Tercer Examen Parcial Minería de Datos Jonathan Emmanuel Hernández Ortiz
Fecha Entrega: 14/11/2024 420054834

1. **¿Cuáles son los desafíos que se tienen que enfrentar en el proceso de Minería de Datos ¿ y que tipo de iniciativas ayudarían a minimizar el impacto?**

Los principales desafíos en el manejo de datos incluyen trabajar con información desordenada, gestionar grandes volúmenes, proteger la privacidad, elegir el modelo adecuado y entender los resultados. Para superarlos, es clave organizar los datos, usar herramientas avanzadas, garantizar la seguridad y comunicar los hallazgos claramente. Aplicar algoritmos estudiados en clase, puede mejorar la precisión y cumplir los objetivos de la Minería de Datos.

2. **El Entrenamiento y Verificación Seleccione una opción:**

Las dos anteriores

3. **Indique 3 características principales de los algoritmos de Clustering: Mezcla Gaussiana y Jerárquico.**

Mezcla Gaussiana

- Agrupa datos basándose en distribuciones estadísticas.
- Permite solapamiento entre grupos.
- Adecuado para datos con formas no lineales.

Jerárquico

- Crea una estructura de árbol.
- No requiere número inicial de grupos.
- Útil para entender relaciones entre grupos.

4. **En los algoritmos de Asociación ¿cómo se calcula la métrica se "Support" y la de "Confidence"?**

Support:

Se refiere a la frecuencia con la que un conjunto específico de ítems aparece en las transacciones. Para calcularla, se divide el número de transacciones que incluyen dicho conjunto de ítems entre el total de transacciones realizadas.

$$Soporte = \frac{Transacciones\ del\ conjunto}{Total\ de\ transacciones}$$

Confidence:

Se trata de la probabilidad de que un ítem esté presente en una transacción dado que otro ítem ya está incluido. Para calcularla, se divide el soporte del conjunto de ítems A y B entre el soporte del conjunto A.

$$Confianza = \frac{Soporte\ del\ conjunto\ A - B}{Soporte\ del\ conjunto\ A}$$

5. **Los algoritmos de Random Forest tienen 3 hiperparámetros principales que deben configurarse antes del entrenamiento, ¿cuáles son?**

Número de árboles (n_estimators):

Define cuántos árboles compondrán el bosque. Un mayor número suele mejorar la precisión, pero aumenta el tiempo de entrenamiento.

Profundidad máxima de los árboles (max_depth):

Limita la altura de los árboles para evitar el sobreajuste.

Número de características por división (max_features):

Especifica cuántas variables se consideran en cada división, equilibrando precisión y velocidad.

6. **Seleccione 2 técnicas que puedes utilizar para eliminar el Overfitting.**

Dropout Regularization:

En redes neuronales, su objetivo es desactivar de forma aleatoria algunas conexiones durante el entrenamiento. Esto ayuda a prevenir el sobreajuste, permitiendo que el modelo generalice mejor en datos no vistos.

Noise Injection:

Añadir ruido a los datos de entrenamiento es una estrategia que fuerza al modelo a generalizar mejor. Esto lo hace más robusto ante pequeñas variaciones o inconsistencias en los datos, mejorando su capacidad de adaptación a nuevos escenarios.

7. **Describa las características de Error Cuadrático medio, Error Absoluto medio y Error Cuadrático Relativo.**

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RSE = (\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)/(\sum_{i=1}^{n}(y_i - \overline{y})^2$$

8. **Describa las características de la técnica del Train-Test Split y el Método K-Folds.**

El **Train-Test Split** es una técnica que separa el conjunto de datos en dos partes: una para entrenamiento (usualmente entre el 70-80%) y otra para prueba (20-30%). La parte de entrenamiento se utiliza para ajustar el modelo, mientras que la de prueba sirve para evaluarlo. Es un método simple y rápido, aunque puede ser menos representativo si el conjunto de datos es pequeño.

Por otro lado, el **K-Folds** divide los datos en k partes iguales. El modelo se entrena y evalúa k veces, utilizando en cada iteración k−1 partes para entrenamiento y 1 parte distinta para prueba. Este método proporciona una evaluación más precisa y generalizable, especialmente cuando se trabaja con conjuntos de datos limitados.

