

Repaso Parcial - Machine Learning para IA

1er Semestre 2024

Ejercicio 1

Se desea construir un árbol de decisión para clasificar frutas, específicamente manzanas y naranjas. Se usarán dos atributos: **Acidez** y **Peso**. Se dispone de un conjunto de entrenamiento con 80 frutas: 40 manzanas y 40 naranjas.

Los datos se resumen en las siguientes tablas:

Acidez	Nº Frutas	Manzanas	Naranjas	Peso	Nº Frutas	Manzanas	Naranjas
Alta	40	20	20	Ligero	50	35	15
Baja	40	20	20	Pesado	30	5	25

Preguntas:

1. Dibuje los dos árboles de decisión posibles, resultantes de dividir el nodo raíz por uno de los dos atributos.

Cada árbol debe incluir la siguiente información:

- a) Número de observaciones en el nodo raíz y hojas.
- b) La distribución de las etiquetas (y) en el nodo raíz y hojas.
- c) Impureza de Gini $H(y; T) = 1 - \sum_c p_c^2$ en el nodo raíz y hojas.
- d) En el nodo raíz: la pregunta realizada.
- e) En las hojas: la predicción.

2. Calcule la impureza de Gini esperada

$$H_P(y; T) = \frac{|T_{\text{true}}|}{|T|} H(y; T_{\text{true}}) + \frac{|T_{\text{false}}|}{|T|} H(y; T_{\text{false}})$$

asociada a cada una de las dos divisiones de la parte anterior.

3. ¿Qué pregunta elegiría en el nodo raíz? Justifique en base a las partes anteriores.
4. Para el árbol elegido en la parte anterior, calcule el error en los datos de entrenamiento.

Ejercicio 2

Usaremos el siguiente conjunto de datos para que un árbol de decisión pueda predecir si los estudiantes aprueban el examen (**Passed** con valores **True** o **False**) según su **GPA** (Grade Point Average, promedio ponderado de calificaciones con valores **High**, **Medium** o **Low**) y si estudiaron o no (**Studied** con valores **True** o **False**).

Student	GPA	Studied	Passed
1	L	F	F
2	L	T	T
3	M	F	F
4	M	T	T
5	H	F	T
6	H	T	T
7	H	F	F

Preguntas:

1. Dibuje el árbol de profundidad 1 resultante de hacer la división con el atributo **Studied**.
2. Calcule la impureza de Gini $H(y; T) = 1 - \sum_c p_c^2$ en el nodo raíz y hojas.
3. ¿Qué métrica debería priorizar en caso de querer darle un curso extra a TODOS aquellos estudiantes que probablemente NO pasen el curso?
4. ¿Cuánto vale la métrica que priorizó en el punto 2 en el árbol dibujado?

Ejercicio 3

Se han entrenado cuatro modelos en el mismo conjunto de datos con diferentes hiperparámetros. En la siguiente tabla se registra el riesgo (costo o error) en el conjunto de Train y Test para cada uno de ellos:

Model	Train Risk	Test Risk
1	0.1	1.8
2	0.4	1.2
3	0.6	0.8
4	1.9	2.3

Sabiendo que cuanto menor sea el riesgo es mejor, conteste las siguientes preguntas:

1. ¿Con cuál modelo se quedaría y por qué?
2. ¿Cuál o cuáles de los modelos parecen estar sobreajustando los datos de entrenamiento? Justifique.

Ejercicio 4

Se ha entrenado un clasificador en un problema de clasificación binaria con etiquetas

0 : negativo y 1 : positivo.

