Exámen Parcial

Parte Teórica

Estudiante Dennis Xiloj

1. ¿Qué es Machine Learning?

El ML es el conjunto de herramientas con base matemática y parte del campo de la Inteligencia Artificial que se basa en encontrar y utilizar los patrones que aparecen en los datos para elaborar predicciones, se le denomina Machine Learning debido a que la máquina identifica estos patrones sin ser explícitamente programada para tal acción.

2. Liste los tipos de aprendizaje de máquina

- Supervisado
- No supervisado

3. ¿Cuál es la diferencia entre AI y ML?

Ambas son disciplinas de las ciencias computacionales y las matemáticas, sin embargo la IA es un área mucho más amplia, ML es parte de la IA y se enfoca en todo el conjunto de algoritmos y programación para el aprendizaje automatizado. La IA por su parte busca simular y duplicar todos los elementos cognitivos del ser humano en general.

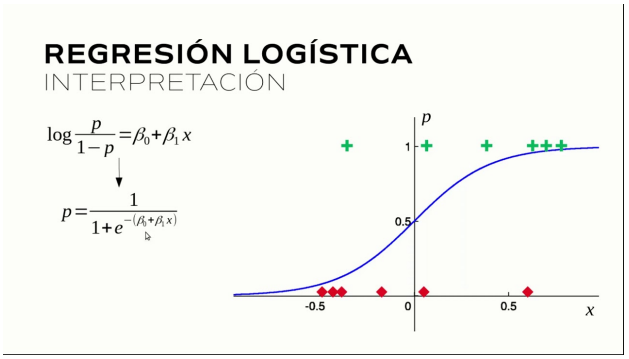
4. Describa brevemente cómo se realiza un modelo de predicción para ML

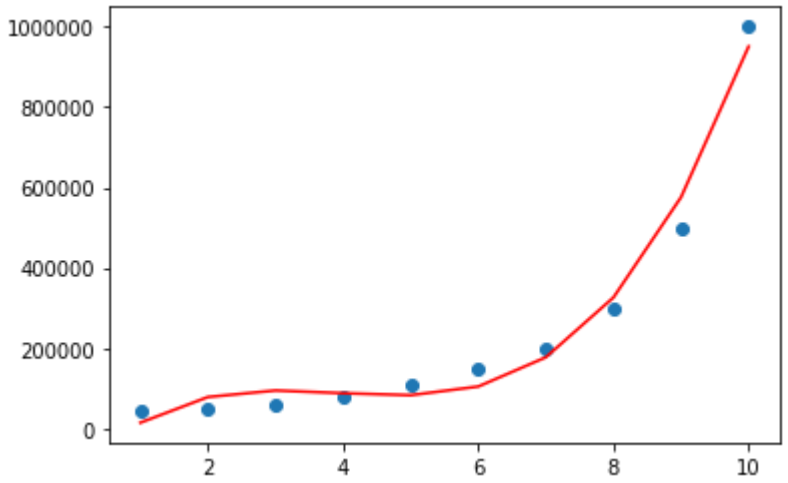
En general siempre se sigue un set de pasos, común a todo proceso de entrenamiento de un modelo:

1. Captura de información
    a. Se busca obtener toda la información necesaria relacionada al modelo que se desea crear
    b. Puede ser de cualquier fuente, siempre que existan suficientes datos
2. EDA
    a. Se realiza una exploración de datos general, en busca de inconsistencias y elementos faltantes
    b. Se identifican las variables y la existencia de suficientes datos
3. Limpieza de datos
    a. Se completan datos si es necesario
    b. Se ordenan las columnas y se generan nuevas si fuera necesario
4. Entrenamiento y elección del algoritmo
    a. Elección de los algoritmos más adecuados
    b. Preparar datos
        i. Se transforman las variables al formato necesario para que el modelo pueda entenderlas.
        ii. Se separan los datos de entrenamiento y de test

c. Entrenamiento de los modelos
    i. Selección de features
    ii. Entrenamiento y tunning del modelo
        1. Selección de features
        2. selección de hiperparámetros
    iii. Evaluación de acuerdo a datos de prueba
d. Elección e implementación final del modelo

5. Describa dos algoritmos de ML y dos de sus aplicaciones.

1. Regresión Logística:
    - Es un algoritmo de aprendizaje supervisado que se utiliza para predecir entre dos resultados discretos en función a los predictores.
    - 

