una sola parte del proyecto.

El formato de entrega será mediante un *minimum viable product* (MVP) en sprints de dos semanas durante el periodo de desarrollo y evaluación de las necesidades según la evolución del producto.

1.6. Planificación del Trabajo

Para la planificación del trabajo se han subdividido las tareas principales y se han mostrado en el siguiente diagrama de Gantt [1.1](#).

- Pec01 - Definición

- Pec01 - Planificación
- Pec02 - Búsqueda de fuentes del estado del arte
- Pec02 - Lectura de estado del arte
- Pec02 - Justificación de estado del arte
- Pec02 - Redacción
- Pec02 - Refinamiento de objetivos
- Pec03 - Planificación de Sprints
- Pec03 - Sprint 1
- Pec03 - Sprint 2
- Pec03 - Sprint 3
- Pec03 - Refinamiento/Redacción
- Pec04 - Revisión de apartados anteriores
- Pec04 - Redacción de nuevos apartados
- Pec05 - Presentación y defensa

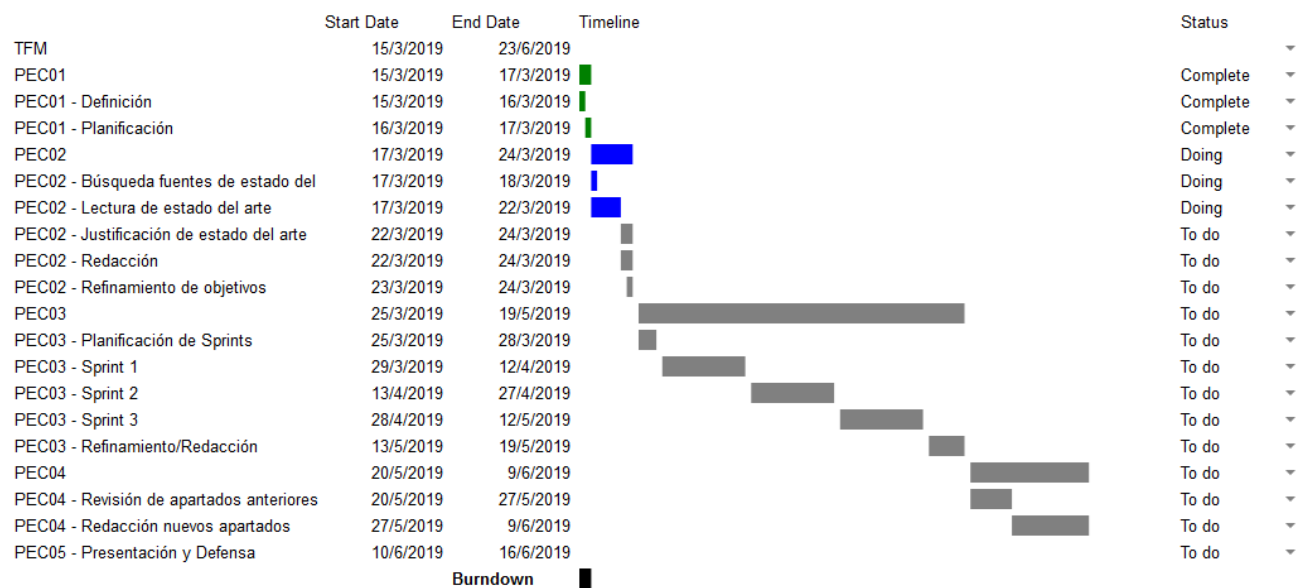


Figura 1.1: Planificación de tareas

Capítulo 2

Estado del Arte

Una de las aplicaciones más comunes dentro de la detección de anomalías es para la detección de intrusos (*Intrusion Detection*) y la creación de sistemas de detección de intrusos (*Intrusion Detection System, IDS*). Se trata de sistemas cuyo objetivo es la monitorización de los sistemas y redes informáticas, con el fin de alertar en caso de que puedan existir brechas de seguridad[1].

Con la ingente cantidad de redes y sistemas que se pueden encontrar a día de hoy es necesario incluir uno de estos sistemas, con el fin de mantener la integridad y disponibilidad de los mismos. Dentro de los IDS, se pueden identificar dos grandes implementaciones de estos sistemas, en los sistemas de detección de intrusos en Host (*Host-Based Intrusion Detection Systems, HIDS*) y los sistemas de detección de intrusos en red (*Network Intrusion Detection Systems, NIDS*).

- **HIDS:** Estos sistemas se caracterizan por implementarse en el Host utilizando información del propio sistema operativo para detectar actos maliciosos [2]. Esta información tiene distintos niveles de información, pero por lo general tienden a ser de bajo nivel sobre operaciones que se pueden estar realizando dentro del sistema. Esta información se consulta dentro de logs, por lo que el análisis de la información es más lento.
- **NIDS:** Para el segundo caso la monitorización se realiza a sistema de red, es decir, de comunicaciones entre distintos nodos y la monitorización de los paquetes que viajan entre ellos [1]. Esta información puede ser consumida en tiempo real, por lo que la reacción ante algún evento es más rápida que en los HIDS que necesitan revisar las acciones.

2.1. Métodos tradicionales de detección de anomalías

En la sección actual se describen algunos de los métodos tradicionales utilizados en IDS, tanto para HIDS como para NIDS.

- Network Security Monitor (NSM): se trata de uno de los primeros sistemas que permitió auditar el tráfico que circulaba dentro de la red [3]. El sistema escucha pasivamente dentro de la red y detecta si existe una conducta sospechosa al desviarse de patrones de conducta. La mayor parte de la monitorización se basa en protocolos estándar como *telnet*, *ftp*, *TCP/IP*, *etc.* por lo que le permitía utilizar una gran cantidad de datos heterogéneos.
- State transition analysis (USTAT): el sistema parte de que el host en un momento se encuentra en un estado seguro y que según las acciones que se realizan sobre el mismo el host cambia de estado, hasta que llega a un estado en el que compromete la seguridad [3]. Este sistema analiza los estados por los que ha pasado la máquina desde el estado seguro al comprometido.
- GrIDS: se trata de un IDS que utiliza un sistema de construcción de grafos basados en la red, donde cada nodo representa a un host y las aristas las conexiones entre los mismos. La representación gráfica de la actividad de la red permite ayudar al espectador en identificar qué está sucediendo [3].
- Haystack: en este caso el IDS se ayuda de métodos estadísticos para la detección de anomalías, definiendo estrategias para usuarios y grupos, además de definir variables del modelo como variables gaussianas independientes [4]. Para la detección se incluyen una serie de intervalos en los valores que en el momento que salen del rango normal, se calcula la distribución de probabilidades y si el *score* o puntuación es demasiado grande se genera una alerta.