La siguiente tabla resume los resultados en un conjunto de validación de 8 observaciones:

Etiqueta Verdadera	0	0	1	0	0	1	1	1
Probabilidad Predicha	0.11	0.18	0.35	0.47	0.55	0.62	0.76	0.94

Preguntas:

1. Para el umbral $\beta = 0.5$ calcule la matriz de confusión y el error.
2. Complete la siguiente tabla:

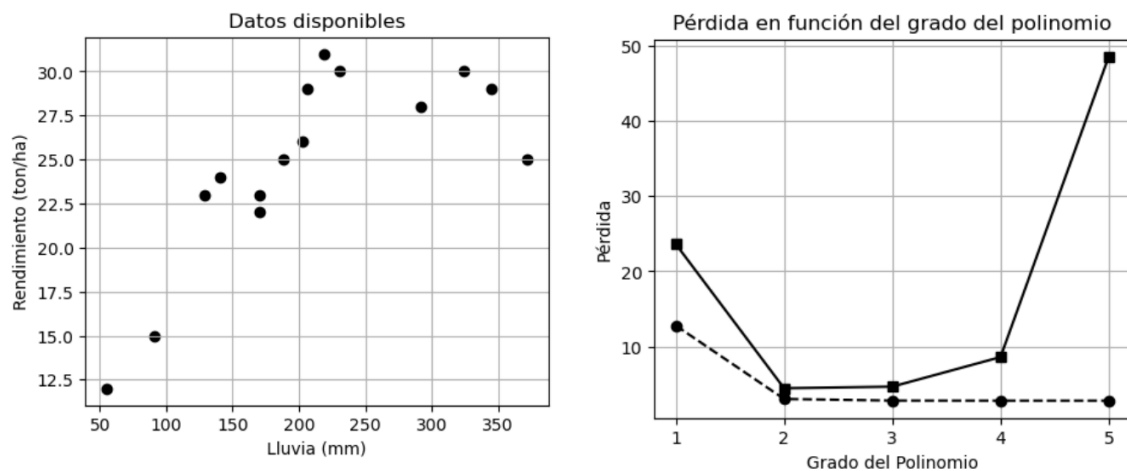
Umbral	Error	Precision	TPR (Recall)	FPR
0.94	0.375	1.00	0.25	0.00
0.76	0.250	1.00		0.00
0.62	0.125	1.00	0.75	0.00
0.55	0.250	0.75	0.75	0.25
0.47	0.375	0.60	0.75	0.50
0.35	0.250	0.67	1.00	
0.18	0.375		1.00	0.75
0.11	0.500	0.50	1.00	1.00

Nota: se predice la etiqueta 1 si la probabilidad predicha es mayor o igual al umbral.

3. Grafique la curva ROC y calcule el AUC.
4. Suponiendo que un falso positivo (FP) tiene un costo de \$100, y un verdadero positivo (TP) tiene una ganancia de \$150, ¿qué valor de umbral elegiría? Justifique.

Ejercicio 5

El gráfico a continuación (izquierda) muestra datos de rendimiento de cultivos de papa (en ton/ha) en función de la lluvia acumulada recibida (en mm):



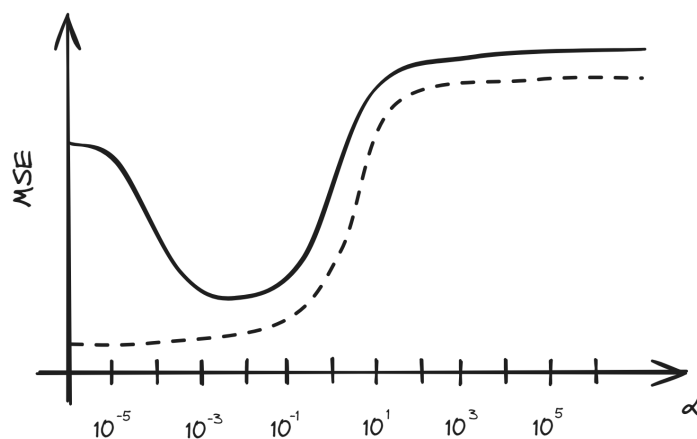
Se desea predecir el Rendimiento en función de la Lluvia:

1. ¿Cree que sería útil estandarizar los datos? Justifique haciendo referencia a este ejemplo en concreto.
2. El gráfico arriba a la derecha muestra las curvas de riesgo en función del grado, tanto en train como en cross-validation:
 - a) Indique cuál es la función de pérdida (loss) más apropiada para este problema.
 - b) Identifique las curvas.
 - c) Explique el propósito de usar cross-validation en este ejemplo concreto.
3. ¿Qué grado elegiría? Justifique.
4. Suponga que decide entrenar con Grado = 6. ¿Qué técnica vista en clase le ayudaría a evitar el sobreajuste? Base su justificación en la explicación de dicha técnica.

