



UNIVERSIDAD
DE GRANADA

Aprendizaje Automático. Complejidad de H y Modelos Lineales

Ruido y complejidad, PLA, Regresión Logística, Clasificación de dígitos, PLA Pocket, Cota de Vapnik-Chervonenkis

Ricardo Ruiz Fernández de Alba

Escuela Técnica Ingeniería Informática y Matemáticas

DECSAI

Universidad de Granada

8 de mayo de 2022

Índice general

Índice general	ii
1 Sobre la complejidad de H y el ruido	1
1.1 Dibujar gráficas de nubes de puntos simuladas	1
1.1.1 Uniformemente distribuidos	1
1.1.2 Siguiendo distribución gaussiana de media 0 varianza dada	2
1.2 Ejercicio 2	2
1.2.1 Dibujo de puntos con etiqueta y recta usada	2
1.2.2 Añadir ruido aleatorio	3
1.2.3 Otras fronteras de clasificación	4
2 Modelos Lineales	7
2.1 Algoritmo de aprendizaje del Perceptrón (PLA)	7
2.1.1 Ejecutar el algoritmo PLA con los datos empleados en el apartado 2a del ejercicio 1.	7
2.1.2 Repetir usando los datos del apartado 2b del ejercicio 1.	9
2.2 Regresión Logística (RL)	10
2.2.1 Gráficas para la primera repetición	13
3 Bonus. Clasificación de Dígitos	14
3.1 Planteamiento del problema de clasificación binaria asociado	14
3.2 Comparación de los modelos lineales estudiados	15
3.2.1 Generar gráficos con la función estimada sobre los datos de entrenamiento y test	15
3.2.2 Calcular E_{in} y E_{test}	17

3.2.3	Repetir inicialización con los pesos obtenidos mediante regresión lineal	17
3.2.4	Obtener cotas sobre el verdadero valor de E_{out} para los 4 métodos .	19
Bibliografía		21

Sobre la complejidad de H y el ruido

En este capítulo, trataremos la dificultad que introduce la aparición de ruido en las etiquetas a la hora de elegir la clase de funciones más adecuadas.

1.1 | Dibujar gráficas de nubes de puntos simuladas

1.1.1 | Uniformemente distribuidos

Consideré $N = 50$, $\text{dim} = 2$, $\text{rango} = [-50, 50]$ con `simula_unif(N, dim, rango)`

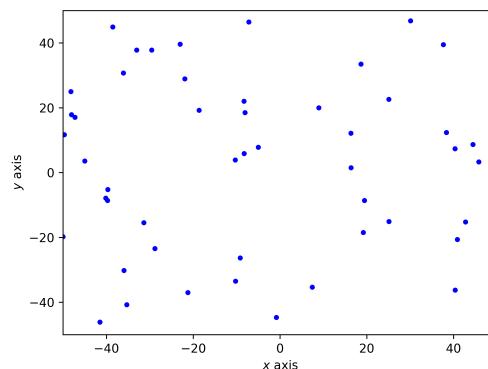


Figura 1.1: Gráfica de nube de puntos uniformemente distribuidos

1.1.2 | Siguiendo distribución gaussiana de media 0 varianza dada

Consideré $N = 50$, $dim = 2$ y $sigma = [5, 7]$ con `simula_gauss(N, dim, sigma)`

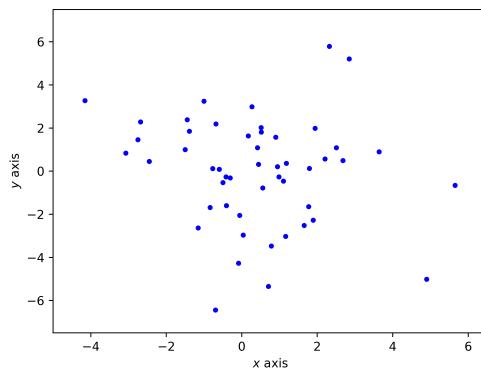


Figura 1.2: Gráfica de nube de puntos en distribución gaussiana.

Al seguir una distribución normal de media cero y varianza sigma, los puntos se acumulan en $[-\sqrt{5}, \sqrt{5}] \times [-\sqrt{7}, \sqrt{7}] \approx [-2.2, 2.2] \times [-2.6, 2.6]$

1.2 | Ejercicio 2

Vamos a valorar la influencia del ruido en la selección de la complejidad de la clase de funciones. Con ayuda de la función `simula_unif(100, 2, [-50, 50])` generamos una muestra de puntos 2D a los que vamos aadir una etiqueta usando el signo de la función $f(x, y) = y - ax - b$, es decir el signo de cada punto con respecto a la recta simulada con `simula_recta()`.

1.2.1 | Dibujo de puntos con etiqueta y recta usada

Dibujamos un gráfico 2D con los puntos clasificados por etiquetas junto con la recta usada para etiquetar.

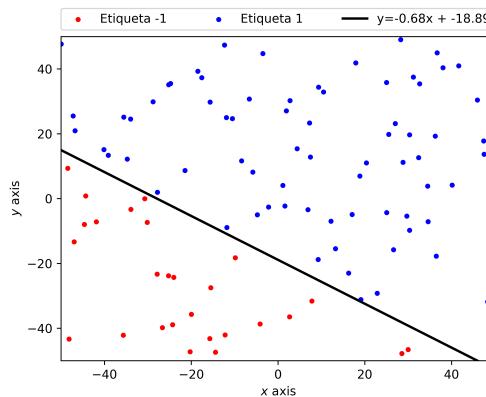


Figura 1.3: Etiquetado de puntos uniformemente distribuidos según recta.

Es claro que si usamos una recta para etiquetar los puntos en dos clases, estos datos están bien clasificados por esta recta.