Los métodos/sistemas listados se desarrollaron durante los años noventa, la tecnología ha evolucionado desde entonces y los sistemas se han vuelto más complejos y más propensos a los ciberataques. Por ello, se han desarrollado nuevas técnicas que se apoyan en el uso de técnicas de minería de datos (*Data Mining*) y las técnicas que se describirán a continuación de *Machine Learning*.

2.2. Machine Learning y la detección anomalías

El *Machine Learning* es una rama de la inteligencia artificial cuya premisa es hacer que la máquina aprenda una tarea sin haber sido específicamente programada para ello. El término fue descrito por Arthur L. Samuel en 1959 en un artículo en el que explica estudios de *Machine Learning* aplicado al juego de las damas [5], utilizando en una primera instancia métodos de aprendizajes más generales, como las redes neuronales de las que hablaremos más adelante, y otro métodos que tendrán que ser parametrizados para sus distintos usos.

La inteligencia artificial se puede entender como la inteligencia ejercida por las máquinas, al contrario que la inteligencia natural inherente a los humanos, que son capaces de realizar tareas cognitivas que permiten potenciar la resolución de las mismas e imitar comportamientos como el aprendizaje [6]. Dentro de la inteligencia artificial se pueden encontrar distintas definiciones según si se centran en el razonamiento o en el comportamiento:

- Actuar humanamente: este enfoque se basa en que las máquinas actúen como humanos más centrado en la interacción de máquinas con personas, más que en la resolución de problemas. Esta interacción se puede ver reflejada en el test ideado por Alan Turing en 1950, en el que se somete a la máquina inteligente.^a un interrogatorio realizado por un humano, donde éste no sabe que está hablando con una máquina, por lo tanto si es incapaz de detectar que se trata de una máquina esta ha actuado como un humano y puede considerarse inteligente.
- Pensar humanamente: esta vertiente se centra en imitar el pensamiento humano, entender cómo funcionan las mentes de los mismos y replicarlo en máquinas. Siguiendo esta corriente, si se consigue imitar el pensamiento humano, una máquina que resuelva el problema utilizará razonamientos humanos, en lugar de solucionar problemas a toda costa, independientemente de como lo realizan los humanos.
- Pensar racionalmente: basado en la lógica formal desarrollada en en siglos XIX y XX que permite formular los problemas en un lenguaje formal, utilizando el razonamiento matemático para resolver éstos.
- Actuar racionalmente: se centra en realizar el comportamiento más efectivo en un momento dado. Antes las distintas situaciones no siempre existe una acción correcta, pero si se puede llegar realizar una acción que minimice los riesgos.

Los enfoques mostrados conllevan sus propios enfoques filosóficos, sin embargo, han influenciado en gran medida a cómo se afrontan los problemas y cómo se han desarrollado las técnicas de inteligencia artificial que están en uso actualmente.

Como se ha comentado con anterioridad, el *Machine Learning* es una rama de la inteligencia artificial, dentro de la cual existe otra rama, *Deep Learning*. Las relaciones entre los términos se puede observar en la imagen 2.1 y como la inteligencia artificial engloba ambas ramas.

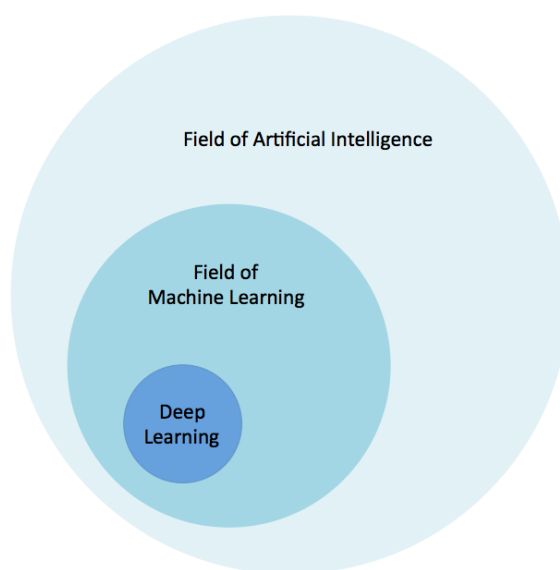


Figura 2.1: Relación entre Inteligencia Artificial, Machine Learning y Deep Learning

El *Deep Learning* se diferencia del *Machine Learning* convencional, en que están formados por modelos con distintas ramas de procesamiento que son capaces de extraer las representaciones de los datos con varios niveles de abstracción [7]. Mucha de la teoría relacionada se puede encontrar en los años 50 y 60, sin embargo, las aplicaciones y la capacidad de cómputo eran limitadas en la época, hasta ahora donde la capacidad de computación ha crecido a gran escala y en conjunto con las nuevas técnicas propuestas, una mayor cantidad de datos y la aplicación de transformaciones no lineales hacen que esta vertiente viva su época dorada.

Una de las implementaciones más representativas del *Deep Learning* se trata de las redes neuronales, las cuales se encuentran basadas en el funcionamiento del cerebro humano, utilizando el concepto de neuronas que pueden realizar cálculos, la conexión entre las mismas, la función de realizar una tarea específica, etc. Añadir, que se encuentra basado y que el cerebro humano es un sistema mucho más complejo que las redes neuronales y tiene muchos más comportamientos [8].

Dentro del *Machine Learning*, los algoritmos pueden dividirse según como se realice el entrenamiento del modelo, afectando también a la evaluación y las aplicaciones del mismo. Estas categorías son: *Aprendizaje Supervisado (Supervised Learning)*, *Aprendizaje No Supervisado (Unsupervised Learning)* y *Aprendizaje Semi-Supervisado (Semi-Supervised Learning)*

2.2.1. Aprendizaje Supervisado

El aprendizaje supervisado se caracteriza por utilizar un conjunto de variables de entrada y encontrar la función que más se aproxime a los valores de salida [9]. Actualmente es una de las metodologías más utilizadas en *Machine Learning* y es una de las más avanzadas en el campo, sin embargo, la necesidad de los valores de salida (variable dependiente o *target*) requiere que el conjunto de datos se encuentre etiquetado, es decir, para las observaciones de las variables de entrada debe existir la variable de salida para poder general el modelo. En muchos casos este etiquetado se debe de realizar manualmente, por lo que es un proceso que requiere mucho tiempo.

