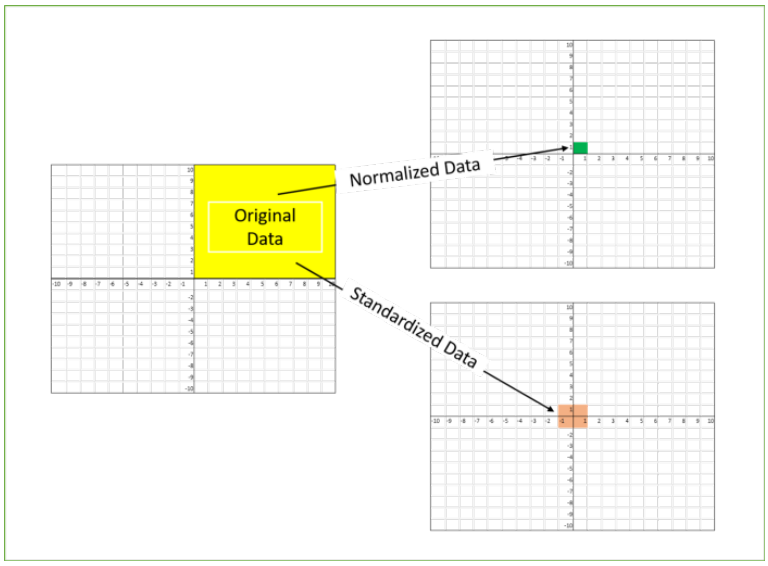
03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa **Tel** 966870140 **Fax** 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

1 MACHINE LEARNING – Escalado

El escalado de características en el aprendizaje automático es uno de los pasos más críticos durante el preprocesamiento de datos antes de crear un modelo de aprendizaje automático. El escalado puede marcar la diferencia entre un modelo de aprendizaje automático débil y uno mejor.

Las técnicas más comunes de escalado de características son la Normalización y la Estandarización.

La normalización se utiliza cuando queremos acotar nuestros valores entre dos números, normalmente, entre [0,1] o [-1,1]. Mientras que la Normalización transforma los datos para que tengan una media cero y una varianza de 1, hacen que nuestros datos no tengan unidades. Consulte el siguiente diagrama, que muestra cómo se ven los datos después de escalarlos en el plano X-Y.



1.1 ¿Por qué necesitamos escalar?

El algoritmo de aprendizaje automático sólo ve números, si hay una gran diferencia en el rango, por ejemplo, unos pocos en miles y unos pocos en decenas, y **asume que los números más altos tienen algún tipo de superioridad**. Así, estos números más significativos empiezan a desempeñar un papel más decisivo en el entrenamiento del modelo.

El algoritmo de aprendizaje automático trabaja con números y no sabe lo que representan. Un peso de 10 gramos y un precio de 10 dólares representan dos cosas completamente diferentes, lo cual es obvio para los humanos, pero para un modelo como característica, trata ambas como si fueran lo mismo.

03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa **Tel** 966870140 **Fax** 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

Supongamos que tenemos dos características de peso y precio, como en la siguiente tabla. El "Peso" no puede tener una comparación significativa con el "Precio". Así que el algoritmo de suposición hace que desde "Peso" > "Precio", por lo tanto "Peso", es más importante que "Precio".

Name	Weight	Price
Orange	15	1
Apple	18	3
Banana	12	2
Grape	10	5

Así, estos números más significativos empiezan a desempeñar un papel más decisivo en el entrenamiento del modelo. Por lo tanto, es necesario escalar las características para que cada una de ellas tenga la misma importancia. Curiosamente, si convertimos el peso a "Kg", "Precio" pasa a ser dominante.

1.2 ¿Cuándo escalar?

El escalado de características es esencial para los algoritmos de aprendizaje automático que **calculan distancias entre datos**. Si no se escala, la característica con un rango de valores más alto empieza a dominar a la hora de calcular distancias, como se explica intuitivamente en la sección "¿por qué?".

El algoritmo de ML es sensible a las "escalas relativas de las características", lo que suele ocurrir cuando utiliza los valores numéricos de las características en lugar de decir su rango.

En muchos algoritmos, cuando deseamos una convergencia más rápida, el escalado es IMPRESCINDIBLE, como en las redes neuronales.

Dado que el rango de valores de los datos brutos varía mucho, en algunos algoritmos de aprendizaje automático, las funciones objetivo no funcionan correctamente sin normalización. Por ejemplo, la mayoría de los clasificadores calculan la distancia entre dos puntos por la distancia. Si una de las características tiene un amplio rango de valores, la distancia rige esta característica en particular. Por lo tanto, **el rango de todas las características debe normalizarse para que cada característica contribuya de forma aproximadamente proporcional a la distancia final.**

La regla general que podemos seguir aquí es que un algoritmo que calcula la distancia o asume la normalidad, escala sus características.

Algunos ejemplos de algoritmos en los que el escalado de características importa son:

03008915 C/ Ferrocaril, 22, 03570 La Vila Joiosa **Tel** 966870140 **Fax** 966870141 http://portal.edu.gva.es/iesmarcoszaragoza

K-nearest neighbors (KNN) con una medida de distancia euclidiana es sensible a las magnitudes y, por tanto, debe escalarse para que todas las características tengan el mismo peso.

K-Means utiliza la medida de distancia euclidiana y en este caso el escalado de las características es importante.

El escalado es crítico cuando se realiza el Análisis de Componentes Principales (PCA). PCA intenta obtener las características con la máxima varianza, y la varianza es alta para las características de alta magnitud y sesga el PCA hacia las características de alta magnitud.

Los algoritmos que no requieren normalización/escalado son los que se basan en reglas. No se verían afectados por ninguna transformación monótona de las variables. El escalado es una transformación monotónica. Ejemplos de algoritmos de esta categoría son **todos los algoritmos basados en árboles**: CART, Random Forests, Gradient Boosted Decision Trees. **Estos algoritmos utilizan reglas (series de desigualdades) y no requieren normalización.**

1.3 ¿Cómo escalar características?

A continuación, se presentan algunas formas en las que podemos realizar el escalado de características.

1) Escalador Min Max
2) Escalador estándar
3) Escalador Max Abs
4) Escalador robusto
5) Escalador por transformador cuantílico
6) Escalador por transformada de potencia
7) Escalador vectorial unitario

Para la explicación, formaremos un DataFrame para mostrar los diferentes métodos de escalado. Ver notebook Adjunto.

El escalado de características es un paso esencial en el preprocesamiento del aprendizaje automático. El aprendizaje profundo requiere el escalado de características para una convergencia más rápida, por lo que es vital decidir qué escalado de características utilizar. Existen muchos estudios comparativos de métodos de escalado para varios algoritmos. Sin embargo, **como la mayoría de los otros pasos del aprendizaje automático, el escalado de características también es un proceso de prueba y error**.