Ejercicio 1 Interpretar un resumen numérico

Abajo se muestran indicadores que caracterizan la distribución de notas de dos clases paralelas de un curso de Inglés. El puntaje máximo es 100.

	Clase 1	Clase 2
Promedio	72	78
Mediana	73	65
Desvío estándar	6	16

1. Bosquejar el histograma (o la densidad) de la distribución de notas de cada clase.
2. ¿En cuál de las dos clases es más probable encontrar un estudiante con nota alta?

Ejercicio 2 Máxima verosimilitud

Considerar la densidad

$$p(x;\theta) = \frac{1}{2}(1+\theta x) \quad -1 \leq x \leq 1$$

El parámetro θ también varía entre -1 y 1.

Suponga que dispone de la siguiente muestra

-0.47 0.89 0.26 -0.59 0.37 0.54 0.87

Implementar el algoritmo Golden Search para hallar el estimador de máxima verosimilitud de θ.

Ejercicio 3 Regresión lineal

Considerar la siguiente tabla de datos sobre el rendimiento de cultivos de papas y el registro de lluvias acumuladas en el período de duración del cultivo:

Datos de entrenamiento		**Datos de validación**	
x = Lluvia (mm)	y = Rendimiento (ton/ha)	x = Lluvia (mm)	y = Rendimiento (ton/ha)
206	29	213	30
188	25	80	16
219	31	391	25
372	25	250	26
345	29	57	9
231	30	303	28
203	26	263	28
170	23	157	25
55	12	72	13
91	15	157	23
292	28	188	26
141	24	216	25
129	23	362	28
170	22	283	33
324	30	308	30

1. Correr una regresión lineal para predecir el rendimiento y en función de la lluvia x.
2. Comparar el MSE en entrenamiento, validación y CV.
3. Determinar el grado óptimo en caso de aplicar una regresión polinomial.
4. Hallar el valor de λ óptimo para la regresión polinomial de grado 5 con regularización.

Ejercicio 4 Multicolinealidad en regresión lineal

Supongamos un problema de regresión lineal bi-variado, es decir con $D = 2$ atributos:

$$y = b + w_1 x^{(1)} + w_2 x^{(2)} + \epsilon$$

Supongamos que los datos $\mathbf{y}$, $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$ han sido estandarizados. Denotemos por

- $r = \text{cor}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right)$
- $[r_1, r_2]^\top = \text{cor}\left([\mathbf{x}^{(1)}, \mathbf{x}^{(2)}], \mathbf{y}\right)$

Se pide:

1. Mostrar que $\widehat{b} = 0$ y los pesos están dados por:

$$\widehat{\mathbf{w}} = \frac{1}{1 - r^2} \begin{bmatrix} r_1 - r r_2 \\ r_2 - r r_1 \end{bmatrix}$$

2. Las correlaciones r, r_1, y r_2 no pueden elegirse arbitrariamente pero basta que satisfagan la relación

$$1 - r_1^2 - r_2^2 + 2 r_1 r_2 r - r^2 \geq 0$$

   Dados r_1 y r_2 hallar un intervalo de valores posibles para r.

3. Mostrar con elecciones adecuadas de r_1, r_2 y r que los pesos del modelo pueden ser abritrariamente grandes en norma.

4. Hacer un bosquejo de las curvas de nivel de $L(\mathbf{w})$ y mostrar cómo las mismas varían con r_1, r_2 y r.

Ejercicio 5 Un juego bayesiano

Carla y Walter están jugando a un juego en el que la primera persona que consigue 6 puntos gana. La forma en que cada punto se decide es la siguiente: se lanza una moneda, si sale cara es un punto para Carla, si sale número es un punto para Walter. Ambos desconocen la probabilidad θ de que en un lanzamiento la moneda sea cara.

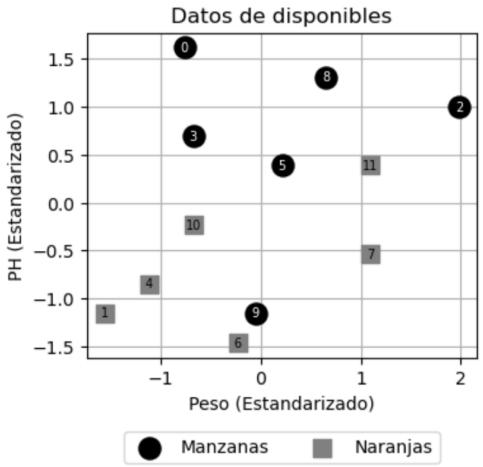
Supongamos que Carla ya está ganando 5 puntos a 3. ¿Cuál es la probabilidad de que Carla gane el juego?

