Método KDD - Daniel Marín López

1. Selección de datos

La selección consiste en seleccionar unos datos a priori que necesitamos para nuestro problema a abordar, es un paso crucial.
Los criterios de selección se dividen en su relevancia, homogeneidad, distribución equilibrada y calidad de los datos. En nuestro caso no sirven la edad, trabajo, lugar de trabajo, etc.

Se recomienda no procesar volúmenes grandes de datos completos sobre todo por el rendimiento y los costos que conlleva realizar estas operaciones, el sobreajuste que puede ocurrir en algunos modelos complejos de IA y el ruido que puede generar patrones irrelevantes o sesgados. Por ejemplo, si los datos de productos de ventas son pasados sin ningún tipo de filtro nos puede proporcionar predicciones incorrectas.

Todo esto nos puede llevar a resultados sesgados o incorrectos que no reflejan la realidad y una falta de generalización que no cubre toda la variedad en nuestro modelos. Por ejemplo, entrenar un modelo con clientes con un buen crédito positivo el modelo aprenderá muy bien solo en este sector.

El dominio de los datos debe proporcionar un contexto necesario para determinar qué variables y registros son verdaderamente reales. Conocer el problema nos lleva a una mejor toma de decisiones y comprensión de los datos.

Al final nos quedamos con un subconjunto de datos adecuados para el problema que estamos tratando, que son los datos objetivo.

2. Preprocesado

Este es el que hemos presentado.

3. Transformación

El proceso de transformación sigue lo siguiente:

1. Datos no etiquetados
2. Agrupación por similitudes
3. Etiquetado de clústeres
4. Clasificación

Las ventajas de reducir variables son:

- Enfoque preciso
- Expresión útil de los datos
- Simplificación de análisis
- Mayor eficiencia

El riesgo de reducción de variables lleva
- Una pérdida de información crítica
- Resultados irrelevantes
- Necesidad de datos relevantes

Los datos son etiquetados por expertos, métodos participativos.

En nuestro caso, en los préstamos los clústeres son
- Los buenos pagadores
- Los malos pagadores
- Boderline

En un sistema de recomendación son historial del cliente, navegación, -

Las etiquetas para nuestros clústeres serían:
- compradores frecuentes de electrónica
- en moda deportiva
- de alta gama
- etc.

Donde se aplica la práctica del clustering:
- análisis de similitudes
- recopilación de datos y comportamiento
- formación de clústeres
- opinión

Conclusión: De cluster a clasificación, reducción inteligente y etiquetado inteligente

4. Minería de datos

La minería de texto es la forma de tratar los datos en un modelo para detectar patrones y de esa manera hacer una toma de decisiones.

Los algoritmos son un conjunto de pasos finitos que su meta es detectar los patrones de nuestro conjunto de datos.

Los algoritmos son de dos tipos:
- Algoritmos de función (Caja Negra): Nos ofrecen un resultado sin saber como lo hacen, hay distintos tipo:
    - Clasificación
    - Regresión
    - Clustering
    - Análisis de secuencias
- Algoritmos de presentación (Caja Blanca):

La clasificación está dentro del aprendizaje supervisado al tener sus datos etiquetados y el clustering entra dentro del aprendizaje NO supervisado al no haber etiquetas.

La función resumen es una versión simplificada de los datos para una mejor interpretación.

El problema principal es el sobreaprendizaje de los datos en nuestro modelo.

El criterio de preferencia es seleccionar una serie de modelos hasta que UNO nos satisfaga el problema, existen distintas métricas como:

- Accuracy
- Recall
- F1-score

La elección del modelo depende del problema que queramos tratar, y para ello las métricas nos ayudarán a tomar esa decisión.

Algunos criterios para encontrar el mejor modelo son:

- Tipo de problema
- Tipo de datos
- Interpretabilidad
- Rendimiento
- Balance de clases
- Escalabilidad

El orden de los algoritmos seguiría un poco en base a la etapa anterior:

1. Clustering
2. Clasificación