Ejercicio 6

Supongamos que está tratando de resolver un problema de regresión, y observa que con un polinomio de grado 6 logra tener MSE muy pequeña en el conjunto de train. Sin embargo la MSE en el conjunto de test observa que es muy grande.

1. ¿Qué nos podría decir del modelo?
2. ¿Qué utilizaría para intentar mejorar el modelo? Justifique.
3. La siguiente gráfica representa la MSE del modelo en función de α el hiperparámetro de regularización:



- a) Identifique cada una de las curvas representadas.
- b) ¿En qué lugares se produce sobreajuste?
- c) ¿Qué valor de α elegiría?

Ejercicio 7

1. Dadas las siguientes gráficas, ¿con qué modelo se quedaría y por qué? Justifique en base a los problemas que identifica en los otros dos modelos.

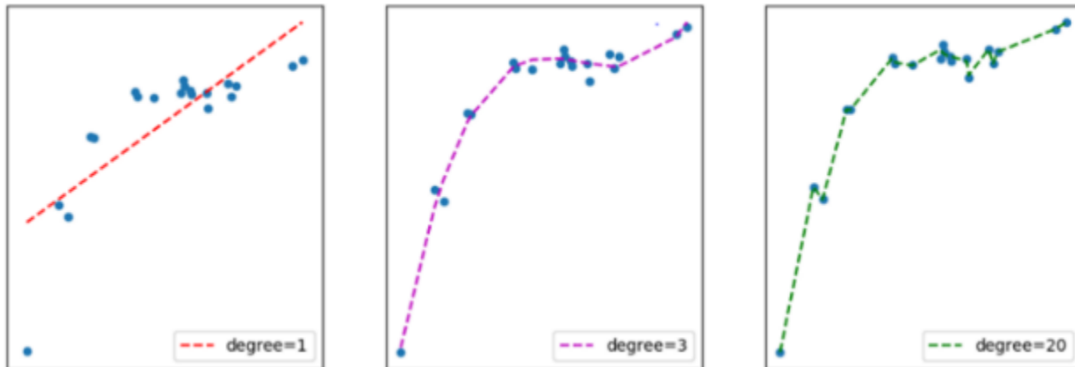
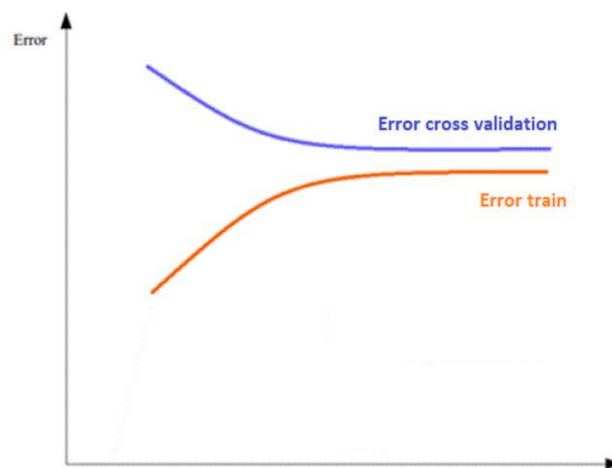


Figura 1: Gráficas de diferentes modelos polinomiales con grados 1, 3 y 20.

2. Defina en sus palabras sesgo (bias) y varianza.
3. ¿Cuál es el objetivo de particionar los datos que se tienen para desarrollar un sistema inteligente en *train*, *test* y *validation*? Desarrolle brevemente.