Por otro lado, cabe mencionar que el uso de métodos de aprendizaje supervisado tienen una mayor efectividad que el semi y no supervisado, sin embargo, la necesidad de tener las etiquetas y el coste que ello supone para grandes cantidades de datos, como es en el caso de la detección de anomalías, provocan que se busquen otras soluciones con aprendizajes semi y no supervisado. Además, suele suceder que en la detección de anomalías, éstas sean la menor parte de todos los datos, en otras palabras, la mayor parte de los datos son normales y solo una pequeña porción se trata de anomalías como tal, provocando que el conjunto de datos esté poco compensado.

Uno de los modelos más sencillos de aprendizaje supervisado es la regresión lineal (simple), ésta permite ajustar a una serie de puntos conocidos , x , y su respuesta y [10], generando una función que permite predecir los valores de y con valores no conocidos de x :

$$y = ax + b \tag{2.1}$$

En este caso se predicen los nuevos valores utilizando la función ajustada, existen distintos métodos para encontrar la recta que ajusta los puntos, siendo uno de los más utilizados el ajuste

por mínimos cuadrados:

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (2.2)$$

$$b = y - a\bar{x} \quad (2.3)$$

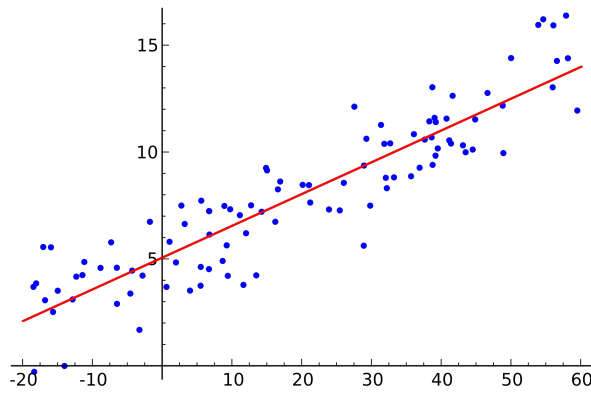


Figura 2.2: Regresión Lineal

Como se puede observar en la imagen 2.2, se ha generado una recta capaz de ajustarse a los puntos y predecir el valor para nuevos puntos. Para generar la función de la recta se ha necesita utilizar la variable dependiente y en el entrenamiento para poder generar la función que aproximará los nuevos puntos.

Durante los siguientes apartados se van a mostrar los distintos modelos utilizados en aprendizaje supervisado para la detección de anomalías, como las redes neuronales artificiales (*Artificial Neural Networks*, *ANN*, *Gradient Boosting Machines* y *Support Vector Machines*)

2.2.1.1. Feedforward Neural Network

Tal y como se ha comentado anteriormente, las redes neuronales intentan imitar el concepto de la conexión de neuronas del cerebro humano, éstas están formadas por una capa de entrada de datos (*Input*), una capa de salida (*Output*) y un conjunto de capas intermedias, también llamadas capas ocultas (*Hidden*) [8]. En las capas intermedias se encuentran las neuronas conectadas entre sí y pueden estar formadas desde una sola capa, hasta n capas, sin embargo, cuanto mayor sea

el número de capas, mayor coste de computación asociado. En la siguiente imagen 2.3 podemos ver como sería una arquitectura de una red neuronal:

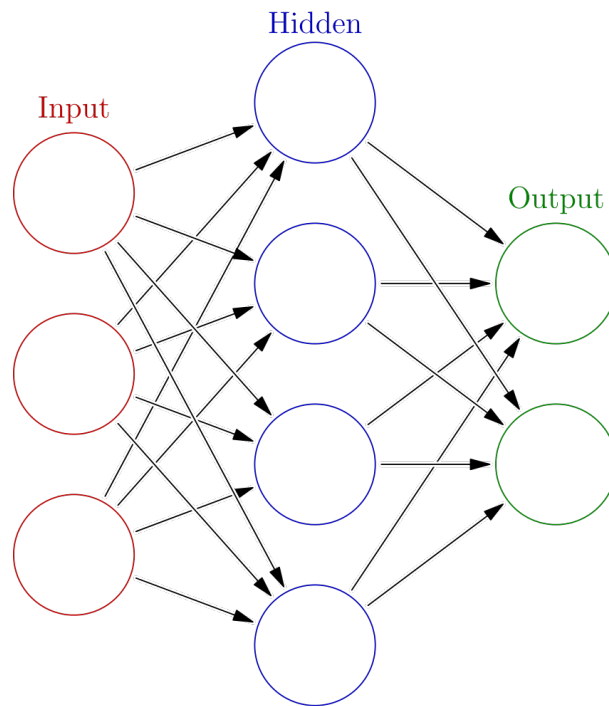


Figura 2.3: Red Neuronal Artificial

Dentro de las redes neuronales existe un caso base conocido como perceptron, un clasificador binario formado por la capa de entrada, una capa oculta de una sola neurona y una salida [11]. El perceptrón ayudará a incluir el concepto de los pesos (*weights*) y el *bias*.

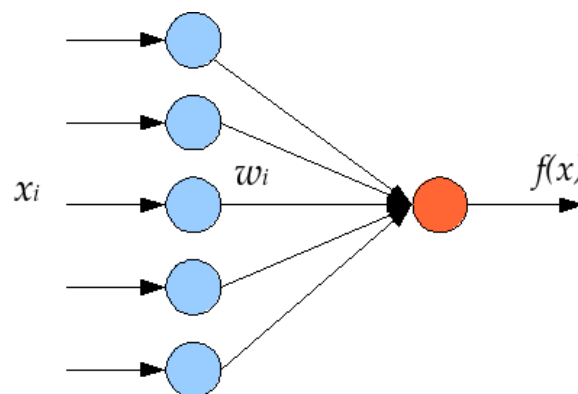


Figura 2.4: Arquitectura de un Perceptron

Los valores de entrada X son multiplicados por una serie de pesos W , en una primera instancia se inician aleatoriamente, y se realiza la suma de esta multiplicación, tras la cual se pasará a una función de activación. La función de activación define el valor de salida de la neurona y que se utilizará como valor de entrada para la siguiente neurona o como es en este caso como el valor de salida final. Un ejemplo de una función de activación es la función escalón, donde los valores menores de cero son iguales a cero y los mayores de cero son igual a uno. El *bias* se trata de un pequeño valor que modifica el output sin interactura con las neuronas. De este modo la clasificación de 0 o 1 en un perceptron sería así:

$$f(x) = g(\sum XW + b) \quad (2.4)$$

Donde $g(x)$ es la función de activación, X el vector de datos de entrada, W el vector de pesos y b el vector de bias.