1.2.2 | Añadir ruido aleatorio

Modifique de forma aleatoria un 10 % de las etiquetas positivas y otro 10 % de las negativas y guarde los puntos con sus nuevas etiquetas. Dibuje de nuevo la gráfica anterior. Ahora habrá puntos mal clasificados respecto de la recta.

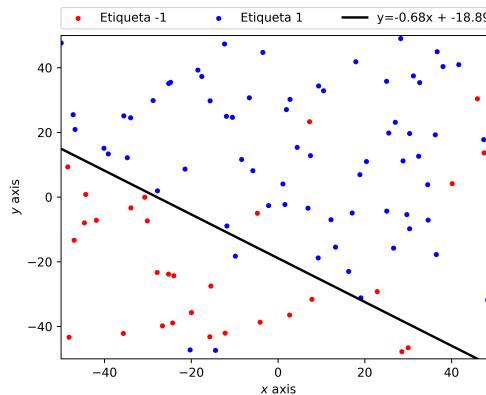


Figura 1.4: Nube de puntos anterior con 10 % de ruido en cada etiqueta.

En efecto, 3 puntos con etiqueta -1 (rojos) ahora tienen etiqueta 1 (son azules). Esto es el 10 % del total (27) redondeado.

1.2.3 | Otras fronteras de clasificación

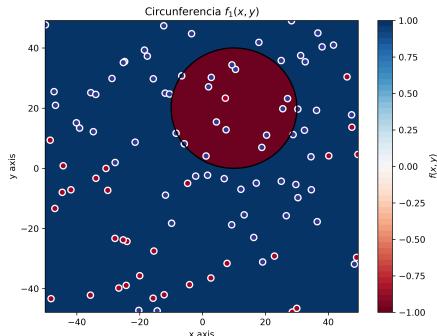
Supongamos ahora que las siguientes funciones (f_1, f_2, f_3, f_4) definen la frontera de clasificación de los puntos de la muestra en lugar de una recta.

Visualizar el etiquetado generado en el apartado 2b junto con la gráfica de cada una de las funciones. Comparar las regiones positivas y negativas de estas nuevas funciones con las obtenidas en el caso de la recta. Argumente si estas funciones más complejas son mejores clasificadores que la función lineal. Observe las gráficas y diga qué consecuencias extrae sobre la influencia de la modificación de etiquetas en el proceso de aprendizaje. Explique el razonamiento.

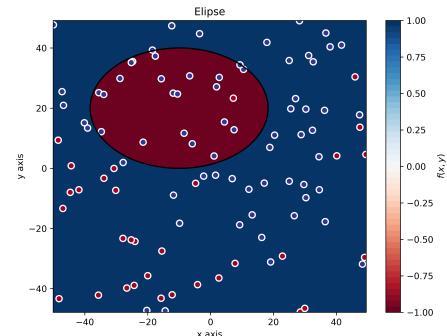
Usando la función `plot_datos_cuad` proporcionada en la plantilla de código, podemos visualizar y comparar las regiones positivas y negativas (azul y rojo respectivamente) que define la frontera dada por $f_i(x, y) = 0$ con $i = 1, 2, 3, 4$.

1.2.3.1 | Circunferencia y Elipse

$$\blacksquare f_1(x, y) = (x - 10)^2 + (y - 20)^2 - 400 \quad \blacksquare f_2(x, y) = \frac{(x+10)^2}{2} + (y - 20)^2 - 400$$



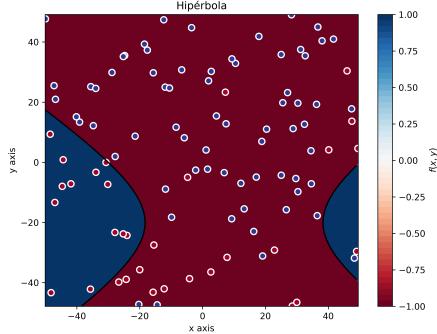
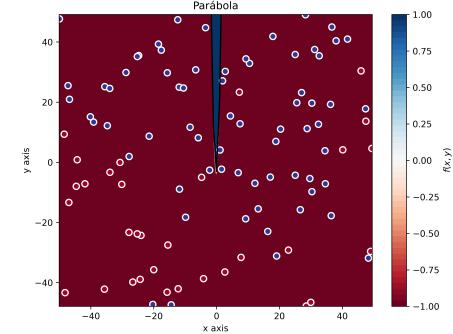
(a) Muestra clasificada por circunferencia f_1



(b) Muestra clasificada por elipse f_2

1.2.3.2 | Hipérbola y Parábola

$$\blacksquare f_3(x, y) = \frac{(x-10)^2}{2} - (y+20)^2 - 400 \quad \blacksquare f_4(x, y) = y - 20x^2 - 5x + 3$$

(a) Muestra clasificada por hipérbola f_3 (b) Muestra clasificada por parábola f_4

Si calculamos el error de clasificación en cada uno de los casos obtenemos:

f_i	Error de clasificación (%)
Recta f	10 %
Circunferencia f_1	41 %
Elipse f_2	49 %
Hipérbola f_3	84 %
Parábola f_4	73 %

Podemos confirmar lo que se podía intuir por las graficas obtenidas. A pesar de ser funciones más complejas que la lineal, no son mejores clasificadores.

En cuanto a la influencia del ruido, no es posible obtener un mejor clasificador que la recta f . Esto se debe a que el ruido ha sido generado aleatoriamente y que es la propia recta f la que se ha usado para etiquetar. Así, el 10 % es cota inferior del de error de clasificación para distinto \mathcal{H} .

De hecho, para funciones demasiado complejas es probable que tengamos un problema de **sobreajuste** (overfitting) impidiendo una buena generalización (alto valor de E_{out}).