Ejercicio 8

La siguiente gráfica representa lo que llamamos Learning Curves:



1. ¿Qué es lo que se representa en el eje de las x de una Learning Curve?
2. ¿Para qué sirven las Learning Curves?
3. ¿Qué problema se detecta en la Learning Curve del dibujo?
4. ¿Qué opción se le ocurre para solucionarlo en un modelo de Regresión Lineal?

Ejercicio 9

Un modelo de Regresión Logística arroja los resultados de probabilidad para la clase “Yes” que se muestran en la tabla siguiente:

Id	Score (sorted)	Real Value
10	0.93	Yes
8	0.90	Yes
7	0.89	Yes
5	0.81	No
1	0.74	Yes
6	0.69	Yes
2	0.59	No
9	0.36	No
3	0.25	No
4	0.01	No

1. ¿Qué umbral elegiría para que no haya falsos positivos?
2. ¿Qué umbral elegiría para que no haya falsos negativos?
3. Al no existir falsos negativos ¿cómo impacta en el PRECISION y en el RECALL del modelo?

Ejercicio 10

Se desea identificar una especie de flor, representada por la variable Y (cuyo valor es 1 si la flor es de la especie buscada y 0 si no), en base al largo de su pétalo (en mm) representado por la variable X . Para ello se realiza una regresión logística $\sigma(f(X))$ con $f(X) = \beta_0 + \beta_1 X$ y σ la función sigmoidea, utilizando un gran dataset de entrenamiento.

En la siguiente tabla se muestra el resultado en un pequeño dataset de validación:

Id	X: largo del pétalo	Y: especie	$\sigma(f(X))$
1	11	0	0.40
2	13	0	0.45
3	20	1	0.62
4	15	0	0.50
5	19	1	0.60
6	22	1	0.67
7	17	0	0.55
8	13	1	0.45
9	12	0	0.43
10	16	1	0.52

1. ¿Qué estima el valor de la columna $\sigma(f(X))$ correspondiente al valor calculado por la hipótesis obtenida?
2. Grafique la columna $\sigma(f(X))$ en función de X .
3. Para dicho dataset, calcule la exactitud (accuracy) para el siguiente conjunto de umbrales (1 si \geq umbral): 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65 y 0.70.
4. ¿Qué valor de umbral elegiría? Justifique.
5. ¿Existe algún umbral que tiene exactitud igual a 1? Justifique.

Ejercicio 11

Considere la siguiente tabla que representa el peso con que quedaron 10 ejemplos luego de la primera iteración del algoritmo AdaBoost:

Ejemplo	1	2	3	4	5	6	7	8	9	10
Peso	0.17	0.17	0.17	0.07	0.07	0.07	0.07	0.07	0.07	0.07

1. ¿Cuánto valía el peso del ejemplo número 4 al inicio (antes de la primer iteración)?
2. ¿Cuánto valía el peso del ejemplo número 2 al inicio (antes de la primer iteración)?
3. Se pide detallar (en la primer iteración y viendo la tabla) cuáles ejemplos quedaron bien clasificados y cuáles no.
4. ¿Qué característica tienen los árboles de decisión que usa el algoritmo AdaBoost?

Ejercicio 12

Se desea saber a qué precio estarán determinadas acciones la semana que viene. Luego de varios intentos Gradient Boosting resultó ser el algoritmo con mejor exactitud en predecir el valor de las mismas.

Para aumentar aún más el rendimiento del algoritmo se propone bajar el learning rate.

1. En caso de aprobarse la medida, ¿qué haría con el número de estimadores? Justifique.
2. ¿Qué consecuencia negativa podría generar esta decisión?
3. Nombre alguna diferencia entre los algoritmos de boosting y de bagging.

Ejercicio 13

En un problema de clasificación donde se aplica ensemble con voto mayoritario como función de agregación, se tienen las siguientes probabilidades para los modelos M1, M2, y M3, en un dataset de test de 6 observaciones.