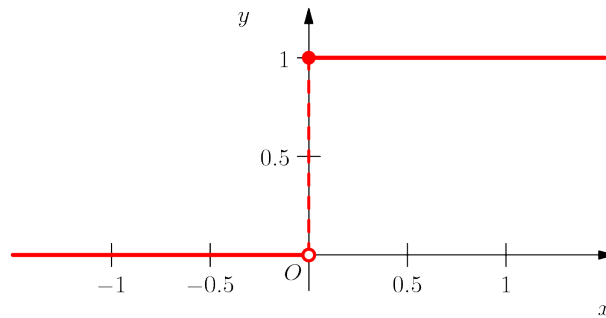


Figura 2.5: Función de activación: Función Escalón

Aumentando la complejidad de la arquitectura, es decir, el número de neuronas, el número de capas y distintas funciones de activación, se consiguen modelos más complejos pero mantienen el enfoque que se realiza en el perceptron.

Para optimizar la salida, se utilizan varias técnicas de optimización que permiten actualizar los pesos W y disminuir el error $E = f(x) - y$, una de las más utilizadas es el *backpropagation*, basado en el descenso del gradiente cuyo objetivo es obtener el gradiente de la función (vector de mayor pendiente) pero utilizar el valor opuesto del mismo multiplicado por un coeficiente de aprendizaje (*learning rate*), una analogía muy utilizada es imaginar como encontrar el punto más bajo del valle, para ello lo mejor es fijarse en que lugar del punto en el que nos encontremos tiene más pendiente negativa y seguir ese camino. El *backpropagation* utiliza este método, pero en lugar de calcular el gradiente sobre la función total (gran coste) se realiza el cálculo de los

gradientes locales, empezando desde el punto final de la función.

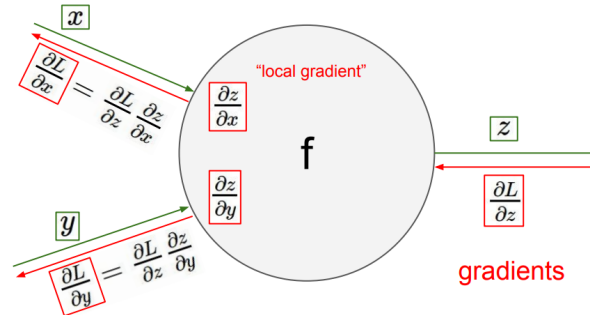


Figura 2.6: Backpropagation

La fórmula de la actualización de los pesos en una ANN sería similar a la siguiente:

$$W_{i+1} = W_i - \gamma dW \quad (2.5)$$

Donde γ es el *learning rate* y $L(x)$ es la función de pérdida.

La última capa, la capa de *outputs*, define el número de neuronas de salida, que puede ser igual al número de clases etiquetadas o una sola neurona en el caso de que existan solo dos clases, clasificación binaria. El valor de las salidas irá definido por la función de activación de la última capa de la capa oculta.

2.2.1.2. Gradient Boosting Machines

Las *Gradient Boosting Machines* son uno de los métodos más utilizados actualmente, dada la gran precisión que aportan a la resolución de los problemas. Estos modelos se basan en realizar el ensamblado de modelos débiles, normalmente árboles de decisión, para generar un predictor "fuerte" de una manera iterativa.

El término *Boosting* se refiere al método utilizado para mejorar el aprendizaje de un modelo mediante el ensamblado de modelos débiles, esto quiere decir que son modelos con una capacidad de predicción mejor que la pura aleatoriedad [12].

Concretamente el *Gradient Boosting* trata de generar un primer árbol para realizar las predicciones, es decir, crea una función $F(x)$ para aproximar y , posteriormente se utiliza la

función de coste definida (puede ser un error cuadrático medio para una regresión, por ejemplo) para ver como de bien ha realizado la predicción. Los residuales obtenidos se utilizan para crear un nuevo modelo que utilice estos residuales con el fin de volver a minimizar el error y finalmente se genera un nuevo modelo que utilice ambos. De esta forma en cada iteración se puede añadir n cantidad de nuevos modelos que ayuden a mejorar las predicciones mediante la corrección de los errores de los modelos previos [13].

Se crea el primer modelo

$$F_1(x) = y \quad (2.6)$$

Se crea un segundo modelo con los residuales

$$h_1(x) = y - F_1(x) \quad (2.7)$$

Se genera un nuevo modelo con los anteriores

$$F_2(x) = F_1(x) + h_1(x) \quad (2.8)$$

Dentro del ajuste a los residuales es donde entra la parte del gradiente, tal y como hemos explicado anteriormente se utiliza el método del descenso del gradiente (*Gradient Descent*) para la optimización de la función de pérdida. Para cada paso que se realiza, en vez de calcular los residuales como en la fórmula 2.7 se ajustan calculando el gradiente de la función de pérdida y ajustando un nuevo modelo, h , con los nuevos residuales y en la generación del nuevo modelo incluirle *gamma* (γ), similar al concepto explicado de *learning rate*:

$$r = -\left[\frac{\partial L(F(x_i), y)}{\partial F(x_i)}\right] \quad (2.9)$$

$$F_2(x) = F_1(x) + \gamma h_1(x) \quad (2.10)$$

En la práctica el uso de estos modelos se ha comprobado que es realmente eficaz para la resolución de problemas. Dentro de estos algoritmos se pueden destacar dos en concreto, cuyas implementaciones se han utilizado en varias soluciones dentro de las competiciones realizadas por Kaggle.

XGBoost

Diminutivo de *eXtreme Gradient Boosting* se trata de una nueva implementación de GBM realizada por Chen, Tianqi, et al [14], que demuestra una mayor escalabilidad y velocidad que los métodos tradicionales, utilizando una menor cantidad de recursos de los sistemas.

