Examen IA

Santiago Nicolas Jadan Mora

Biomedicina

1. **¿Cuáles son los 3 tipos de machine learning?**

- Aprendizaje supervisado.
- Aprendizaje no supervisado.
- Aprendizaje de refuerzo según la naturaleza de los datos que recibe.

2. **¿Qué es el sobreajuste de un modelo, y porqué se puede producir?**

El sobreajuste es un concepto en la ciencia de datos, que ocurre cuando un modelo estadístico se ajusta exactamente a sus datos de entrenamiento. Cuando esto sucede, el algoritmo desafortunadamente no puede funcionar con precisión contra datos invisibles, frustrando su propósito. La generalización de un modelo a nuevos datos es, en última instancia, lo que nos permite usar algoritmos de machine learning cada día para hacer predicciones y clasificar datos.

3. **¿Para evitar el sobreajuste existen diferentes técnicas, intenta explicar 3 posibles métodos para arreglarlo?**

• Métodos de Ensamble: Los métodos de aprendizaje por ensamble consisten en combinar múltiples clasificadores, como árboles de decisión, y fusionar sus predicciones para obtener un resultado más preciso. Los dos enfoques más conocidos son el empaquetado y el refuerzo. En el empaquetado, se generan muestras aleatorias con reemplazo del conjunto de entrenamiento, y varios modelos se entrenan de forma independiente. Luego, se promedian o se toma la mayoría de las predicciones para obtener una estimación más robusta, útil para reducir la variación en datos ruidosos.

• Entrenar con Más Datos: Aumentar el tamaño del conjunto de entrenamiento al incluir más datos puede mejorar la precisión del modelo, permitiéndole analizar mejor las relaciones entre las variables de entrada y salida. Sin embargo, este enfoque es más efectivo al agregar datos limpios y relevantes. Si se incluyen datos irrelevantes o complejos, el modelo puede sobreajustarse y perder generalización.

• Selección de Características: La selección de características implica identificar las variables más importantes en el conjunto de entrenamiento y eliminar las redundantes o irrelevantes. Esto ayuda a simplificar el modelo y destacar las tendencias dominantes en los datos, pero debe diferenciarse de la reducción de dimensionalidad, ya que no altera las características originales, solo las selecciona.

Supongamos que estamos investigando la cromatografía de gases acoplada a espectrometría de masas para obtener datos de la concentración de los distintos componentes de una nueva droga la cual está causando efectos negativos en la sociedad,

4. **¿Con qué tipo de ML se podría calibrar el equipo?**

Con una regresión se podría relacionar lo que el equipo obtuvo con los valores de compuesto que tenga la droga. Para hacer eso debería crearse un conjunto de entrenamiento con los valores conocidos del compuesto y las señales del equipo.

5. **Hemos identificado los picos de interés, resulta que se basa en el LSD, entre otros componentes, cuales se han identificado mediante comparativa entre bases de datos de drogas. ¿Con qué técnica**

podríamos observar los distintos tipos de componentes a través de los componentes identificados? Pista: tenemos demasiadas variables y no podemos relacionarlas entre sí una por una.

En este caso de que existen muchos datos, lo óptimo seria realizar un PCA como primer paso. Con esto podremos de cierta manera analizar menos datos o grupos más cerrados, se puede visualizar y analizar mejor la estructura y patrones presentes en los componentes identificados.

6. **Con cuáles de los siguientes algoritmos podríamos clasificar los componentes de la droga, ¿teniendo en cuenta que se tiene un banco de datos con diversas drogas y sus perfiles moleculares?: regresión logística binomial, regresión logística multinomial, regresión logística multinomial regualraizada, knn, SVM, regresión lineal múltiple, k-means, k-medianas, regresión por mínimos cuadrados parciales, regresión por componentes principales. Explica qué tipo de ML son y porque sí y porqué no.**
   Pista: Hemos realizado un análisis de correlación de Kendall (debido a que los residuos de la regresión parcial entre cada metabolito no cumplían la normalidad de los datos)

