Hoja de Ejercicios y Problemas tipo examen 2

Ismael Sagredo Olivenza

Colaboración en este material de Belén Díaz Agudo y otros profesores de ISIA

Problema 01

Tenemos un dataset con datos almacenados sobre las precipitaciones desde hace 10 años y queremos predecir el nivel de precipitación en el futuro en función del dia del año (entero), la humedad (real), la cantidad de nubosidad (baja, media, alta), la temperatura (real) y la presión atmosférica (real).

Hemos aplicado diferentes técnicas de Machine Learning y nos ha generado los siguientes resultados de accuracy:

- KNN: K=5, 88% distancia euclídea.
- Decision Tree: 90%, profundidad del árbol 5.
- Random Forest: 98%
- Perceptrón Multicapa: 10 neuronas ocultas en la primera capa y 5 en la segunda, factor de aprendiszaje de 0,2, valores normalizados 0-1: 90%

A los investigadores les ha sorprendido que el perceptrón multicapa de resultados tan bajos.

1. ¿Qué podriamos hacer para mejorar los resultados del perceptrón?
2. Con los datos del ejercicio, razona cual sería el modelo que elegirías.
3. En KNN el valor categórico ha sido codificado como 0, 1, 2 y el resto de atributos para calcular la diferencia se realizan restando sus valores sin ninguna transformación. ¿Cómo podríamos mejorar KNN para conseguir mejores resultados?

Problema 02

En un problema de Machine Learning, hemos utilizado la librería train_test_split para dividir el conjunto de entrenamiento y de test. En concreto se ha utilizado con los siguientes valores train_test_split(data, y, train_size = 0.8, random_state = 1)

Sin embargo, haciendo pruebas vemos que en función del random_state o del tamaño de los datos de entrenamiento el resultado varia hasta un 5% de precisión.

1. ¿Cómo podríamos abordar este problema?

Problema 03

Queremos detectar a partir de imágenes de rayos X si existe o no un tumor cerebral. Para ello disponemos de un dataset con 100.000 imágenes de 1024x1024 en escala de grises.

1. Explica los pasos que harías para preparar los datos
2. A priori, cual crees que sería el mejor modelo y porqué.
3. Suponiendo que se han creado dos modelos y conm 2000 datos de validación, el modelo A y el modelo B

La matriz de confusión de A es:

	Tumor	**No-tumor**
Tumor	800	200
No-Tumor	100	900

La matriz de confusión de B es:

	Tumor	**No-tumor**
Tumor	900	100
No-Tumor	200	800

¿Qué modelo elegirías?

Problema 04

Hemos conectado al videojuego Tetris un módulo que captura los eventos del mando, que se van registrando a la vez que se va haciendo una foto del estado del jugador. La idea es construir un sistema que permite jugar de forma autónoma a un agente que debe jugar igual que el jugador para el modo versus, donde un jugador juega contra una IA. Para ellos almacenamos en un histórico el estado del juego y la acción que ha realizado el jugador que puede ser una de las siguientes:

- Mover Izquierda (MI)
- Mover derecha (MD)
- Rotar pieza (R)
- Bajar pieza (B)
- Guardar pieza (G)

El estado está compuesto por una matriz con las celdas ocupadas por piezas, el nivel actual, la posición de la pieza que cae actualmente, la siguiente pieza y la pieza almacenada de repuesto.

Si queremos utilizar KNN para seleccionar la acción más similar a la realizada por el usuario….

1. ¿Como codificarías los datos para que KNN pueda funcionar correctamente y que medida de distancia usarías?
2. Para saber si el agente imita correctamente al jugador, comprobamos que la secuencia de acciones realizadas por el agente sea similar a la del jugador para un estado de juego dado (El Tetris es determinista asumiendo la misma semilla aleatoria) De esta forma tendremos la siguiente secuencia de acciones:

MI|MD|R|B|MI|MI|B|G|MD|MI|MD|R|R|B

¿Qué medida de distancia utilizarías para calcular el error?

Problema 05

Dada la siguiente tabla con los datos de entrenamiento de un algoritmo ID3 y utilizando la Ganancia y la Entropía

$$Entro(s) = \sum_{n=1}^{c} -p_i * Log_2 p_i$$

$$Gan(S, A) = Entropia(S) - \sum_{v \in V(A)}^{|A|} \frac{|sv|}{|s|} Entro(S_v)$$

Donde p_i es la probabilidad del atributo i-ésimo, S es el conjunto de todos los ejemplos y A el conjunto de todos los atributos y V(A) los valores de todos los atributos.

A1	A2	A3	A4	Clase
2	0	3	A	Bueno
1	1	1	B	Malo
1	2	0	B	Bueno
0	2	1	B	Bueno
2	1	0	A	Malo
3	0	1	B	Malo

1. Calcular cuál será el atributo que primero elegirá ID3

Problema 06

Si quiero extraer las características de una imagen para poder aplicar un algoritmo de ML sobre una versión reducida de dicha imagen ¿Qué técnicas puedo aplicar y que ventajas e inconvenientes tiene cada una?

Problema 07

A priori (ya que no se está totalmente seguro hasta no implementarlos), ¿Qué algoritmo clasificaría mejor un problema que no es linealmente separable? Explica porqué.

- Un árbol de decisión
- Una Red de neuronas profunda compuesta únicamente por capas convolucionales.
- Un Perceptrón multicapa con función de activación ReLu

Problema 08

Tenemos un problema de clasificación que tiene 5 clases. Queremos diseñar una red de neuronas, pero no tenemos claro cual será el modelo a elegir. Lo que si queremos es que la salida de la red nos devuelva la probabilidad de que un elemento pertenezca a cada una de las clases. Describe como sería la capa de salida de la red neuronal, cuantas neuronas de salida y de que tipo sería.

Problema 09

1. ¿Dónde aplicarías clustering? Pon un ejemplo.
2. Si utilizas K-Means, ¿Cómo podrías determinar el valor de K?

Problema 10

Dado los siguientes datos:

A1	A2	A3	A4
2	0	3	2
1	1	1	0
2	0	2	2
0	2	1	1
2	1	0	1
2	0	3	0

Describe que clases generaría un algoritmo de clustering aglomerativo con distancia de Manhattan

---

Problema 11

Tenemos un NPC en un videojuego de lucha y queremos que aprenda a luchar de forma eficiente contra una IA ya programada. No disponemos de un dataset con información de partidas previas. Qué algoritmo utilizarías y describe brevemente y de forma aproximada como lo implementarías.

Problema 12

¿Qué podemos hacer si el espacio de estados de un algoritmo de Q-Learning es demasiado grande? Razona la respuestae indica si se te ocurre al menos dos posibles soluciones.