LightGBM

Desarrollado por Microsoft [15] se propone mejorar los problemas encontrados en algoritmos como *XGBoost* cuando existe una alta cantidad de variables, esto es ocasionado por la necesidad de evaluar el punto óptimo para la división de los árboles de decisión. Propone su propia forma de efectuar la ganancia de información (por división) y demuestra una mejora en el tiempo de cómputo sobre los modelos convencionales.

Algunas de las competencias donde se han utilizado los algoritmos (también ensamblándolos) son las siguientes:

- Home Credit Default Risk: Can you predict how capable each applicant is of repaying a loan? [16].
- Corporación Favorita Grocery Sales Forecasting: Can you accurately predict sales for a large grocery chain? [17].
- Google Analytics Customer Revenue Prediction: Predict how much GStore customers will spend [18].

2.2.1.3. Support Vector Machines

Se trata de otro algoritmo enfocado a la clasificación mediante la separación de los puntos por un hiperplano [10]. Por ejemplo, para un set de datos etiquetados con dos categorías (clasificación binaria) se busca encontrar la línea (hiperplano) que mejor separe ambos puntos, como puede observarse en la imagen 2.7 existen distintos hiperplanos (H_1, H_2, H_3) que separan las distintas clases, como puedo observarse tanto H_1 como H_2 son capaces de separar las dos clases por completo, mientras que H_3 fallaría en la tarea de clasificación .

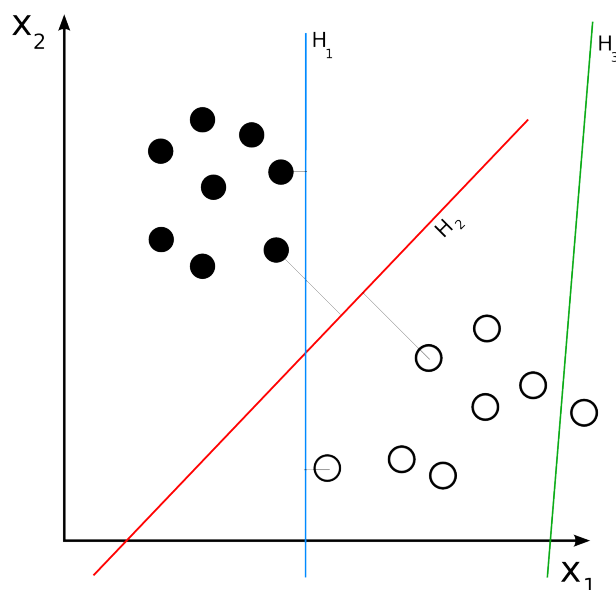
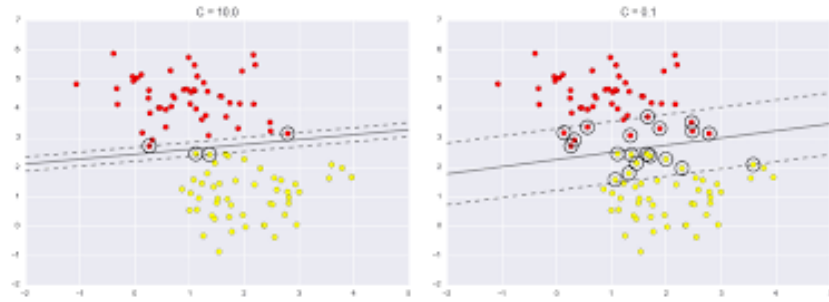


Figura 2.7: Hiperplanos SVM

Se puede definir el hiperplano como la "línea" que puede dividir los datos entre las dos clases, es decir, se pretende separar el espacio en dos mitades. Esta "línea" para un problema de p dimensiones sería de $p - 1$ dimensiones, en otras palabras, para un problema de dos dimensiones sería una línea, para uno de tres sería un plano y así sucesivamente hasta n dimensiones. Esta separación indica que según donde se encuentren los puntos con respecto al hiperplano se clasificarán como una clase u otra.

$$\beta_0 + X_1\beta_1 + X_2\beta_2 + \dots + X_n\beta_n = 0 \quad (2.11)$$

Como se ha mencionado en la imagen 2.7, existen varios hiperplanos, sin embargo, se debe de buscar el hiperplano óptimo de los posibles, siendo este el que se encuentra más alejado de los puntos de entrenamiento, es decir, es el hiperplano con mayor margen entre los puntos y el hiperplano. Sin embargo, si solo nos basamos en encontrar el mejor hiperplano según los márgenes se puede dar el caso de que un nuevo punto afecte en gran medida al hiperplano, por maximizar los márgenes, por ello se utiliza un parámetro C que permite "ablandar" el clasificador y permitir que ciertos puntos puedan encontrarse entre el hiperplano y el margen, la mala clasificación de unos pocos puntos permitidos evita sobreentrenar el modelo y permitir generalizar mejor.

Figura 2.8: Hiperplanos para distintos valores de C

Por ahora, se ha hablado de separar los datos linealmente y en dos clases, sin embargo, ese no es siempre el caso, podemos tener más de dos clases o por ejemplo tenemos datos que no sean separables por una "línea", pensemos en dos conjuntos de puntos que forman una serie de circunferencias concéntricas (2.9), para este caso una línea nunca logrará separar bien los puntos. Para esta última, la *SVM* puede utilizar los *kernels*, se trata de funciones que permiten cambiar la dimensionalidad del espacio para ayudar a la separación de las clases.

Kernel lineal

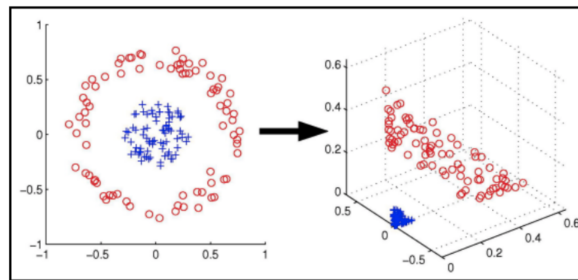
$$K(x_i, x_{i'}) = \sum_{j=1}^p x_{ij} x_{i'j} \quad (2.12)$$

Kernel polinomial

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^p x_{ij} x_{i'j}\right)^d \quad (2.13)$$

Kernel radial

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^p (x_{ij} x_{i'j})^2\right) \quad (2.14)$$

Figura 2.9: Efecto de aplicación de *Kernel* radial en SVM

Para el caso de tener más de dos clases se pueden aplicar dos enfoques distintos:

- Clasificación uno contra uno: para este caso se crean una SVM por cada par de clases y realiza la clasificación para ellos, luego el resultado es elegido observando cuantas veces el punto cae en una categoría para las distintas SVM clasificándolo a la clase más frecuente.
- Clasificación uno contra todos: en este caso se realiza una SVM por clase, de modo que se compare la clase escogida con el resto y valorando en cual de los modelos generados es más apta la clasificación del punto.