Real Value	M1	M2	M3	Output
1	0.45	0.80	0.75	
1	0.49	0.54	0.70	
0	0.30	0.20	0.50	
0	0.20	0.50	0.30	
1	0.48	0.50	0.51	
0	0.30	0.49	0.50	

Si se aplica en todos los casos el umbral 0.5 (mayor o igual a 0.5 para la categoría "1" o menor a 0.5 para la categoría "0"):

1. ¿Se cumple que la predicción final (columna Output) tiene mayor accuracy que cualquiera de los modelos individuales? Justificar.
2. ¿Cuál de los tres modelos usados como input por el ensemble tiene el menor accuracy?
3. ¿Cuál de los tres modelos usados como input por el ensemble tiene el mayor accuracy?

Ejercicio 14

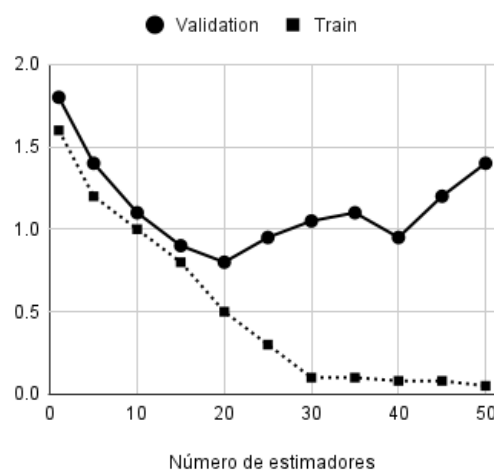
Se quiere resolver un problema de clasificación binaria, para lo cual se analizan dos algoritmos de ensemble: Algoritmo A y Algoritmo B. Uno de ellos corresponde a una técnica de *bagging* mientras que el otro a una técnica de *boosting*.

1. Explique las principales diferencias entre ambos tipos de técnicas de ensemble.
2. Analice las siguientes afirmaciones:
 - a) "Bagging es efectivo cuando los estimadores base tienen poca correlación."
 - b) "Como boosting ataca el problema de varianza alta, si los estimadores base son árboles de decisión es preferible que tengan profundidad máxima."

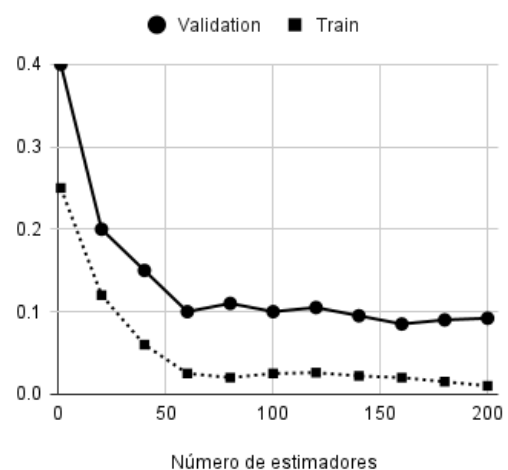
Indique, justificando su respuesta, si estas afirmaciones son correctas.

3. Considere las siguientes gráficas que muestran la evolución (en función del número de estimadores) de la pérdida (para el Algoritmo A) y el error (para el Algoritmo B), tanto en *train* como en *validation*:

Curvas de Loss para el Algoritmo A



Curvas de error para el Algoritmo B



A partir de estas gráficas identifique cada algoritmo con su respectiva técnica de ensemble. Justifique su respuesta.

4. Para cada uno de los algoritmos A y B, analice la siguiente afirmación:

"La curva de validación (pérdida o error) tenderá a disminuir si entrenamos con una cantidad mayor de estimadores (mas allá del límite mostrado en la gráfica)."

Indique, justificando su respuesta, si esta afirmación es correcta.

Ejercicio 15

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.
4. Identifique cada uno de los ensembles.