Finalmente, nos preguntamos qué ocurriría si usamos alguna de estas funciones para clasificar los datos y luego añadimos el 10 % de ruido. Por ejemplo, etiquetando con la circunferencia es claro que resultan datos no linealmente separables. El error de clasificación con una función más sencilla (recta) será notablemente superior y de igual manera que con el clasificador lineal, será imposible encontrar circunferencia que obtenga menor error de clasificación al 10 %.

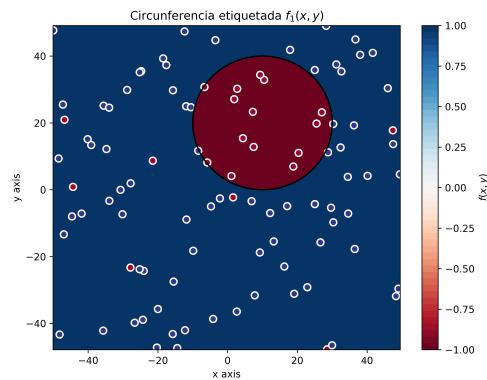


Figura 1.7: Muestra etiquetada por circunferencia con 10 % de ruido

Modelos Lineales

2.1 | Algoritmo de aprendizaje del Perceptrón (PLA)

Implementar la función `ajusta_PLA(datos, label, max_iter, vini)` que calcula el hiperplano solución a un problema de clasificación binaria usando el algoritmo PLA. La entrada `datos` es una matriz donde cada ítem con su etiqueta está representado por una fila de la matriz, `label` el vector de etiquetas (cada etiqueta es un valor $+1$ o -1), `max_iter` es el número máximo de iteraciones permitidas y `vini` el valor inicial del vector. La función devuelve los coeficientes del hiperplano.

2.1.1 | Ejecutar el algoritmo PLA con los datos empleados en el apartado 2a del ejercicio 1.

Iniciar el algoritmo con:

- el vector cero y,
- con vectores de números aleatorios en $[0, 1]$ (10 veces).

Anotar el número medio de iteraciones necesarias en ambos para converger.

Se deben mostrar en una tabla cada uno de los pesos iniciales empleados, los finales (obtenidos tras el proceso de entrenamiento), y el porcentaje de error de clasificación.

Valorar el resultado relacionando el punto de inicio con el número de iteraciones.

El algoritmo de aprendizaje del perceptrón (PLA) implementado en `ajusta_PLA` consiste en actualizar los pesos por cada punto mal clasificado de la muestra de acuerdo a la siguiente regla:

$$w(t+1) = w(t) + x_i \cdot y_i$$

Esto se hace hasta que el vector de pesos no cambie en toda una época (recorrido de la muestra) o se alcance el número máximo de iteraciones `max_iter`. La implementación realizada devuelve el histórico de pesos, el número de iteraciones dadas y el error de clasificación.

Para visualizar el proceso de convergencia de PLA, se ha implementado una clase `Animacion`. En la ejecución del código se muestra esta animación para el vector inicial nulo en los datos sin ruido, donde tras 75 iteraciones se encuentra un clasificador con 0 % de error.

Vector inicial	Iteraciones	Coeficientes w	Error de clasificación
[0.000, 0.000, 0.000]	75	[661.00, 23.20, 32.39]	0 %
[0.574, 0.349, 0.057]	257	[1115.57, 43.48, 62.12]	0 %
[0.229, 0.664, 0.497]	43	[464.23, 15.39, 23.75]	0 %
[0.519, 0.175, 0.571]	231	[1078.52, 39.47, 53.76]	0 %
[0.997, 0.817, 0.594]	71	[664.00, 23.15, 31.90]	0 %
[0.976, 0.902, 0.596]	76	[661.98, 24.90, 36.20]	0 %
[0.032, 0.094, 0.065]	59	[558.03, 19.36, 29.71]	0 %
[0.452, 0.375, 0.975]	274	[1145.45, 40.28, 60.81]	0 %
[0.168, 0.973, 0.767]	235	[1089.17, 39.45, 53.53]	0 %
[0.824, 0.633, 0.669]	257	[1148.82, 39.90, 60.95]	0 %
[0.477, 0.013, 0.353]	74	[673.48, 22.59, 31.35]	0 %
Promedio	157.7		0 %

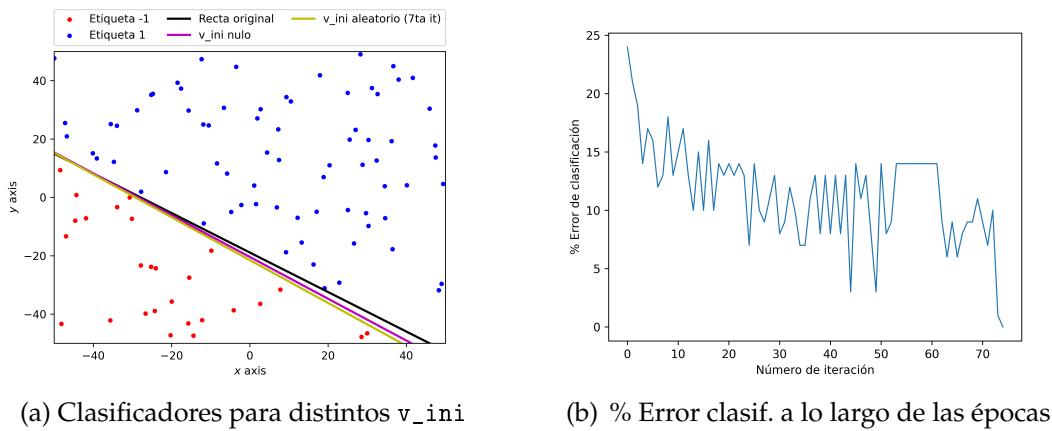
Cuadro 2.1: Resultados de ejecución de PLA **muestra sin ruido** para vector inicial nulo y generados aleatoriamente (10 iteraciones)

Podemos observar en los resultados, que al ser los datos **linealmente separables**, PLA converge en todos los casos a una recta con error de clasificación cero en una media de ≈ 158 iteraciones (vectores iniciales aleatorios).