2.2.2. Aprendizaje No Supervisado

Este nuevo tipo de aprendizaje se distingue porque no necesita ese conjunto de etiquetas que se utiliza por observación, es decir, en este caso no se tiene el valor de la clase o el valor numérico correspondiente a cierto conjunto de datos. Para este caso no se enfoca a la predicción de los valores, debido a que no tenemos esa variable *target* para la que queremos desarrollar la función que la aproxime lo máximo posible, en este caso, el objetivo principal es intentar descubrir nuevos patrones o información oculta dentro de los datos [10].

Muchas veces el uso del aprendizaje no supervisado no se encuentra con el simple objetivo de predecir, si no que puede ser parte de un proceso de análisis de los datos, de modo que permitan encontrar más información útil de los mismos, como por ejemplo, se puede realizar una segmentación de usuarios, de modo que posteriormente se pueda estudiar que causa esta segmentación de los mismos y mejorar los servicios que se les puedan ofrecer.

El uso del aprendizaje no supervisado esta creciendo cada vez más, dado que en la mayor parte de los casos los datos no vienen etiquetados, tal y como se ha comentado anteriormente el etiquetado de los mismos es un proceso costoso y además mientras que en el supervisado se puede "supervisar" si el resultado es correcto en la fase de entrenamiento, mientras que para el no supervisado, no se tiene esta capacidad de conocer cual es el valor real, dado que es imposible saber cual es la respuesta correcta.

2.2.2.1. Análisis de Componentes Principales

Se trata de uno de los algoritmos más conocidos cuya finalidad es explicar en una serie de "componentes", siendo la cantidad de éstos menor a la cantidad de variables de los datos iniciales, que permiten explicar la mayor parte de la varianza de los datos originales. Estos "componentes" se trata de un conjunto de nuevas variables que son capaces de explicar una gran parte de la varianza de un conjunto mayor, en otras palabras, permite obtener gran parte de la varianza explicada en una dimensionalidad menor.

En las siguientes imágenes se ha utilizado un set de datos sintéticos de dos dimensiones, en el que se muestra como en la imagen de la izquierda 2.10 se encuentran nubes de datos que cambian en función de los ejes x e y y como la aplicación del análisis de componentes principales 2.11, *PCA* de ahora en adelante, genera nuevas variables (dos componentes principales) que tienden a explicar la varianza de las variables originales. En este caso se puede apreciar como la aplicación del algoritmo "mueve" las observaciones, haciendo que así sean más identificables en a lo largo de los ejes.

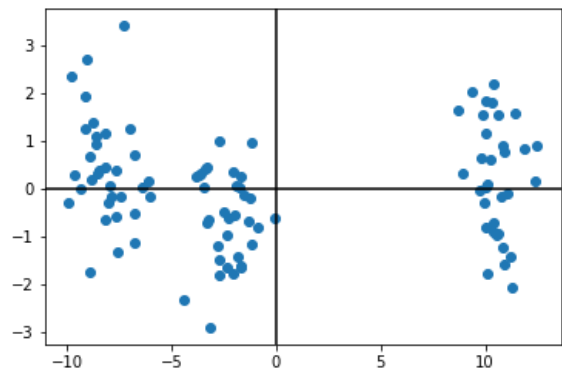
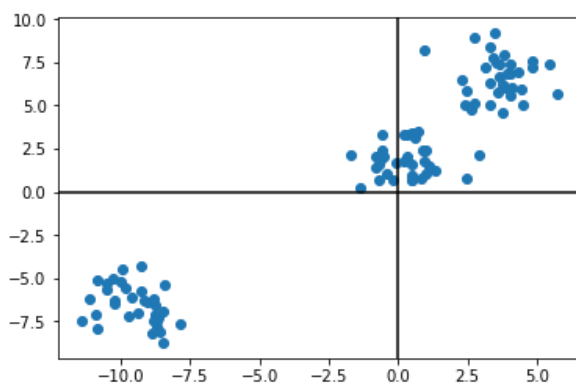


Figura 2.10: Datos de dos variables sin PCA

Figura 2.11: PCA aplicado a dos variables

Este ejemplo se trata de un caso sencillo, por lo general esta técnica también se puede utilizar para visualizar los datos de una gran dimensionalidad en dos o tres variables, utilizando ese número de componentes principales, que se tratan de una combinación lineal de las variables anteriores [10].

Para realizar el cálculo del PCA, se puede realizar mediante el uso de álgebra lineal, más en concreto con una descomposición de autovalores y autovectores que siguen los siguientes pasos

para una matriz X con p variables y n observaciones [19]:

Se calcula la media de cada una de las columnas

$$\overline{X} = \overline{X_1}, \overline{X_2}, \dots, \overline{X_p} \quad (2.15)$$

Se normaliza el valor de las columnas restando la media de cada columna a todos los valores de la matriz

$$X_{norm} = X - \overline{X} \quad (2.16)$$

Se genera la matriz de covarianza con la matriz normalizada, siendo E la media, por cada par de variables [20]

$$Cov[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])] \quad (2.17)$$

De la matriz de covarianza se calculan los autovalores y autovectores

$$\lambda, \nu < -X_{cov} \quad (2.18)$$

Una vez obtenidos los autovectores y autovalores, se seleccionarán una cantidad m que será la cantidad de componentes principales elegidos

$$B = [\lambda_1 \dots \lambda_m, \nu_1 \dots \nu_m] \quad (2.19)$$

Y con estos calculamos la proyección P de la matriz original X con m componentes principales

$$P = B^T \cdot A \quad (2.20)$$

2.2.2.2. Agrupación

Tal y como el propio nombre indica, actualmente se utilizan métodos de agrupación en el que podemos encontrar grupos o *clusters*. Esta agrupación se realiza mediante la similitud entre observaciones, donde la similitud puede ser la distancia euclídea entre las observaciones [10], útil cuando todas las variables son numéricas. En la imagen 2.12 podemos ver un ejemplo donde datos (sin etiquetar, imagen de la izquierda) se analizan y se agrupan (imagen de la derecha). En este caso la agrupación es clara, sin embargo, la agrupación se puede realizar con datos más abstractos y menos representables dada su alta dimensionalidad.