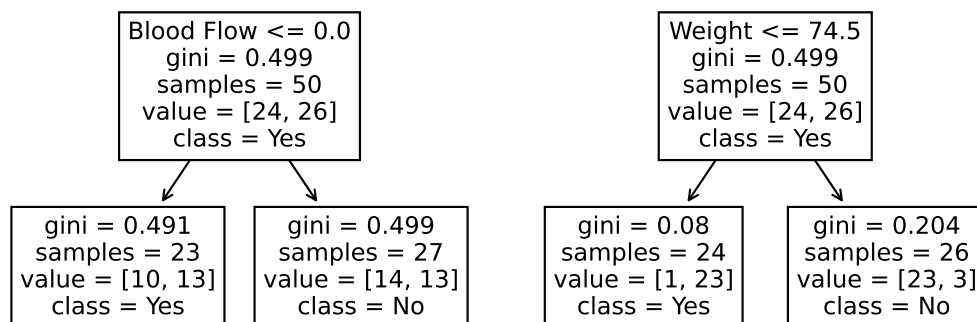
Ejercicio 16

La siguiente tabla resume las primeras tres líneas de un conjunto de datos para predecir la variable **HD** (Heart Disease):

Id	HD	Blood Flow	Blocked Arteries	Chest Pain	Weight
1	No	1	-1	-1	70
2	Yes	-1	1	1	91
3	No	-1	-1	1	85
⋮	⋮	⋮	⋮	⋮	⋮

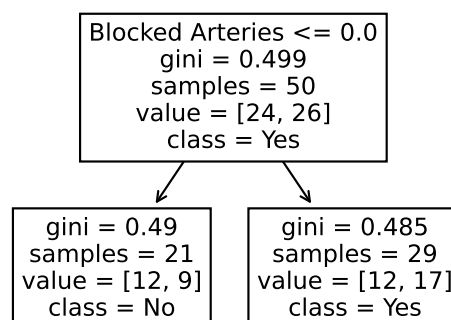
Cuadro 1: Datos disponibles.

- Indique de qué tipo de problema de ML se trata.
- Para resolver el problema se eligió un algoritmo de ensemble que en cada iteración utiliza diferentes muestras (bootstrap) de los datos y a la vez selecciona aleatoriamente un subconjunto de atributos para entrenar un árbol de decisión.
 - Indique cuál es el algoritmo.
 - Explique cuál es el propósito de elegir un subconjunto de atributos al azar.
- Para el algoritmo de la parte 2., se decide entrenar con árboles de profundidad 1. Los diagramas a continuación muestran los dos árboles posibles para la primera iteración, en la cuál los atributos seleccionados aleatoriamente son **Blood Flow** y **Weight**:



Indique cuál es el árbol elegido por el algoritmo. Justifique.

- Suponga que el árbol elegido en la segunda iteración del algoritmo de la parte 2. es el mostrado en el siguiente diagrama:



Calcule el error promedio del ensemble (de 2 iteraciones) para las tres observaciones mostradas en el Cuadro 1.

Ejercicio 17

El cuadro a continuación representa las predicciones de 5 modelos base en un problema de clasificación binaria en donde la variable a predecir es Y con valores 0 o 1:

id	h_1	h_2	h_3	h_4	h_5	Y
1	0	0	0	0	0	0
2	0	1	1	1	0	1
3	1	0	1	1	1	1
4	1	1	1	1	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	1
7	1	1	1	1	1	1
8	1	0	0	1	1	0
9	1	0	1	0	0	1
10	0	0	1	0	0	0

Se pide:

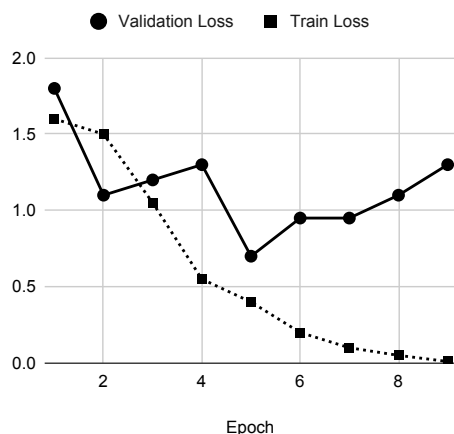
1. Calcule el error de cada uno de los modelos base y su valor promedio.
2. Calcule las predicciones del modelo ensemble que aplica voto mayoritario.
3. Calcule el error del ensemble de la parte anterior.
4. Compare el promedio calculado en 1. con el error del ensemble calculado en 3. Explique el resultado obtenido en dicha comparación.

Ejercicio 18

Considere la siguiente red de tipo MLP:

Layer (type)	Output Shape	Param #
D_1 (Dense)	(None, 6)	??
A_1 (Activation)	(None, 6)	0
D_2 (Dense)	(None, D)	84
A_2 (Activation)	(None, D)	0
D_3 (Dense)	(None, 3)	??
A_3 (Activation)	(None, 3)	0
Total params: 141		
Trainable params: 141		
Non-trainable params: 0		

Validation Loss and Train Loss



1. Calcule la dimensión de salida D de D_2.
2. Calcule la dimensión de entrada de la red.

3. Dada la siguiente salida de A_1:

```
[[0.0, 0.0, 7.5, 0.2, 0.0, 0.0],
 [6.0, 4.6, 0.7, 0.0, 0.3, 0.0],
 [3.8, 0.0, 9.2, 1.0, 0.0, 0.0],
 [5.7, 4.9, 0.0, 0.5, 0.0, 0.1]]
```

- a) ¿cuál es la función de activación?
- b) ¿cuál es el tamaño del batch?

4. Dada la siguiente salida de A_3:

```
[[0.25, 0.25, 0.50],
 [0.40, 0.35, 0.25],
 [0.05, 0.65, 0.30],
 [0.85, 0.02, 0.13]]
```

- a) ¿cuál es la función de activación?
- b) ¿de qué tipo de problema de ML se trata?
- c) ¿qué función de loss usaría?

5. Si para la salida de la parte 4 se sabe que la variable target es:

```
[[0, 0, 1],
 [0, 1, 0],
 [0, 1, 0],
 [1, 0, 0]]
```

¿cuántos aciertos hay en este batch?

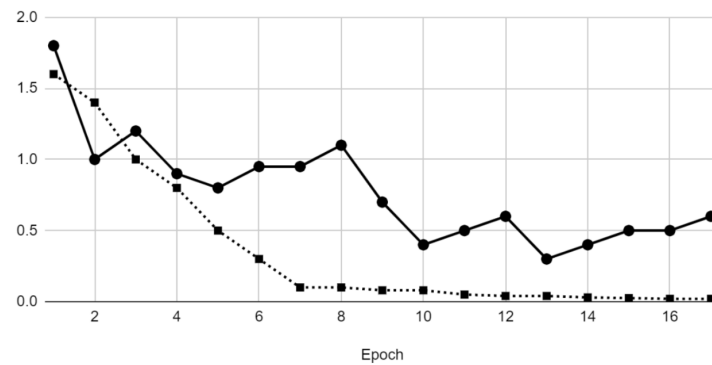
6. En la gráfica se muestran las curvas de Validation Loss (línea sólida) y Train Loss (línea punteada) de un entrenamiento de la red neuronal inicializado en 12 épocas y con Early Stopping.

A partir de la misma determine:

- a) El valor de la paciencia
- b) La época en la que se obtiene el mejor modelo

Ejercicio 19

En la siguiente gráfica se muestran las curvas de loss en Validation y Train de un entrenamiento de una red neuronal inicializado en 20 épocas y con Early Stopping.



Se pide:

1. Identifique en la gráfica las curvas de Validation y Train.
2. Determine en la época en la que se obtiene el mejor modelo.
3. Determine el valor de la paciencia.