Si hallamos la desviación típica del número de iteraciones, obtenemos 94.17, un valor considerablemente alto que indica la alta sensibilidad de la convergencia (nº de iteraciones en llegar al 0 % de error) con respecto al vector inicial. Podemos afirmar por tanto, que escoger el vector nulo como vector inicial resulta ser una buena heurística.

En las siguientes figuras, vemos los clasificadores obtenidos para el vector nulo, el vector generado aleatoriamente en la iteración 7 y el clasificador original. Le sigue una gráfica para visualizar la disminución del error de clasificación a lo largo de las iteraciones.

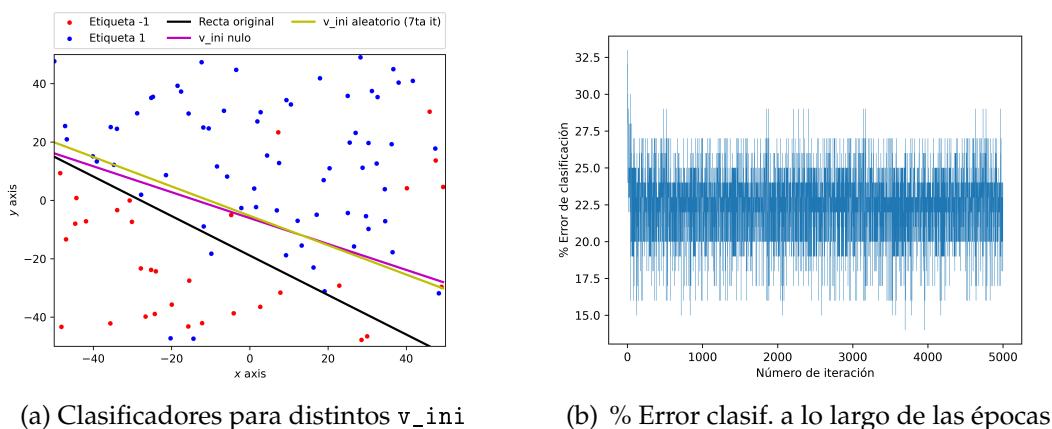
Figura 2.1: Muestra linealmente separable (sin ruido)



2.1.2 | Repetir usando los datos del apartado 2b del ejercicio 1.

¿Observa algún comportamiento diferente? En caso afirmativo diga cuál y las razones para que ello ocurra.

Figura 2.2: Muestra con ruido (no linealmente separable)



En este caso, partimos de la muestra de datos con 10 % de ruido. Así, PLA no tiene garantizada una convergencia pues los datos **no son linealmente separables** y la cota inferior del error de clasificación para las rectas obtenidas con PLA es justo ese 10 %.

Por tanto, las ejecución se para al alcanzar el máximo número de iteraciones (épocas) que se ha fijado en 5000.

Vector inicial	Iteraciones	Coeficientes w	Error de clasificación
[0.000, 0.000, 0.000]	5000	[339.00, 24.81, 55.86]	24 %
[0.492, 0.730, 0.469]	5000	[348.49, 17.77, 42.43]	24 %
[0.457, 0.138, 0.011]	5000	[349.46, 16.66, 44.13]	24 %
[0.758, 0.320, 0.984]	5000	[339.76, 21.61, 56.29]	22 %
[0.220, 0.339, 0.524]	5000	[336.22, 17.72, 43.85]	25 %
[0.755, 0.464, 0.125]	5000	[336.75, 17.32, 53.11]	27 %
[0.313, 0.505, 0.674]	5000	[339.31, 15.10, 26.87]	20 %
[0.770, 0.130, 0.023]	5000	[333.77, 27.72, 49.69]	20 %
[0.519, 0.810, 0.013]	5000	[343.52, 12.09, 38.81]	22 %
[0.672, 0.687, 0.449]	5000	[331.67, 12.87, 21.73]	19 %
[0.915, 0.644, 0.005]	5000	[341.91, 21.10, 56.13]	22 %
Promedio	5000		22.5 %

Cuadro 2.2: Resultados de ejecución de PLA **muestra con ruido** para vector inicial nulo y generados aleatoriamente (10 iteraciones)

Al contrario que en el caso anterior (muestra sin ruido), ejecutar el algoritmo con vector nulo como vector inicial produce un error de clasificación mayor que la media obtenida para los 10 vectores iniciales generados aleatoriamente (24 % vs 22.5 %). Además, en este caso, por lo comentado anteriormente, el vector inicial no influye en el número de iteraciones.

En cuanto a la variación del error por número de iteraciones, podemos observar grandes oscilaciones, estando concentrados la mayoría en el intervalo $22.5 \pm 2.5\%$. El mínimo se alcanza entorno a la iteración 4000 donde el error llega a ser 14 %, bastante cerca de la cota inferior del 10 % antes mencionada.

2.2 | Regresión Logística (RL)

En este ejercicio emplearemos nuestra propia función objetivo f y un conjunto de datos D para ver cómo funciona regresión logística. Consideraremos $d = 2$ para que los datos sean fácilmente visualizables, y emplearemos $X = [0, 2] \times [0, 2]$ con probabilidad uni-

forme de elegir cada $x \in X$. Elegir una línea en el plano que pase por X como la frontera que separa la región en donde y toma valores $+1$ y -1 .

Para ello, seleccionar dos puntos aleatorios de X y calcular la línea que pasa por ambos.