9. **Describa las principales características del Principio de Longitud Mínima.**

Nos dice que, en un conjunto de datos, el mejor modelo es aquel que logra la mejor compresión de los datos. Busca el equilibrio entre la complejidad del modelo y su capacidad para explicar los datos, eligiendo el modelo que minimiza la longitud combinada de la descripción del modelo y la de los datos cuando se codifican con dicho modelo. Es útil para evitar el sobreajuste, ya que penaliza modelos excesivamente complejos que no aportan mejoras significativas en la compresión de los datos.

10. **Ya se realizaron pruebas de los algoritmos de minería con datos históricos y se obtuvieron resultados interesantes para el cliente, detalle las actividades siguientes que se deberían de realizar para concluir un proyecto de Minería de Datos.**

Para concluir un proyecto de Minería de Datos, se debe: validar los resultados, interpretar y analizar los hallazgos, elaborar un informe final, implementar el modelo en producción, monitorear y ajustar su desempeño y realizar una presentación al cliente con conclusiones y recomendaciones.

11. **¿Qué actividades y qué técnicas de Minería de Datos utilizarías para optimizar el uso de la información del cliente cuando te diga que tiene 5 años de historia de transacciones que equivaldría a 30 millones de registros**

y además necesitarás considerar archivos clave como Clientes, Inventarios, Costos, etc.?

Para optimizar la información del cliente, primero limpiaría los datos eliminando valores atípicos y duplicados, luego realizaría un análisis exploratorio para identificar patrones y correlaciones. Usaría técnicas de clustering como **K-means** para segmentar clientes, regrsión logística o **Random Forest** para modelos predictivos, y **Apriori** para reglas de asociación entre productos. Finalmente, aplicaría análisis de series de tiempo para predecir demanda y ajustar inventarios, aprovechando los datos históricos de transacciones e inventarios.

12. **Crear un diagrama de arquitectura de infraestructura para Minería de datos donde se muestre tecnologías como Big Data, capacidades Real Time, Machine Learning y ambiente Cloud.**

Contestado en la hoja de examen

13. **Durante la fase de entendimiento de Negocio, ¿qué información considerarías para evaluar los siguientes temas:**

a) Supuestos y Limitantes

Calidad y disponibilidad de datos, recursos tecnológicos, tiempo, presupuesto y restricciones legales.

b) Riesgos y Contingencias

Datos deficientes, fallos técnicos, falta de experiencia; mitigados con planes de respaldo, capacitación y ajustes presupuestarios.

14. **Se tiene una reunión con el gerente del área de Ventas y marketing en una empresa como el Palacio de Hierro interesado en el proceso de Minería de Datos, a) Indique qué preguntas le realizarías para conocer su requerimiento, y b) qué comentarios le darías para que tenga confianza en el proceso.**

a)

1. ¿Qué objetivos clave busca alcanzar con el análisis de datos, como mejorar la segmentación o aumentar la conversión de clientes?
2. ¿Qué tipo de datos están disponibles (ventas, clientes, inventarios) y cómo se almacenan actualmente?

Esto para saber su posible uso o cómo podemos atacar el problema.

b)

1. Este proceso le permitirá identificar patrones valiosos para optimizar estrategias y mejorar resultados.
2. Garantizamos que toda la información será tratada con confidencialidad y se adaptará a sus necesidades específicas.

15. Se tienen el registro de 10,000 pacientes que fueron sometidos a una prueba para conocer si tienen algún tipo de diabetes, ¿qué algoritmos predictivos se propondrían para realizar el estudio de impacto, y qué técnica de validación se propondría?

Algoritmos Predictivos:

1. **Regresión Logística:** Ideal para predecir resultados binarios (diabetes: sí o no).

2. **Random Forest:** Ofrece precisión al manejar conjuntos de datos grandes y complejos, identificando patrones relevantes.

Técnica de Validación:

K-Fold Cross Validation: Divide los datos en varios subconjuntos para entrenar y validar el modelo, asegurando resultados confiables y minimizando el sobreajuste.