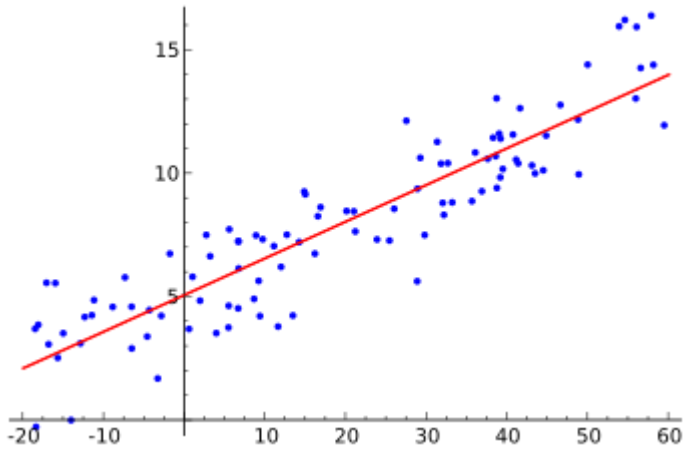
    - Aplicación: Se puede usar en escenarios donde el resultado buscado es un sí o no, verdadero o falso, bueno o malo. Por ejemplo al determinar si un crédito será satisfactorio o no, o según varios estudios un paciente tiene o no una enfermedad.
2. Regresión polinómica:
    - Este tipo de algoritmo se usa para predecir el valor de una variable contínua a través de un polinomio de grado n.
    - 

    - Aplicación: Este algoritmo se puede usar sobre cualquier variable de naturaleza contínua, como el ritmo de crecimiento de un árbol según su edad, la producción de litros de leche en una granja, etc.

6. ¿Cómo se calcula los θs en una regresión lineal?

La estimacion de los parametros se puede calcular por el método de mínimos cuadrados, que consiste en hallar los valores B0 y B1 que hacen mínima la suma de los cuadrados de las desviaciones entre los valores observados de la variable dependiente, y los valores estimados de la misma, de forma que el error entre los valores estimados sea siempre la misma a lo largo de la línea formada por la función resultante:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon$$



7. ¿Qué es "ACCURACY" en un modelo de ML?

Es la razón entre el número de predicciones correctas entre el total de muestras y es un número entre 0 y 1, siempre se busca que sea el mayor posible.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

8. ¿Cómo funciona K-FOLD CROSS VALIDATION?

Es un algoritmo iterativo que divide los datos entre k-grupos, luego se entrena al algoritmo usando uno de los grupos como prueba y el resto como datos de entrenamiento, este proceso se repite k-veces, k-*folds*. Este proceso genera k estimaciones cuyo promedio se usa como modelo final.

9. ¿Cómo se determina si el problema es lineal o no lineal en algoritmos de ML?

- Cuando son solo dos variables a relacionar, se puede comprobar graficando, creando un plot de las variables que queremos comparar
- Cuando son mas de dos variables, se debe calcular el coeficiente de correlación entre las variables a evaluar, si el resultado es mayor a 0.7 para todas las variables, nos indica que la relación es lineal.

10. ¿Cómo se selecciona el "K" correcto para el algoritmo K-Nearest?

Usualmente se hace mediante fuerza bruta, iterando sobre diferentes valores de k y comparando los resultados mediante algún scorer, usualmente accuracy, para al final elegir el mejor de todos.

11. Usted está interpretando el resultado de un algoritmo de ML que categoriza pacientes de enfermedades terminales, la matriz de confusión devuelve el valor TP con 8 y FN con 15, ¿qué podría concluir del modelo?

	Predecidos verdaderos	Predecidos falsos
En realidad verdadero	8	
En realidad falso	15	

Con estos valores podemos concluir que no es un buen modelo ya que está identificando como positivo a muchos más casos que son en realidad falsos vs los que en realidad son positivos. calculando la sensibilidad sería: TP/(TP+FN) = 0.35, un valor muy bajo.

12. Usted está interpretando un modelo de ML que tiene un F1-SCORE de 1, ¿qué podría concluir del modelo?

El F1-Score con valor 1 representa que el modelo es perfecto, sin embargo esto es muy poco probable en la práctica, este valor en realidad nos podría indicar que el modelo está extremadamente sobre-entrenado o tomamos la misma variable dependiente dentro del arreglo de predictores, por lo que se debe evaluar muy detenidamente.

13. ¿Utilizaría un Decision Tree para obtener un resultado predictivo y asi clasificar a la variable dependiente? Justifique su respuesta.
Depende del caso, no es necesariamente el mejor o el peor, se debe evaluar en cada caso específico de acuerdo a los valores esperados

14. Explique los indicadores obtenidos de la matriz de confusión.

1. True Positive (TP): cuántos valores son realmente verdaderos y se predijeron verdaderos.
2. False negativo (FN): cuántos valores reales son positivos y la predicción dice que son negativos.
3. Falso verdadero (FP): El valor real es negativo y la prueba predijo positivo.
4. Falso falso (TN): El valor real es negativo y la prueba predijo negativo.