Padrón: _____ Nombre y Apellido:_____

Parcial 27-10-2022

Ejercicio 1

Indique qué método de **reducción de la dimensionalidad** (PCA, MDS, t-SNE, ISOMAP) aplica mejor según el caso:

a) Se sospecha que los datos están ubicados en una variedad, dentro del espacio de alta dimensionalidad:_____
b) Se privilegia preservar las distancias entre puntos, incluso para métricas no euclidianas:_____
c) Se requiere entender cuán dispersos están los datos y sobre todo, sobre qué ejes o variables:_____
d) Se requiere proyectar un espacio de alta dimensionalidad en tan solo 2 dimensiones, manteniendo los clusters del conjunto original:_____

Ejercicio 2

Indique si las siguientes afirmaciones sobre sobre **SVM** son verdaderas o falsas (V / F) :

a) SVM es un algoritmo que permite buscar un *Support Vector Classifier* (un clasificador de margen blando) en un espacio de dimensión más pequeña, que el espacio de entrada original:_____
b) SVM, a través de uno de sus Kernels es capaz de encontrar soluciones en un espacio de dimensiones infinitas:_____
c) SVM **no** realiza un mapeo real del espacio de entrada al nuevo espacio, sino que solo computará las relaciones entre pares de observaciones como si estuviesen en el nuevo espacio, lo que comúnmente se llama, *kernel trick*: _____
d) La función Kernel polinómica, necesita un parámetro **d**, llamado grado del polinomio que indica el grado más pequeño del polinomio a usar:_____

Ejercicio 3

Determinar si las siguientes afirmaciones sobre **K-means** son verdaderas o falsas:

a) Es un algoritmo no supervisado de Clustering ya que partimos de un conjunto de datos etiquetado previamente. _____
b) El parámetro K corresponde a la cantidad de clusters o grupos que se formarán._____
c) El parámetro K corresponde a la cantidad de centroides que se calcularán._____
d) Es un algoritmo supervisado de Clustering ya que partimos de un conjunto de datos etiquetado previamente. _____

Padrón: _____ Nombre y Apellido:_____

Ejercicio 4

Indique a qué conjunto (CRUZ o CIRCULO) pertenece el punto negro según el parámetro K en el modelo de **K-Nearest-Neiboogrs**:

a) K = 1 _____
b) K = 4 _____
c) K = 8 _____

Ejercicio 5

Indique si las siguientes afirmaciones sobre **AdaBoost** y **Random Forest** son verdaderas o falsas (V/F):

a) AdaBoost se entrena mediante la técnica de Bagging:_____
b) En AdaBoost todos los árboles son completos mientras que Random Forest sólo entrena tocones (árboles con un solo nodo):_____
c) En Random Forest cada árbol vota según su peso relativo:_____
d) En AdaBoost se utiliza la técnica de Bootstrap aggregating para partir el dataset de entrenamiento en N datasets más pequeños:_____

Ejercicio 6

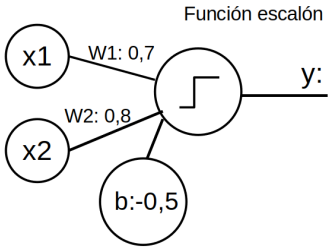
Indique si las siguientes afirmaciones sobre **Redes Neuronales** son verdaderas o falsas (V/F):

a) El perceptrón simple no permite dividir el espacio linealmente:_____
b) Las redes SOM necesitan un conjunto de datos balanceado y correctamente etiquetado para su entrenamiento: _____
c) Para poder entrenar una red neuronal con el algoritmo **Backpropagation**, las funciones de activación de las neuronas deben si o si ser derivables: _____
d) Backpropagation es un método alternativo al método de descenso por gradiente:_____

Padrón: _____ Nombre y Apellido:_____

Ejercicio 7

Dado el siguiente perceptrón, calcule los valores de salida en cada uno de los casos:



$$u(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ 1 & \text{si} \quad x \geq 0 \end{cases}$$

a. x1= 0, x2=0 : _____
b. x1= 1, x2=1 : _____
c. x1= 0.5, x2=0.5 : _____
d. x1= -0.25, x2=1 : _____

Ejercicio 8

Indique si las siguientes afirmaciones sobre implementación de **Redes Neuronales** son verdaderas o falsas (V/F):

a) Dropout es un método de regularización que evita el sobre-ajuste: _____
b) Habitualmente Adam suele ser más rápido que RMSProp y que Momentum como optimizador: _____
c) La regularización L2, L1 nos evita tener que entrenar la red neuronal con Backpropagation : _____
d) La función Softmax permite ponderar la salida como un puntaje o probabilidad de la certeza en la clasificación realizada: _____

