TEMA 7

MINERÍA DE DATOS Y ANALISIS PREDICTIVO

Andrei Alexandru Miu

Índice

Conceptos básicos 3

Técnicas de modelización 3

Preprocesamiento de datos 4

Modelización y aprendizaje 4

Inferencia y aplicación de modelos 5

Casos de aplicación 5

Conceptos básicos

a) ¿Qué es la minería de datos y cuál es su principal objetivo en el análisis de Big Data?

Puede definirse como el conjunto de metodologías, procesos y tecnologías para el descubrimiento no trivial de información relevante, normalmente subyacente en grandes volúmenes de datos y su consiguiente aplicación e integración dentro de las operaciones del negocio con el fin de mejorar el rendimiento y soporte la toma de decisiones.
b) Explica la diferencia entre análisis descriptivo y análisis predictivo dentro del contexto de minería de datos.

El análisis descriptivo se enfoca en comprender y resumir los datos históricos, mientras que el análisis predictivo se basa en modelos estadísticos para prever eventos futuros.

Técnicas de modelización

a) ¿Cuál es la diferencia entre un modelo supervisado y un modelo no supervisado?

El aprendizaje supervisado se lleva a cabo mediante un mecanismo de inspección que permite evaluar su calidad en 2 aspectos:

- Como de bien detecta el patrón que subyace en los datos de partida
- Que capacidad tiene de generalizar ese patrón sobre nuevos datos

El aprendizaje no supervisado tiene como objetivo descubrir esas relaciones mediante la identificación de agrupaciones y coocurrencias en los datos de la forma que más desatendida posible.

Es decir, en el aprendizaje supervisado sabemos lo que estamos buscando, mientras que en el no supervisado no.

b) Clasifica los siguientes ejemplos dentro de modelos supervisados o no supervisados:

- Segmentación de clientes según patrones de compra.

  No supervisado

- Predicción de la probabilidad de que un cliente cancele su suscripción.

  Supervisado

- Detección de fraude en transacciones bancarias.

  Depende si se usa para detectar anomalías (no supervisado) o si se cuenta con datos históricos etiquetados (supervisado)

- Identificación de grupos de productos que suelen comprarse juntos.

  No supervisado

Preprocesamiento de datos

a) ¿Por qué es importante la etapa de preprocesamiento en un proyecto de minería de datos?

La etapa de preprocesado consiste en preparar los datos para las tareas de modelización posteriores. Se compone de una serie de fases de transformación cuyo objetivo es publicar, en el formato adecuado, una selección de observaciones tratadas y cuya calidad podamos asegurar.

Aunque todas estas fases tienen su importancia, la de filtrado y compresión es especialmente relevante. Su principal motivación es una reducción en el tamaño de los datos de cara a su posterior modelización.
b) Menciona al menos tres técnicas utilizadas en el preprocesamiento de datos y su utilidad.

- **Selección**: Acceso a los sistemas origen y extracción de datos.
- **Exploración**: Estudio de las características de los datos.
- **Limpieza**: Detección y corrección de problemas de calidad.

Adicionalmente, existen técnicas como:

- **Muestreo**: Permite disminuir el numero de observaciones mediante la extracción de un subconjunto representativo de estas.
- **Muestreo estratificado**: Se emplea cuando es necesario asegurar que la muestra mantendrá la misma proporción de ocurrencias respecto a uno o varios atributos de interés.

Modelización y aprendizaje

a) ¿Qué significa el término "sobreajuste" (overfitting) en un modelo de aprendizaje supervisado?

Un modelo que aprenda excesivamente bien los datos iniciales, pero pierda esa capacidad de generalización cuando se le presentan datos nuevos. A eso se le dice que esta sobre ajustado, mientras que aquel que ni siquiera tiene la capacidad de detectar patrón alguno en los datos iniciales está subajustado.

b) ¿Qué estrategias pueden utilizarse para evitar el sobreajuste?

Una de las estrategias podría ser el aumento del conjunto de datos, si se dispone de mas datos, el modelo aprende patrones mas generales y menos específicos.

Inferencia y aplicación de modelos

a) ¿Qué diferencia hay entre la inferencia por lotes y la inferencia en tiempo real en la puesta en producción de un modelo?

Inferencia por lotes con persistencia: El modelo es aplicado a un conjunto de observaciones en bloque. Los datos de entrada residen típicamente en un fichero o en una tabla de una base de datos, a donde van a parar también los resultados de la inferencia. Esta puede ejecutarse bajo demanda o bien de forma planificada mediante un proceso periódico. La latencia en la aplicación del modelo no es importante.

Inferencia por lotes sin persistencia: Caso similar al anterior con la salvedad de que no hay el requerimiento de persistir los datos. Cada vez que se accede los resultados son actualizados.

Inferencia en tiempo real: Este escenario se da principalmente cuando se desea aplicar, con una latencia mínima, un modelo a observaciones individuales que todavía no tiene persistencia. Estas acaban de ser generadas por una aplicación a la que hay que devolver los resultados en tiempo real para soportar una acción de negocio.

b) ¿Por qué es importante monitorear un modelo una vez que ha sido desplegado en producción?

Para garantizar su rendimiento y precisión a lo largo del tiempo, ya que con el tiempo los datos de entrada pueden cambiar y hacer que el modelo se vuelva menos preciso, incluso se puede enfrentar a valores atípicos, faltantes o fuera de rango, y eso hay que detectarlo y corregirlo antes.

Casos de aplicación

a) Explica cómo la minería de datos podría aplicarse en un hospital para mejorar la atención a los pacientes.

Puede mejorar la atención a los pacientes al analizar grandes volúmenes de información para identificar patrones, optimizar procesos y tomar decisiones basadas en datos, como, por ejemplo:

- Detección temprana de enfermedades y/o diagnósticos.
- Optimización del flujo de pacientes y la gestión de recursos.
- Detección de posibles fraudes o errores en registros.

b) Imagina que una empresa de e-commerce quiere mejorar sus recomendaciones de productos. ¿Qué tipo de modelo de minería de datos recomendarías y por qué?

Para un e-commerce, el mejor enfoque creo que sería un modelo híbrido, el cual combine filtrado colaborativo, basado en contenido y reglas de asociación, ya que permite recomendaciones más precisas y personalizadas, mejorando así la experiencia del usuario y aumentando las ventas.