   - Regresión Logística Binomial: Este método no se puede emplear ya que sirve cuando existen dos posibles clasificaciones. En este caso no son solo dos.
   - Regresión Logística Multinomial: Si pudiera ser útil en este caso, porque a diferencia del anterior este relaciona varias variables del sistema.
   - Regresión Logística Multinomial Regularizada: También podría servir este método, dado que funciona igual que el anterior, con la diferencia de que este la regulariza para evitar un sobreajuste.
   - KNN (K-Nearest Neighbors): Es bueno para problemas de clasificación, podría utilizarse para clasificar los componentes de la droga según los componentes de otros en la base.
   - SVM (Support Vector Machine): Es bueno también para problemas de clasificación multiclase y podría ser útil para clasificar los componentes de la droga basándose en la separación óptima entre las diferentes clases en el espacio de características.
   - K-Means y K-Medias: Estos algoritmos son utilizados para agrupar dato, pero no se deberían de usar para clasificación multiclase, ya que no proporcionan etiquetas para cada una.
   - Regresión por Mínimos Cuadrados Parciales (PLS): Es una técnica de regresión que puede utilizarse para modelos de clasificación multiclase.
   - Regresión por Componentes Principales (PCR): Es otra técnica de regresión que no es adecuada para clasificación multiclase, ya que está diseñada para problemas de regresión.

Los algoritmos más adecuados serían la Regresión Logística Multinomial, la Regresión Logística Multinomial Regularizada, KNN y SVM. Cada uno de estos algoritmos puede manejar problemas de clasificación multiclase y proporcionar una etiqueta para cada componente de la droga en función de sus componentes.

El análisis de correlación de Kendall realizado previamente ayuda a identificar relaciones y patrones importantes entre las variables antes de aplicar los algoritmos de clasificación.

7. **¿Qué tienen que ver los residuos con la normalidad en ciertos modelos? ¿Qué tiene que ver la función objetivo con los residuos de un modelo?**

Estudiar el comportamiento de los residuos $u_i = Y_i - \hat{Y}_i$ es de vital importancia para el análisis de regresión, pues varios de los supuestos del Modelo Clásico de Regresión Lineal (MCRL) hacen énfasis en los residuos, es por esto por lo que se recurre a herramientas que nos permitan verificar si se cumplen estos supuestos y así, aumentar la confiabilidad sobre las conclusiones que se hagan a partir del modelo planteado.

La función objetivo en un modelo matemático es la expresión que se busca optimizar, es decir, minimizar o maximizar. En el contexto de modelos matemáticos que involucran residuos, la función objetivo puede estar relacionada con la minimización de costos, la maximización del ahorro del impacto ambiental y la maximización de la satisfacción al cliente.

8. **Siguiendo la misma línea, hay veces que un modelo se supone que es normal e independiente. Intenta explicar con 3 ejemplos de algoritmos que cumplan estos supuestos.**

a) Regresión Lineal: Se asume que los residuos siguen una distribución normal con media cero y varianza constante. La normalidad de los residuos es verificada mediante gráficos QQ y pruebas estadísticas como la prueba de Shapiro-Wilk. La independencia de los residuos se puede verificar mediante la inspección de gráficos de autocorrelación de los residuos.

b) Regresión Logística: En el caso de la regresión logística, que se utiliza para problemas de clasificación binaria, se asume que los residuos siguen una distribución binomial. La normalidad de los residuos es menos relevante en este contexto, pero se debe verificar la independencia de los residuos para asegurarse de que no exista dependencia serial en los datos.

c) ARIMA (AutoRegressive Integrated Moving Average): Es un modelo para el análisis de series de tiempo. En ARIMA, se supone que los residuos siguen una distribución normal, y la independencia de los residuos es esencial para asegurar que no haya patrones de dependencia temporal no capturados por el modelo.