Impleméntese RL con Gradiente Descendente Estocástico (SGD) del siguiente modo:

- Inicializar el vector de pesos con valores 0.
- Parar el algoritmo cuando $\|w(t+1) - w(t)\| < 0.01$, donde $w(t)$ denota el vector de pesos al final de la época t . Recuérdese que una época es un pase completo a través de los N ejemplos de nuestro conjunto de datos.
- Aplicar una permutación aleatoria de $\{1, 2, \dots, N\}$ a los índices de los datos, antes de usarlos en cada época del algoritmo.

A continuación, empleando la implementación anterior, realícese el siguiente experimento:

- Seleccione $N = 100$ puntos aleatorios $\{x_n\}$ de X y evalúe las respuestas $\{y_n\}$ de todos ellos respecto de la frontera elegida.
- Ejecute RL para encontrar la función solución g , y evalúe el error E_{out} usando para ello una nueva muestra de datos (> 999). Se debe escoger experimentalmente tanto el learning rate (tasa de aprendizaje η) como el tamaño de batch.
- Repita el experimento 100 veces, y calcule los valores promedio de E_{out} , de porcentaje de error de clasificación, y de épocas necesarias para converger.

Recordamos que la Regresión Logística es un modelo lineal de clasificación, donde ahora la función hipótesis es de la forma

$$h(x) = \theta(w^T X) \quad (2.1)$$

con $\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$ la llamada **función logística o sigmoide** que toma valores en $[0, 1]$. Estos se interpretan de forma probabilística para un evento binario. A través del método de máxima verosimilitud (Yaser S. Abu-Mostafa (2012)), obtenemos una medida de error en la muestra:

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^N \ln(1 + e^{y_n w^T x_n}) \quad (2.2)$$

Intuitivamente, el error se acerca más a cero cuanto mayor es cada $y_n w^T x_n$ (positivo). Esto implicaría que $\text{sign}(w^T x_n) = y_n$, es decir, un número mayor de instancias x_n bien clasificadas.

Para entrenar el modelo, aplicamos **gradiente descendente estocástico (SGD)** en su versión **batch de 1 elemento**. Es decir, por cada época hallamos una permutación de las instancias x_n (índices en la implementación), computamos el gradiente de cada instancia y actualizamos los pesos. La expresión del gradiente para batch de tamaño $M = 1$ es (Yaser S. Abu-Mostafa (2012)):

$$\nabla E_{in}(w) = \frac{-1}{M} \sum_{n=1}^M \frac{y_n x_n}{1 + e^{y_n w^T x_n}} = \frac{-y_n x_n}{1 + e^{y_n w^T x_n}} \quad (2.3)$$

La implementación del algoritmo se encuentra en la función `sgdRL(X, y, lr, max_iter, epsilon=0.01)` y se ha seguido el criterio de parada descrito en el enunciado $\|w(t+1) - w(t)\| < 0.01$ parametrizando 0.01 como `epsilon` por si se quisiera experimentar con otros valores en un futuro.

En cuanto a la realización del experimento, se ha decidido generar la muestra de test X_{test} con 1000 elementos (> 999) para no incrementar demasiado el tiempo de ejecución en las 100 repeticiones. Todas las llamadas al algoritmo de aprendizaje de Regresión Lineal se han realizado con un **learning rate** $\eta = 0.01$, **un número máximo de 1000 iteraciones y tamaño de batch de 1 elemento**.

En efecto, se ha comprobado experimentalmente que valores mayores de η (0.2, 0.4, 0.6) implican un mayor número de iteraciones debido al criterio de parada (la distancia entre los vectores de pesos tarda más en disminuir), aunque menor error E_{in} .

Por otro lado, se implementó inicialmente el algoritmo de aprendizaje de regresión logística en su versión SGD minibatch con tamaño de batch superiores a 1 (2, 4, 16, 32, 64), ofreciendo menor número de épocas aunque mayor error.

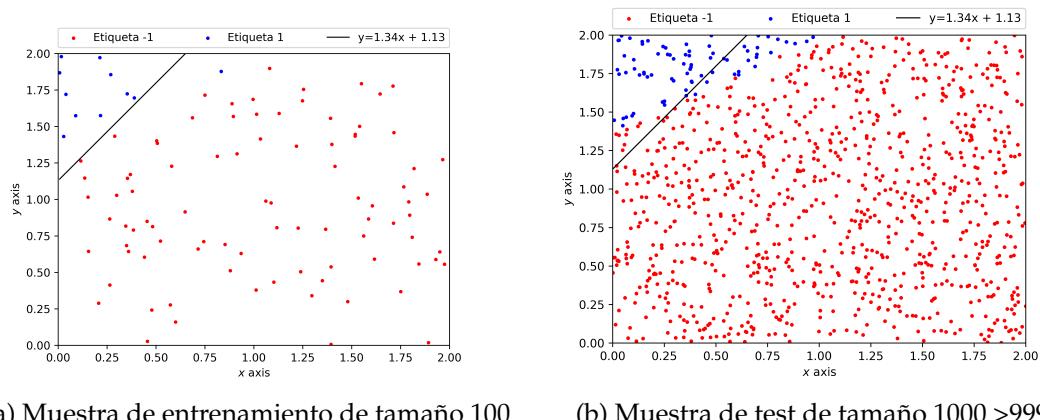
Repetición	E_{out}	E_{out}^{clas} (%)	Épocas para converger
0	0.086	3.2 %	311
1	0.089	1.3 %	355
2	0.118	2.6 %	500
3	0.124	1.5 %	383
4	0.133	2.7 %	451
...
96	0.092	2.6 %	374
97	0.090	2.1 %	426
98	0.146	5.3 %	476
99	0.111	1.0 %	410
Promedio	0.118	2.52 %	413.5

Cuadro 2.3: Resultados de ejecución de 100 experimentos de Regresión Logística

Además, las épocas extremas alcanzadas en el experimento son de 198 y 576. Y la desviación típica es considerable ($\sigma = 77.97$ éps.) que asociamos al componente aleatorio.