Ejercicio 9

Se tiene un conjunto de datos de 250 filas y 10 columnas y se quiere entrenar un árbol de decisión para resolver un problema de clasificación. Se genera un conjunto de entrenamiento (80%) y un conjunto de test (20%) y se quiere evaluar la performance del modelo en entrenamiento utilizando k-fold cross validation con k=5. Responder las siguientes preguntas:

a) ¿Cuántas veces se entrenará el árbol de decisión? _____
b) ¿Con cuántos registros se entrenará el árbol (en cada entrenamiento)? _____
c) ¿Cuántos registros tendrá un conjunto de validación? _____
d) ¿Cuántas veces se entrenaría el árbol si la proporción train/test fuera 70/30?_____

Padrón: _____ Nombre y Apellido:_____

Ejercicio 10

Dado un conjunto de datos de entrenamiento se quieren optimizar los hiperparámetros de un árbol de decisión. Para ello se propone utilizar GridSearch Cross Validation

```
#parámetros a optimizar
params_grid = {'criterion':['gini','entropy'],
               'ccp_alpha':[0.001, 0.0173, 0.0336, 0.05],
               'max_depth':[2,4,5]}

#GridSearchCV
gridcv = GridSearchCV(estimator=arbol ,
                      param_grid=params_grid,
                      scoring='accuracy',
                      cv=10 )
```

Responder las siguientes preguntas:

a) ¿Cuántos juegos de parámetros se van a evaluar con el método GridSearchCV? ____

b) ¿Cuántas veces se va a evaluar cada set de parámetros? ____

c) Muestre dos ejemplos de juegos de parámetros evaluados

```
{'param1': valor, 'parm2': valor, … , 'paramN': valor}
```

1)_____

2)_____

d) ¿Qué representa el parámetro ccp_alpha? ¿Para qué se utiliza?

_____

_____

_____

e) ¿Qué representa el parámetro max_depth? ¿Para qué se utiliza?

_____

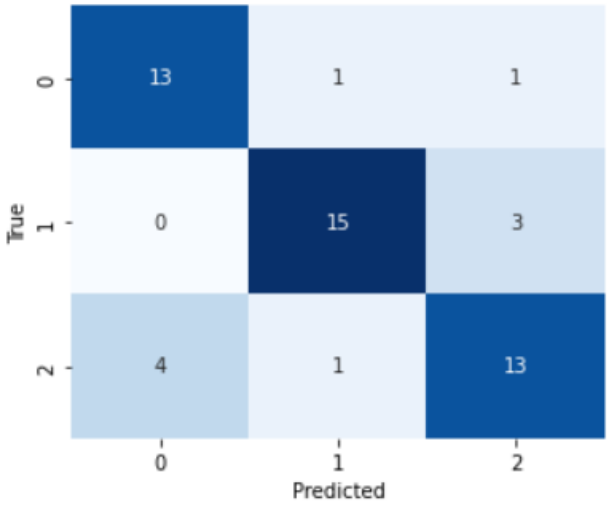
_____

_____

Padrón: _____ Nombre y Apellido:_____

Ejercicio 11

Se entrenó un modelo de clasificación para detectar la especie de una flor : clase 0, clase 1 y clase 2. Luego se evaluó el modelo en los datos de test y se obtuvo la siguiente matriz de confusión:



Se pide:

a) Calcular la métrica Precision para la clase 0: _____
b) Calcular la métrica Recall para la clase 1: _____
c) Calcular el accuracy del modelo: _____

Nota: Los resultados pueden expresarse como fracción

Ejercicio 12

Para un proyecto de ciencia de datos, se entrena un modelo predictivo y se evalúa su performance con la métrica F1-Score. Se obtienen los siguientes resultados:

-F1-Score en entrenamiento: 0.75
-F1-Score en test: 0.31

Indique cuáles si las siguientes afirmaciones son verdaderas o falsas (V/F):

a) El modelo generaliza muy bien para datos nuevos._____
b) El modelo podría estar sobreajustando a los datos de entrenamiento (overfitting)._____
c) Hay un problema en los datos de test, se debería evaluar en otro conjunto._____
d) El modelo podría estar subajustando a los datos de entrenamiento (underfitting)._____

Padrón: _____ Nombre y Apellido:_____

Ejercicio 13

Explique brevemente las principales diferencias entre métodos de regresión y de clasificación.

Ejercicio 14

Explique brevemente las principales diferencias entre métodos de ensamble de tipo bagging y los de tipo boosting.

Ejercicio 15

Describa brevemente una técnica de preprocesamiento de datos aplicada en el TP1.Ejemplifique