Ejercicio 6 K vecinos más cercanos

Se desea implementar el algoritmo de K vecinos más cercanos para clasificar **Manzanas** y **Naranjas** en base a su **Peso** y **PH**.

El gráfico a continuación (derecha) muestra los datos disponibles estandarizados. En el lado izquierdo, se presenta el conjunto de datos segmentado en tres folds. Cada fila simboliza una observación, y su ID correspondiente está alineado con el gráfico situado a la derecha. Las columnas, por otro lado, exhiben los IDs de los 8 puntos que son parte de los dos folds que no contienen la observación de la fila en cuestión, todos ellos ordenados por su proximidad a dicha observación. Por ejemplo, en la primera fila se visualizan las observaciones de los folds 2 y 3. Están dispuestas en un orden que va desde la más cercana a la más lejana respecto a la observación con ID=0.

	ID	1ero	2do	3ro	4to	5to	6to	7mo	8vo
Fold 1	**0**	8	5	10	11	4	7	9	6
	1	4	10	6	9	5	7	11	8
	2	11	8	7	5	10	9	6	4
	3	10	5	8	4	11	9	7	6
Fold 2	**4**	1	10	9	3	0	11	8	2
	5	11	3	8	10	9	0	2	1
	6	9	10	1	3	11	8	0	2
	7	11	9	2	10	8	3	1	0
Fold 3	**8**	5	2	0	3	7	4	6	1
	9	6	4	7	1	5	3	0	2
	10	4	3	5	1	6	7	0	2
	11	5	7	2	3	0	6	4	1



Se pide:

1. Calcular el error de 3-fold cross-validation para los valores impares de K. ¿Cuál de estos valores de K elegiría?
2. Analizar la descomposición del error en este ejemplo.

Ejercicio 7 Regresión logística

Se tiene los siguientes datos:

Paciente	Glucosa (mg/dL)	Tiene Diabetes (1 = sí, -1 = no)
A	90	-1
B	160	1
C	100	-1
D	200	1
E	130	1

El objetivo del ejercicio es implementar el algoritmo de *descenso de gradiente* para estimar los parámetros óptimos de la regresión logística.

1. La función de pérdida en regresión logística es

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ln\left(1 + e^{-y_i\theta^\top x_i}\right)$$

en donde $\theta = (b, w)$ es el parámetro.

Mostrar que el gradiente de L con respecto a θ, que se define como:

$$\nabla L(\theta) = \frac{\partial L(\theta)}{\partial \theta}$$

está dado por

$$\frac{\partial L(\theta)}{\partial \theta_j} = \frac{1}{N}\sum_{i=1}^{N} -\frac{y_i x_{ij} e^{-y_i\theta^\top x_i}}{1 + e^{-y_i\theta^\top x_i}}$$

2. Podemos actualizar los coeficientes de forma iterativa usando la fórmula del descenso de gradiente:

$$\theta_{t+1} = \theta_t - \alpha\nabla L(\theta_t)$$

Donde α es la tasa de aprendizaje. Las iteraciones se realizan hasta alcanzar una tolerancia deseada. Calcular usando este procedimiento el valor $\widehat{\theta}$ del parámetro óptimo en este ejemplo.

Ejercicio 8 Árboles de decisión

Se desea construir un *árbol de decisión* para clasificar dos variedades de tomates: **Cherry** y **Perita**. Se usarán dos atributos: **Dulzor** y **Tamaño**. Se dispone de un conjunto de entrenamiento con 90 tomates: 45 Cherry y 45 Perita.