9. **A veces se aplican a los anteriores algoritmos unos parámetros extra, modificándolos, ayudando en ciertas violaciones a la independencia de las variables ¿Cuáles son las 3 modificaciones que se le pueden añadir a un modelo de regresión lineal múltiple con tal de evitar lo anterior?**

    - Transformación de variables: Aplicar transformaciones como logaritmos o raíces cuadradas a las variables puede estabilizar la varianza de los residuos y reducir la dependencia entre las variables.
    - Selección de características: Seleccionar cuidadosamente las variables más relevantes en lugar de incluir todas las disponibles puede ayudar a eliminar la multicolinealidad y mejorar la independencia.
    - Regularización: La Regresión Ridge y la Regresión LASSO son técnicas de regularización que introducen penalizaciones para evitar el sobreajuste y mejorar la generalización del modelo, reduciendo la dependencia entre variables.

10. **¿Qué es conjunto de entrenamiento, prueba y validación del modelo, cuál porcentaje pondrías en cada conjunto? (Supongamos que tenemos que crear un dispositivo con nariz artificial, con alguno de los métodos que has respondido, se ha realizado en Python, y tiene que pasar 3 fases relacionadas con el anterior)**

    - Conjunto de entrenamiento: se emplea para adiestrar el modelo, ajustando sus parámetros y aprendiendo patrones de los datos. Aquí, el modelo establece relaciones entre las características de entrada y las etiquetas de salida.
    - Conjunto de prueba: se utiliza para evaluar el rendimiento del modelo después de haber sido entrenado con el conjunto de entrenamiento. Contiene datos que el modelo nunca ha visto previamente, permitiendo medir su capacidad de generalización a datos nuevos y desconocidos. El rendimiento en el conjunto de prueba proporciona una estimación de cómo se comportará el modelo en situaciones reales.
    - Conjunto de validación: es empleado para ajustar hiperparámetros y validar el rendimiento final del modelo. Tras entrenar el modelo con el conjunto de entrenamiento y ajustar los hiperparámetros utilizando el conjunto de validación, se evalúa su rendimiento final en el conjunto de prueba.

El porcentaje de estos conjuntos puede variar según el tamaño y características del conjunto de datos. Generalmente se usa aproximadamente el 70-80% de los datos al conjunto de entrenamiento, y el resto se divide entre el conjunto de validación y el conjunto de prueba, asignando aproximadamente un 10-15% a cada uno de ellos.

11. **El equipo que estamos trabajando, no ha detectado bien, la señal es muy mala en algunas partes, a pesar del procesado que se hizo. Por lo tanto, se decide asignar valores faltantes. ¿Qué dos métodos podríamos aplicar a este problema?**

- Imputación por Modelado: Utilizar información de otras variables para predecir los valores faltantes. Se construye un modelo basado en datos completos y luego se aplica para predecir los valores faltantes utilizando técnicas como regresión lineal, regresión logística o k-vecinos más cercanos (k-NN).
- Imputación por Media o Mediana: Reemplazar los valores faltantes con el promedio o la mediana de la variable respectiva. Es una estrategia simple y efectiva, adecuada para datos no normales o con valores atípicos.

12. **Intenta explicar en qué consiste la matriz de confusión en ML, en especial por qué es tan importante.**

La matriz de confusión de un modelo es muy importante debido a que nos proporciona la información para saber el rendimiento de un modelo de clasificación para así poder interpretarla. Existen cuatro parámetros clave en la matriz de confusión que son.

- Verdaderos positivos
- Falsos positivos
- Verdaderos negativos
- Falsos negativos

Estas métricas ayudan a evaluar el rendimiento del modelo y comprender cómo está realizando la clasificación en cada clase.

13. **En qué consiste el ML no supervisado, y cómo nos podría ayudar a encontrar nuevos patrones en la droga.**

El aprendizaje automático no supervisado es un tipo de aprendizaje automático en el que el algoritmo busca patrones en los datos sin tener etiquetas predefinidas. A diferencia del aprendizaje supervisado, donde el algoritmo aprende a partir de ejemplos etiquetados, en el aprendizaje no supervisado el algoritmo debe encontrar patrones y relaciones en los datos por sí mismo.

En la búsqueda de nuevos patrones en la droga, el aprendizaje no supervisado podría sería útil para identificar grupos o clusters de datos similares, lo que podría indicar la presencia de patrones o tendencias en los datos. Por ejemplo, se podría utilizar un algoritmo de clustering para agrupar drogas con propiedades químicas similares y así identificar nuevas relaciones entre ellas.