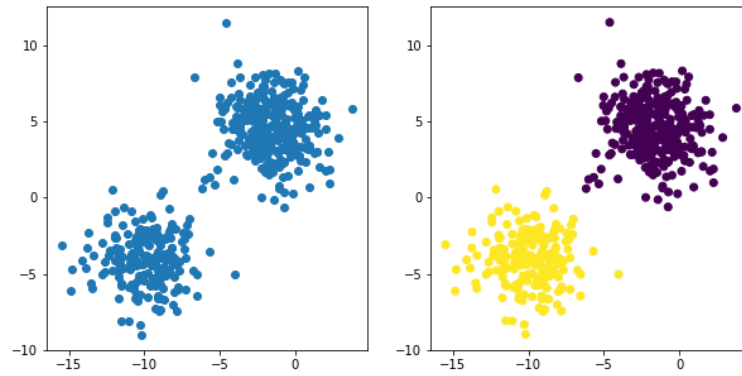


Figura 2.12: Ejemplo de agrupamiento

Como se ha comentado, la similitud entre observaciones ayuda a encontrar que grupos son más similares, por lo tanto dependiendo del problema se pueden utilizar distintas distancias como la euclídea o manhattan, siendo estas de las más utilizadas:

Distancia euclídea

$$d(X_1, X_2) = \sqrt{(X_{11} - X_{21})^2 + (X_{12} - X_{22})^2 + \dots + (X_{1n} - X_{2n})^2} \quad (2.21)$$

Distancia manhattan

$$d(X_1, X_2) = \sum_{i=1}^n |X_{1i} - X_{2i}| \quad (2.22)$$

Estas distancias se aplican a datos con variables numéricas por lo que no son adecuadas cuando existe variables discretas, sin embargo, se pueden realizar transformaciones en estas

variables a numéricas para poder utilizarlas. En el caso donde la mayor parte de las variables sean categóricas esta transformación no daría buenos resultados, por lo que se pueden utilizar distancias específicas para variables discretas, como *Goodall*, *Lin* o *Smirnov* [21].

Dentro de los algoritmos de clusterización los más conocidos son los siguientes:

- KMeans o K-medias: es un algoritmo de clusterización en el cual se indica a priori el número de clústers K que se quieren realizar [10]. Una vez definido se generan k centroides aleatorios, se calculan las distancias de estos centroides al resto de puntos de modo que se asocia a cada punto el centroide que más cercano se encuentre (se puede decir que se le asigna una clase o grupo) y finalmente se actualiza el centroide utilizando los valores medios de todos sus puntos asociados. Se vuelven a calcular las distancias, asignar de nuevo los centroides y la actualización de éstos hasta que no se realizan grandes cambios, es decir, converge o hasta un número máximo de rondas.
- DBSCAN: este algoritmo se basa en la generación de los grupos mediante la identificación de áreas de alta densidad (de puntos) frente a zonas de poca densidad [22]. En este caso no se necesita indicar el número de clústers a priori, pero sí un número mínimo de observaciones como para poder considerar que se trata de un nuevo grupo. En este algoritmo no todas las observaciones se agrupan, si no cumplen la cantidad mínima estas observaciones se pueden considerar como *outliers* o anomalías.
- Algoritmos jerárquicos: en este tipo de algoritmos se define una medida de disimilitud (distancia euclídea, por ejemplo) entre los distintos puntos (considerados como un grupo cada uno), de modo que aquellos puntos con la menor disimilitud se agrupan en un nuevo clúster, generando un nuevo grupo de puntos. Con este grupo de puntos surge una nueva medida que permite identificar la disimilitud entre los mismos, *linkage*. Por lo general existen distintos modos de *linkage* siendo los más comunes: completo (se calcula la disimilitud entre todos los puntos de los grupos y se selecciona la mayor), *single* (se selecciona la menor), media (se hace la media de todas las disimilitudes) y centroide (se calcula la disimilitud entre los centroides de los grupos)

Tal y como se ha comentado anteriormente, generalmente se suelen utilizar conjuntos de varias técnicas que permitan mejorar los resultados obtenidos utilizando una técnica sola, un ejemplo en la detección de anomalías es el uso de PCA y KMeans, arrojando buenos resultados en dispositivos IOT [23].

2.2.2.3. Autoencoders

Los autoencoders son redes neuronales con una determinada estructura, de modo que no se necesite un set de datos etiquetados (variable dependiente o *target*), el objetivo de esta red neuronal es reconstruir los datos de entrada en la capa de salida. Esto se realiza utilizando el mismo número de neuronas de entrada en la salida, de modo que utilizando las técnicas de las redes neuronales, como la optimización por descenso del gradiente y el *backpropagation* se puedan generar nuevas variables (*features*) que reconstruyen los datos de entrada [24].

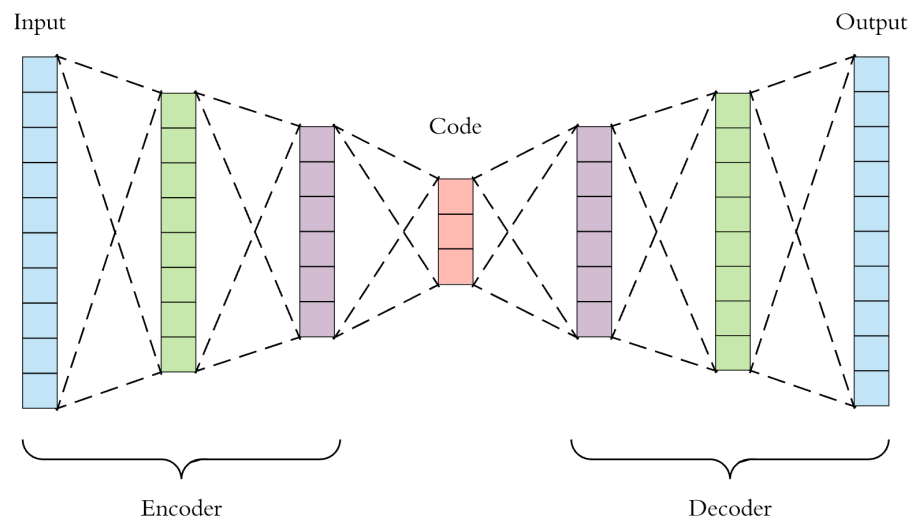


Figura 2.13: Arquitectura de un Autoencoder

Como se puede observar en la imagen 2.13 las dimensiones de entrada son iguales en la salida, y las capas intermedias se pueden dividir en dos partes:

- *Encoder*: permite codificar la entrada a una dimensionalidad menor, de modo que se genera una representación de menor dimensionalidad de la entrada y se generan las nuevas *features*. La capa intermedia se puede utilizar como representación de menor dimensionalidad, similar a la reducción de dimensionalidad del PCA, que puede utilizarse como variables para otro algoritmo.
- *Decoder*: decodificador utilizando la capa intermedia donde se pretende reconstruir los datos de entrada, utilizando las *features* generadas por el *Encoder* hasta llegar a la dimensionalidad de entrada.

Utilizando este método es posible, a partir de datos sin etiquetar, entrenar una red neuronal que permita reconstruir fielmente los datos de entrada, pero en el caso de las anomalías esta reconstrucción no sería del todo fiel (o similar a datos comunes), en otras palabras, el error de la reconstrucción sería mayor que el de una representación normal. De este modo se permite identificar picos en el error que muestren que la reconstrucción no es fiel y que la observación en cuestión se pueda tratar de una anomalía.

En la imagen 2.14 se puede observar un ejemplo del uso de autoencoders que permite clasificar una transacción fraudulenta (clase 1) de una no fraudulenta (clase 0). En este caso la reconstrucción de los datos no fraudulentos representan un error muy pequeño, en comparación a sus contrapartidas, y se puede observar como la mayor parte de estos se encuentran debajo del umbral definido. Aquellos datos anómalos no acaban de reconstruirse con total fidelidad aumentando el error definido (en este caso el error cuadrático medio). No todos son identificados correctamente, pero si permite aumentar la detección de los mismos, en algunos casos donde la cantidad de datos es muy grande y la cantidad de anomalías es muy pequeña este tipo de enfoques puede ayudar a disminuir la cantidad de trabajo necesario para confirmar el fraude, un ejemplo sería en el blanqueo de capitales donde la mayor parte de casos no constituyen blanqueo, reducir el número de casos a revisar puede ahorrar una gran cantidad de costes y tiempo.

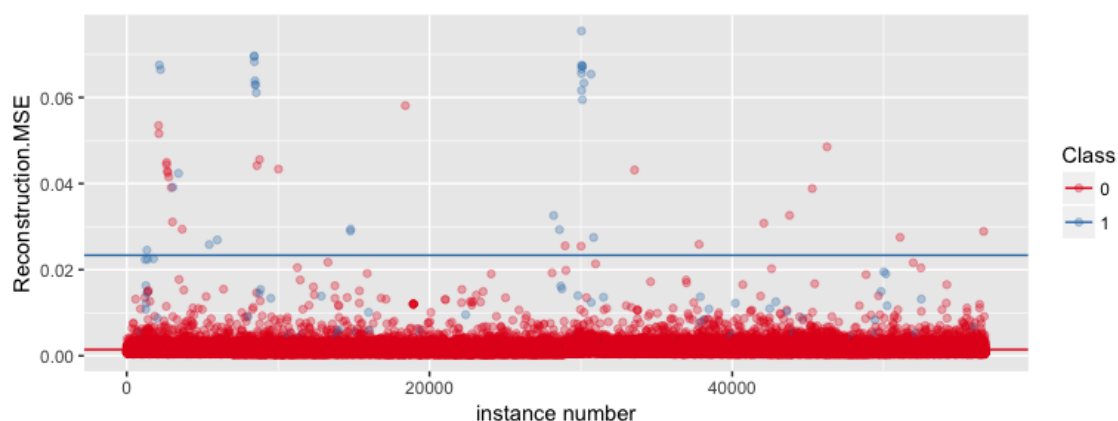


Figura 2.14: Ejemplo de detección de anomalías con Autoencoders

2.2.2.4. Generative Adversarial Networks

2.2.3. Aprendizaje Semi-Supervisado

2.2.3.1. Autoencoders

2.2.3.2. Convolutional Neural Networks

2.2.3.3. Generative Adversarial Networks

Bibliografía

- [1] Rebecca Bace and Peter Mell. Intrusion detection systems. *National Institute of Standards and Technology (NIST)*, 2001.
- [2] Giovanni Vigna and Christopher Kruegel. Host-based intrusion detection. 2005.
- [3] Stefan Axelsson. Research in intrusion-detection systems: a survey. *Department of Computer Engineering, Chalmers University of Technology*, 1998.
- [4] Pedro Garcia-Teodoro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28, 2009.
- [5] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226, 2000.
- [6] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.
- [7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [8] Simon Haykin. *Neural networks*, volume 2. Prentice hall New York, 1994.
- [9] Qiong Liu and Ying Wu. *Supervised Learning*. Springer US, Boston, MA, 2012.
- [10] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning – with Applications in R*, volume 103 of *Springer Texts in Statistics*. Springer, 2013.
- [11] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

- [12] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. 96:148–156, 1996.
- [13] Ben Gorman. A kaggle master explains gradient boosting, 2017. Last accessed 27 March 2019.
- [14] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 2016.
- [15] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. pages 3146–3154, 2017.
- [16] Home credit default risk, 2018. Last accessed 27 March 2019.
- [17] Corporación favorita grocery sales forecasting, 2018. Last accessed 27 March 2019.
- [18] Google analytics customer revenue prediction, 2019. Last accessed 27 March 2019.
- [19] How to calculate principal component analysis (pca) from scratch in python, 2018. Last accessed 04 April 2019.
- [20] Covariance matrix, 2019. Last accessed 04 April 2019.
- [21] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 243–254. SIAM, 2008.
- [22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 1996.
- [23] Anonymous. Inferring and characterizing internet-scale iot probing campaigns by leveraging a novel data dimensionality reduction technique. *ACSAC*, 2018.
- [24] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.