2.2.1 | Gráficas para la primera repetición

Figura 2.3: Regresión logística aplicada a dos muestras uniformes



Se concluye tras el experimento que gradiente descendente estocástico para Regresión Lineal obtiene buenos resultados (97.5 % de precisión media en test) en un número razonable de iteraciones (promedio de 413).

Bonus. Clasificación de Dígitos

Clasificación de Dígitos. Considerar el conjunto de datos de dígitos manuscritos, y seleccionar las muestras de los dígitos 4 y 8. Extraer las características de intensidad promedio y simetría en la manera que se indicó en el ejercicio 3 de la práctica anterior.

3.1 | Planteamiento del problema de clasificación binaria asociado

Debe considerar el conjunto de entrenamiento como datos de entrada para aprender la función g .

El siguiente conjunto $\{\mathcal{X}, \mathcal{Y}, \mathcal{D}, f : \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{A}_i, \mathcal{H}, g\}$ define 4 problemas de clasificación binaria con $i = 1, 2, 3, 4$ siendo los elementos:

- $\mathcal{X} = \{1\} \times \mathbb{R}^2$
- $\mathcal{Y} = \{-1, 1\}$ donde -1 representa el dígito 4 y 1 el 8.
- $\mathcal{D} = \{(x_n, y_n) : n = 1, \dots, N\} \subset \mathcal{X} \times \mathcal{Y}$ es el conjunto de N datos de entrenamiento dentro del espacio de entrada con su correspondiente salida (problema de clasificación binaria, aprendizaje supervisado).
- $f : \mathcal{X} \rightarrow \mathcal{Y}$ función objetivo, completamente desconocida que aproximaremos a partir de los datos de entrenamiento.
- $\mathcal{H} = \{h \in \mathbb{R}^3 \rightarrow \mathbb{R} : h(x) = \text{sign}(w^T x), w \in \mathbb{R}^3\}$: conjunto de hipótesis a partir del cual estimaremos $g \in \mathcal{H}$ suponiendo distribución de probabilidad \mathcal{P} en $\mathcal{X} \times \mathcal{Y}$ de acuerdo a la que se obtiene las N instancias de \mathcal{D} de forma independientemente e idénticamente distribuida.

■ \mathcal{A}_i : algoritmo de aprendizaje

\mathcal{A}_1 : Algoritmo de la Pseudoinversa para Regresión Lineal

\mathcal{A}_2 : Algoritmo de Aprendizaje del Perceptrón (PLA)

\mathcal{A}_3 : Algoritmo SGD para Regresión Lineal (RL)

\mathcal{A}_4 : Algoritmo PLA-POCKET

■ $g \in \mathcal{H}$: función del conjunto de hipótesis que usamos para aproximar f

■ \mathcal{L}_i : función de pérdida (usada por \mathcal{A} para aproximar g)

\mathcal{L}_1 : Error cuadrático medio

$\mathcal{L}_2 = \mathcal{L}_4$: Error de clasificación

\mathcal{L}_3 : Error de entropía cruzada (véase apartado Reg. Log.)

3.2 | Comparación de los modelos lineales estudiados

Compárense los modelos de regresión lineal, PLA, RL y PLA-Pocket.

3.2.1 | Generar gráficos con la función estimada sobre los datos de entrenamiento y test

