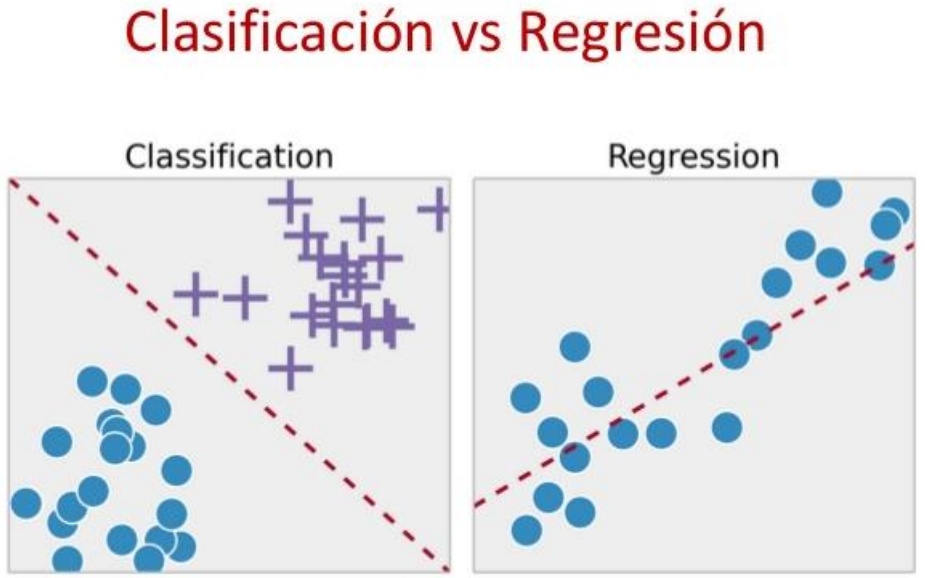
Clasificación Vs Regresión

POR: JOSÉ CARLOS MURILLO

Clasificación vs Regresión

La regresión tiene el objetivo de predecir valores continuos (Números pues, como el 1, 2.3, 3.1416 etc…), Y la clasificación tiene la tarea de asignar una clase, es decir predecir a que clase pertenece un conjunto de datos, aquí es muy importante entender que en los problemas de clasificación los valores son discretos .



Clasificación

Cuando usamos clasificación, **el resultado es una clase, entre un número limitado de clases**. Con clases nos referimos a categorías arbitrarias según el tipo de problema.

Ejemplo:

- ¿comprará el cliente este producto? [sí, no]
- ¿tipo de tumor? [maligno, benigno]
- ¿es este comportamiento una anomalía? [sí, no]
- ¿nos devolverá este cliente un crédito? [sí, no]
- ¿qué deporte estás haciendo? tal y como lo detectan los relojes inteligentes [caminar, correr, bicicleta, nadar]

Regresión

Cuando usamos regresión, **el resultado es un número**. Es decir, el resultado de la técnica de machine learning que estemos usando será un valor numérico, dentro de un conjunto infinito de posibles resultados.

Ejemplos:

- Predecir por cuánto se va a vender una propiedad inmobiliaria
- Predecir cuánto tiempo va a permanecer un empleado en una empresa
- Estimar cuánto tiempo va a tardar un vehículo en llegar a su destino
- Estimar cuántos productos se van a vender

Arboles de Decisión

Ventajas

En comparación con otros algoritmos, los árboles de decisión requieren menos esfuerzo para la preparación de datos durante el pre-procesamiento.

No requiere normalización de datos.

No requiere escalar los datos.

Un modelo de árboles de decisión es muy intuitivo y fácil de explicar.

Desventajas

Un pequeño cambio en los datos puede causar un gran cambio en la estructura del árbol.

A veces el cálculo puede ser mucho más complejo en comparación con otros algoritmos.

El entrenamiento del árbol de decisión es relativamente costoso ya que la complejidad y el tiempo que se toma es más.

El algoritmo del árbol de decisión es inadecuado para aplicar regresión y predecir valores continuos.

Árbol de Clasificación: Índice Gini

La impureza de Gini es una medida de cuán a menudo un elemento elegido aleatoriamente del conjunto sería etiquetado incorrectamente si fue etiquetado de manera aleatoria de acuerdo a la distribución de las etiquetas.

Alcanza su mínimo (cero) cuando todos los casos del nodo corresponden a una sola categoría de destino.

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$

Árbol de Regresión: MSE

En estadística, el error cuadrático medio (ECM) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El ECM es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

Árbol de Regresión: Score

El Score nos representa el coeficiente de determinación (R^2)

El resultado puede variar entre 0 y 1, esto significa que mientras más cerca esté del uno estará más ajustada a la variable que intentas probar, mientras que en el caso contrario, es decir, cuanto más se acerca a 0 menos fiable será ya que estará menos ajustado el modelo.