¿Qué es la regresión lineal?

La regresión es un método para modelar un valor objetivo basado en predictores independientes. Este método se utiliza principalmente para pronosticar y descubrir la relación de causa y efecto entre variables. Las técnicas de regresión difieren principalmente en función del número de variables independientes y el tipo de relación entre las variables independientes y dependientes.

Para otros usos de este término, véase Función lineal (desambiguación).

Ejemplo de una regresión lineal con una variable dependiente y una variable ***independiente.***

En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε.

La regresión lineal simple es un tipo de análisis de regresión donde el número de variables independientes es uno y existe una relación lineal entre la variable independiente (x) y la dependiente (y)

¿Qué es la regresión logística?

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. ***Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores.***

El algoritmo de regresión logística también utiliza una ecuación lineal con predictores independientes para predecir un valor.

Necesitamos que la salida del algoritmo sea una variable de clase, es decir, 0-no, 1-sí. (ejemplo Spam/ no spam)

¿Un árbol de decisión?

Es un modelo de predicción utilizado en diversos ámbitos. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

¿Random forest? (o random forests)

También conocidos en castellano como "'Bosques Aleatorios"' es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de **bagging** que construye una larga colección de árboles no correlacionados y luego los promedia.

¿Qué es Support Vector Machine?

Las máquinas de vectores de soporte o máquinas de vector soporte (del inglés Support Vector Machines, SVM) son un conjunto de algoritmos de aprendizaje supervisado.

Estos métodos están propiamente relacionados con problemas de clasificación y regresión. *Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra.*

Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama **vector soporte.** Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase.

Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) ***que puede ser utilizado en problemas de clasificación o regresión***. Una buena separación entre las clases permitirá una clasificación correcta.

Dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos) pertenece a una categoría o a la otra.

La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior.

En ese concepto de "separación óptima" es donde reside la característica fundamental de las SVM: este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. Por eso también a veces se les conoce a las SVM como clasificadores de margen máximo.

Los modelos basados en SVM están estrechamente relacionados con las redes neuronales. Usando una función kernel, resultan un método de entrenamiento alternativo para clasificadores polinomiales, funciones de base radial y perceptrón multicapa.

Kernel

La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano N-dimensional.

Desafortunadamente los universos a estudiar no se suelen presentar siempre en casos idílicos de dos dimensiones, sino que un algoritmo SVM debe tratar con ***a) más de dos variables predictoras, b) curvas no lineales de separación, c) casos donde los conjuntos de datos no pueden ser completamente separados, d) clasificaciones en más de dos categorías.***

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio de funciones Kernel ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de la máquinas de aprendizaje lineal. Es decir, mapearemos el espacio de entradas X a un nuevo espacio de características de mayor dimensionalidad

Tipos de funciones Kernel (Núcleo)

- Polinomial-homogénea
- Perceptron
- Función de base radial

SVR. Regresión

Una nueva versión de SVM para regresión

La idea básica de SVR consiste en realizar un mapeo de los datos de entrenamiento $x \in X$, a un espacio de mayor dimensión F a través de un mapeo no lineal $\varphi: X \rightarrow F$, donde podemos realizar una regresión lineal.

Support Vector Regression es una variante del modelo de análisis Support Vector Machine utilizado para clasificar, sin embargo, con esta variante el modelo de vector soporte se utiliza como un esquema de regresión para predecir valores.

Este algoritmo se basa en buscar la curva o hiperplano que modele la tendencia de los datos de entrenamiento y según ella predecir cualquier dato en el futuro. Esta curva siempre viene acompañada con un rango (máximo margen), tanto del lado positivo como en el negativo, el cual tiene el mismo comportamiento o forma de la curva.

Los datos pueden ser lineales o no lineales, ya que al final el modelo se ajusta al comportamiento de los datos, lo importante es que se cumplan ciertos criterios para que

el modelo obtenga resultados óptimos. Algunos de los criterios que se deben considerar son los siguientes:

- Los datos deben estar limpios, por lo que se deben preprocesar con anterioridad.
- No es adecuado para conjuntos de datos grandes ya que el tiempo de entrenamiento puede ser alto.

SVC (Soporte de Vectores para Clasificación)

SVC es la clase principal ofrecida por Scikit-learn y la función de coste viene determinada por el parámetro C . El parámetro kernel especifica el tipo de kernel a usar. Las implementaciones nativas son "linear", "poly", "rbf" y "sigmoid". T

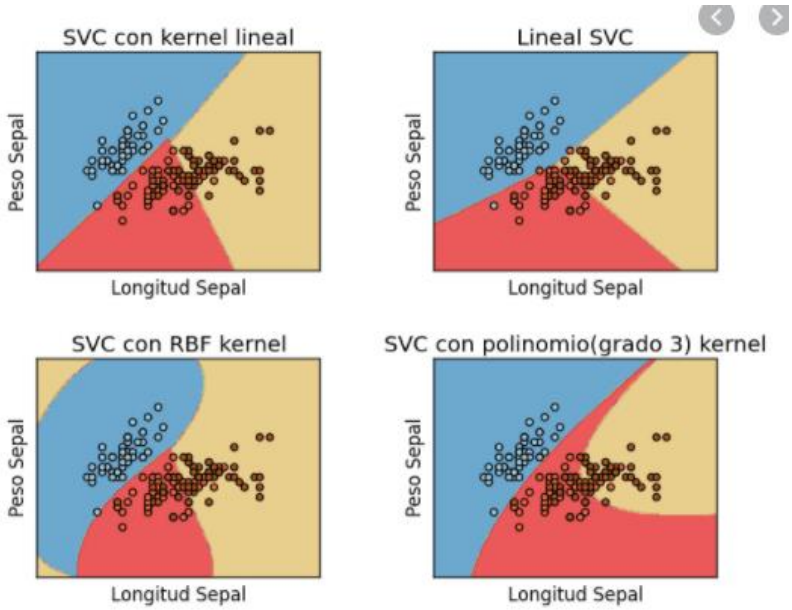

Imagen de algunos tipoes de SVC