Parcial - Machine Learning para Inteligencia Artificial
13 de Julio 2023

Duración: **3 horas**
Con material: **NO**
Puntaje mínimo / máximo: **0 / 30 puntos**
Importante:

- Escribir con letra clara y prolija.
- Las respuestas deben ser completas, precisas y claras, sin ambigüedad.
- La ausencia de alguno de los puntos anteriores puede implicar la quita de puntos.

Ejercicio 1

Se utilizó el dataset de **Palmer Penguins** para predecir la variable **species** en función de los atributos **flipper length** (mm) y **body mass** (g). La tabla siguiente describe brevemente los datos:

Item	flipper length	body mass
count	342	342
mean	200.9	4201.8
std	14.1	802.0
min	172	2700
25%	190	3550
50%	197	4050
75%	213	4750
max	231	6300

Species	Count
Adelie	151
Gentoo	123
Chinstrap	68
Total	342

Se dividió el conjunto de datos disponibles ($N = 342$) en un conjunto de entrenamiento *Train* ($N_{train} = 273$) y un conjunto de *Test* ($N_{test} = 69$). Con el objetivo de aplicar el algoritmo de *K-Vecinos más cercanos (KNN)* se utilizó el método *Repeated Holdout* con 25 repeticiones y un tamaño para el conjunto de validación del 25% de *Train*.

Figura 1: resultado de *Repeated Holdout*.

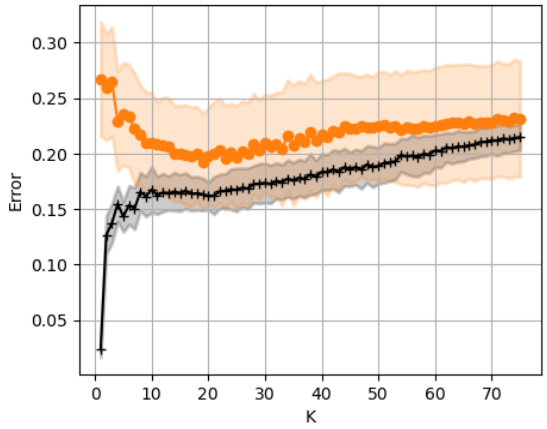
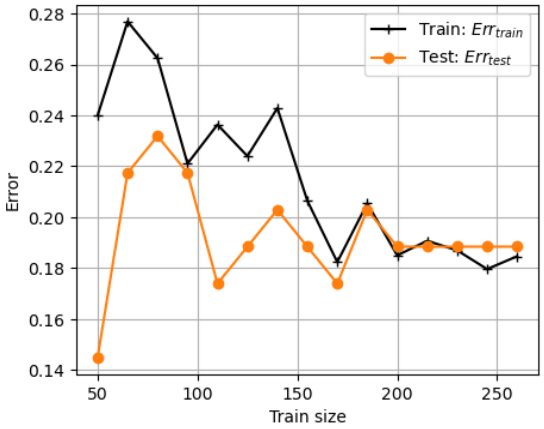


Figura 2: *Learning Curve* para el valor elegido de K



1. Explique en qué consiste el algoritmo de K-Vecinos más cercanos (KNN), indicando cuáles son los hiper-parámetros que determinan su sesgo inductivo.

2. Dado el resumen de los datos disponibles, indique qué tipo de preprocesamiento considera adecuado para KNN.
3. Explique el objetivo de dividir el conjunto de datos disponibles en *Train* y *Test*. Relacione las componentes de error en la descomposición *Sesgo-Varianza-Error irreducible* con éstos subconjuntos de datos.
4. Explique las curvas de la Figura 1 indicando qué representa cada una, por qué tienen la forma mostrada y cuál es el objetivo de realizar *Repeated Holdout*.
5. Indique qué valor de K elegiría a partir de la Figura 1. Justifique su respuesta.
6. A partir de lo mostrado por las curvas de la Figura 2, ¿qué conclusiones se obtienen sobre el desempeño del algoritmo con el valor de K elegido?

Ejercicio 2

Considere los siguientes pseudo-códigos de dos algoritmos de ensemble:

Ensemble A

Entrada: Conjunto de datos de entrenamiento *D*
Salida: Clasificador ensemble

1. Repetir *K* veces:

- **1.1.** Muestrear aleatoriamente con reemplazo un subconjunto de entrenamiento *D'* de tamaño *N*=|*D*| a partir de *D*.
- **1.2.** Entrenar un clasificador base *C* utilizando *D'*.
- **1.3.** Agregar *C* al conjunto de clasificadores base del ensemble.