Los datos se resumen en las siguientes tablas:

Dulzor	**No tomates**	**Cherry**	**Perita**
Alto	55	30	25
Bajo	35	15	20

Tamaño	**No tomates**	**Cherry**	**Perita**
Pequeño	60	40	20
Grande	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 $H(y; S) = 1 - \sum_c p_c^2$ 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

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

   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

Ejercicio 9 Baggin vs Random Forest

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

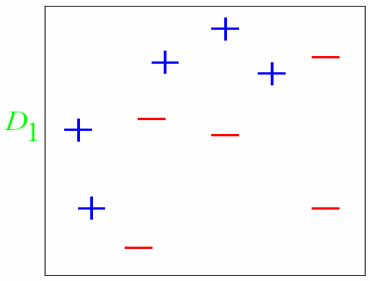
1. Repetir K veces:
   1.1 Muestrear aleatoriamente con reemplazo un subconjunto de entrenamiento D' de tamaño $N = |D|$ a partir de D.
   1.2 Seleccionar aleatoriamente un subconjunto de atributos M de tamaño m (donde m es menor al número total de atributos).
   1.3 Entrenar un clasificador base C utilizando 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 10 Boosting

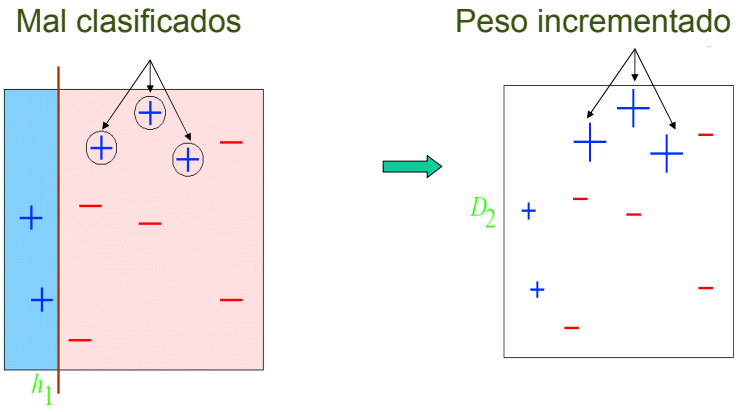
Considere el dataset D_1 de la siguiente figura, de la forma $\{(\boldsymbol{x}_i, y_i)\}$ con $\boldsymbol{x}_i \in \mathbb{R}^2$, e $y_i \in \{+1, -1\}$ para todo $i \in [1, n]$



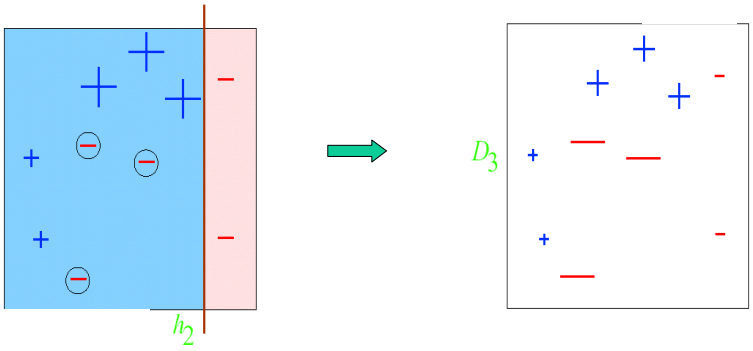
Inicialmente todos los datos tienen el mismo peso: $w_i(1) = \frac{1}{n}, i \in [1, n]$

Se pide:

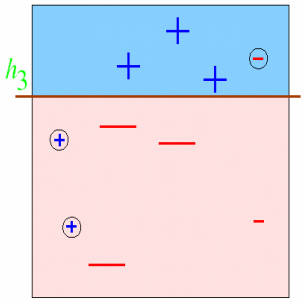
1. Primera ronda: justificar que el árbol de profundidad 1 (stump) h_1 dado por el algoritmo es



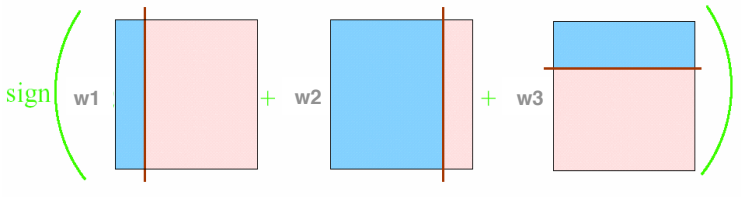
2. Determinar el error ϵ_1 y el voto α_1 de dicho árbol en el ensemble.
3. Calcular los nuevos pesos de las observaciones $w_i(2), i \in [1, n]$.
4. Segunda ronda: justificar que el árbol de profundidad 1 (stump) h_2 dado por el algoritmo es



5. Tercera ronda: justificar que el árbol de profundidad 1 (stump) h_3 dado por el algoritmo es



6. Concluir el modelo final


sign
w1
+
w2
+
w3