Figura 3.1: Regresión Lineal

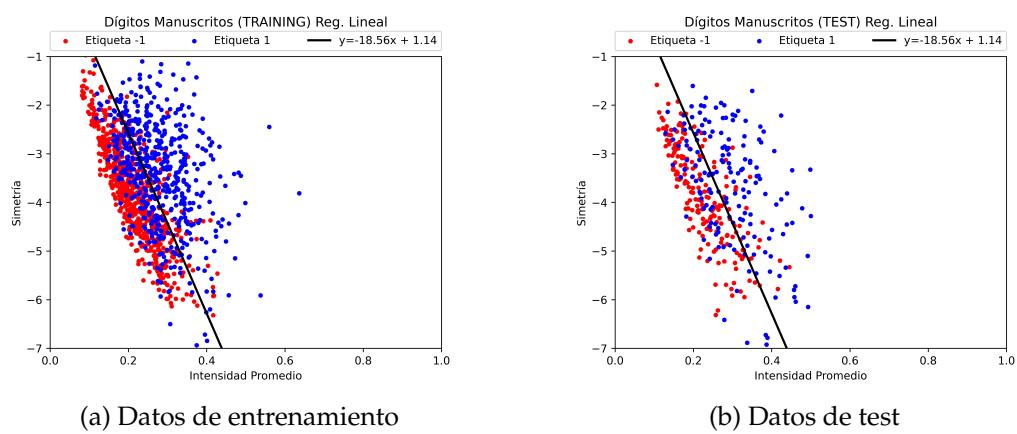


Figura 3.2: PLA - Algoritmo de aprendizaje del perceptrón

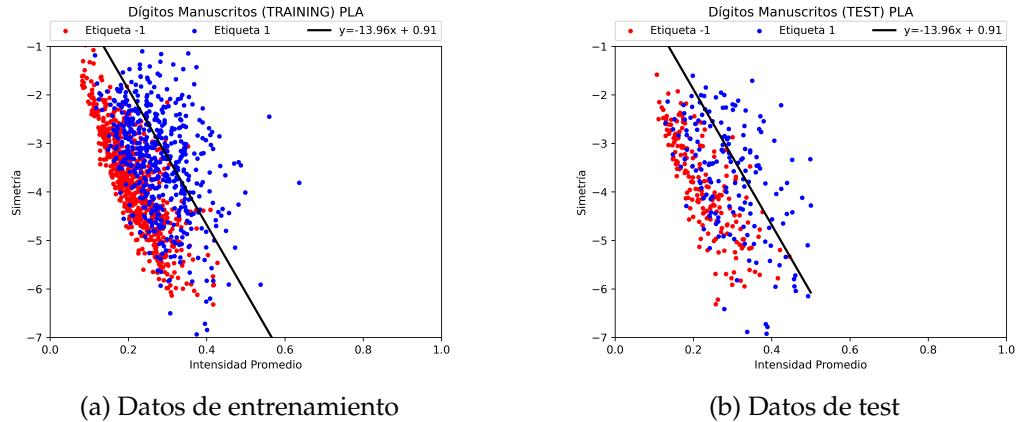


Figura 3.3: RL - Regresión Logística para clasificación lineal

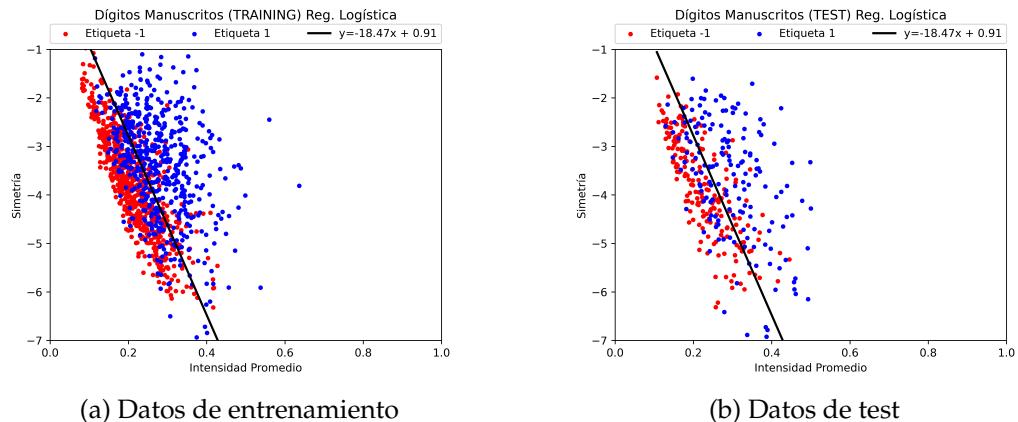
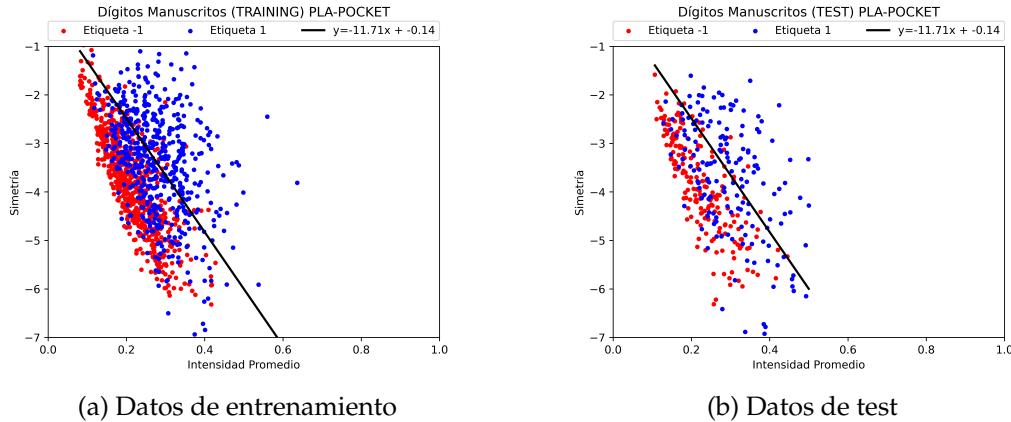


Figura 3.4: Algoritmo PLA-POCKET



Nos detenemos en **PLA-POCKET** para comentar que su implementación es parecida a la de PLA, con la salvedad de que se devuelve el vector de pesos de salida que mejor E_{in} proporcione en las iteraciones. Es decir, se reemplaza sucesivamente w si se calcula otro con menor error *in-sample* asociado.

3.2.2 | Calcular E_{in} y E_{test}

Algoritmo	E_{in}	E_{test}	E_{in}^{clas} (%)	E_{test}^{clas} (%)	Iteraciones
Regresión Lineal	0.643	0.709	22.8 %	25.1 %	—
PLA	0.309	0.301	30.9 %	30.1 %	1000
Regresión Logística	0.464	0.527	22.2 %	26.0 %	552
PLA-POCKET	0.264	0.281	26.4 %	28.1 %	36

Cuadro 3.1: Comparación de los modelos lineales estudiados

3.2.3 | Repetir inicialización con los pesos obtenidos mediante regresión lineal

Si se emplean los pesos obtenidos con regresión lineal para inicializar los otros tres métodos (RL, PLA, PLA-pocket), ¿se observa alguna mejora en los resultados a algún nivel? Justifique su respuesta

Denominamos w_{lin} a los pesos obtenidos tras aplicar regresión lineal al conjunto de datos de entrenamiento.