2. Devolver el clasificador ensemble con voto mayoritario.

Ensemble B

Entrada: Conjunto de datos de entrenamiento *D*
Salida: Clasificador ensemble

1. Repetir *K* veces:

- **1.1.** Muestrear aleatoriamente con reemplazo un subconjunto de entrenamiento *D'* de tamaño *N*=|*D*| a partir de *D*.
- **1.2.** Seleccionar aleatoriamente un subconjunto de atributos *M* de tamaño *m* (donde *m* es menor al número total de atributos).
- **1.3.** Entrenar un clasificador base *C* utilizando *D'* y *M*.
- **1.4.** Agregar *C* al conjunto de clasificadores base del ensemble.

2. Devolver el clasificador ensemble con voto mayoritario.

Se pide:

1. Al utilizar el voto mayoritario como clasificador ensemble, ¿en qué aspecto de la descomposición de sesgo-varianza se espera mejorar en comparación con los clasificadores base individuales?
2. ¿Qué condiciones se deberían cumplir para que lo postulado en el punto 1. logre efectivamente obtener mejores resultados?
3. ¿Cuál es el propósito de la selección aleatoria de subconjuntos de atributos en la construcción del clasificador Ensemble B? Justifique su respuesta.

Ejercicio 3

Se desea predecir si un estudiante pasará o no un examen basado en dos variables: el número de horas de estudio (X_1) y la cantidad de horas de sueño (X_2) la noche anterior al examen. El resultado deseado está representado por la variable Y, donde $Y = 1$ si el estudiante pasa el examen y $Y = 0$ si no lo pasa. Para esto, se utiliza un modelo de regresión logística con la función de decisión dada por $f(X) = \sigma(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$, en donde σ representa la función sigmoidea.

A continuación se muestra un pequeño dataset de validación con los resultados obtenidos:

Id	X_1	X_2	Y	f(X)
1	3	4	0	0.03
2	4	5	0	0.11
3	5	5	0	0.24
4	4	7	0	0.30
5	5	6	1	0.37
6	6	6	0	0.59
7	6	7	1	0.72
8	7	6	1	0.78
9	7	8	1	0.92
10	8	7	1	0.94

1. Indicar de qué tipo de problema de Machine Learning se trata y qué representa cada una de las columnas de la tabla.
2. Realizar un scatterplot de X_1 y X_2, indicando con un círculo las observaciones para las cuales $Y=0$ y con una cruz aquellas donde $Y=1$. Hacer un bosquejo de la curva $f(X)=0.5$, para esto haga un análisis intuitivo de la ubicación de la curva.
3. Para cada uno de los siguientes umbrales [1.0, 0.72, 0.59, 0.37, 0.0] calcular la matriz de confusión para el dataset de validación. Usar la convención en la cual $\hat{Y}=1$ si y sólo si $f(X) \geq umbral$.
4. Hacer un gráfico detallado de la curva ROC usando los umbrales mencionados en la parte anterior.
5. Calcular el área bajo la curva ROC (AUC-ROC) y determinar qué tan bueno es el rendimiento del modelo en términos de su capacidad para distinguir entre las clases.

Ejercicio 4

Se desea construir un árbol de decisión para predecir si un correo electrónico es spam o no spam, basado en dos atributos: "Longitud del correo" (en palabras) y "Número de enlaces" presentes en el correo. Se dispone de un conjunto de entrenamiento con 50 correos electrónicos, donde 30 son spam y 20 no son spam. Los datos se resumen en las siguientes tablas:

Tabla 1: Longitud del correo

Longitud	Correos	Spam	No Spam
Corto	28	25	3
Largo	22	5	17

Tabla 2: Número de enlaces

Enlaces	Correos	Spam	No Spam
Bajo	25	15	10
Alto	25	15	10

Se pide:

1. Describa el procedimiento mediante el cual se elige la pregunta a realizar en el nodo raíz de un árbol de decisión. Indique cuáles son los criterios comúnmente utilizados.
2. A partir de las Tablas 1 y 2, ¿qué atributo elige para particionar la raíz del árbol de decisión? Dibuje los dos árboles posibles como ayuda en la justificación de la elección.

Ejercicio 5

Este ejercicio contiene las siguientes preguntas sobre el obligatorio, responder brevemente en un **máximo de media página**:

1. ¿Qué métricas de clasificación utilizó y por qué?
2. ¿Cuál fue su metodología para seleccionar y validar modelos?
3. ¿Qué algoritmos le dieron mejores resultados?