Figura 3.5: PLA inicializado con w_{lin}

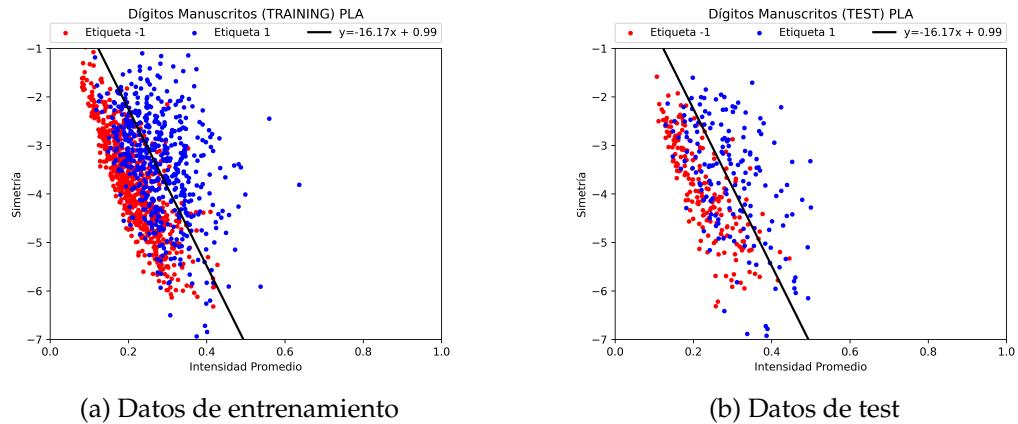


Figura 3.6: RL inicializado con w_{lin}

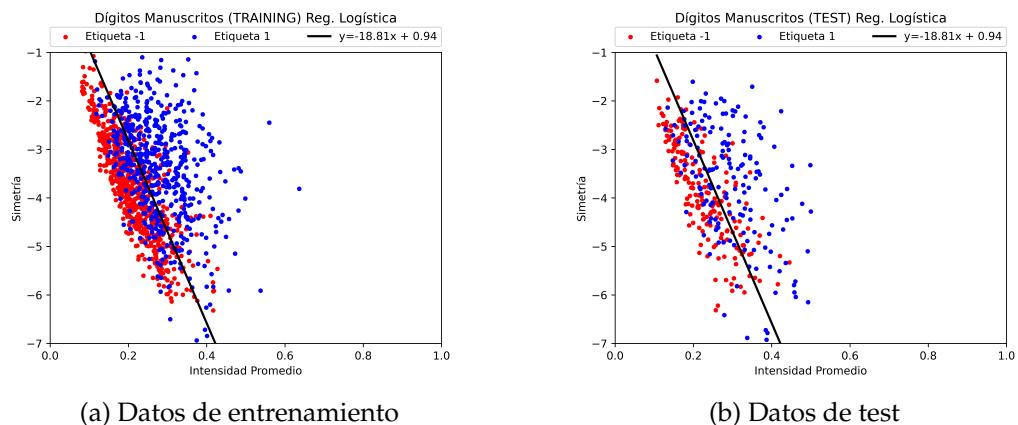
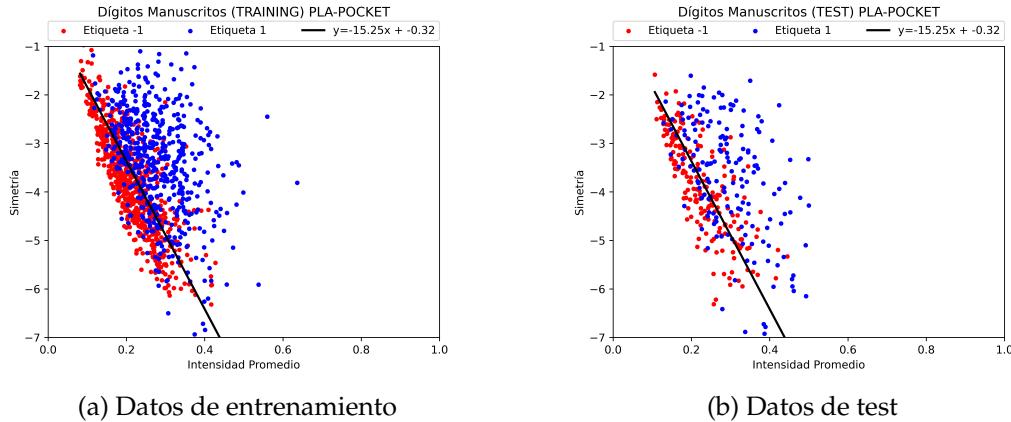


Figura 3.7: PLA-POCKET inicializado con w_{lin}


Algoritmo	E_{in}	E_{test}	E_{in}^{clas} (%)	E_{test}^{clas} (%)	Iteraciones
PLA	0.252	0.251	25.2 %	25.1 %	1000
RL	0.465	0.531	22.2 %	26.0 %	557
PLA-POCKET	0.246	0.322	24.6 %	32.2 %	2

 Cuadro 3.2: Comparación de modelos lineales iterativos inicializados con w_{lin}

En efecto, **observamos mejora en los resultados de PLA inicializado con w_{lin}** . En este caso se clasifica correctamente al (75 %) frente al (70 %) de precisión obtenido fuera de la muestra con vector inicial nulo. Sin embargo, el error out of sample aumenta ligeramente en Regresión Logística y PLA-POCKET con esta inicialización (w_{lin}) frente al vector nulo.

3.2.4 | Obtener cotas sobre el verdadero valor de E_{out} para los 4 métodos

Calcúlense dos cotas: una basada en E_{in} y otra basada en E_{test} . Usar una tolerancia $\delta = 0.05$. ¿Que cota es mejor? Justifique la respuesta.

3.2.4.1 | Cota de Vapnik-Chervonenkis

Recordamos que para $\delta > 0$, obtenemos cota de E_{out} proporcionada por la expresión:

$$E_{out}(g) \leq E_{in} + \sqrt{\frac{8}{N} \log \left(\frac{4[(2N)^{d_{VC}} + 1]}{\delta} \right)} \quad (3.1)$$

con probabilidad $1 - \delta$ siendo d_{VC} la dimensión de Vapnik-Chervonenkis del modelo y N el tamaño de \mathcal{D}_{test} .

Para los 4 problemas de clasificación descritos en el primer apartado, $d_{VC} = 3$ de acuerdo al conjunto de hipótesis \mathcal{H} y el tamaño de la muestra de test de clasificación de dígitos es $N = 366$. Tomando $\delta = 0.05$, obtenemos

- Pseudoinversa: $E_{out} \leq 0.9548$
- PLA (w_{lin}): $E_{out} \leq 0.9788$
- RL: $E_{out} \leq 0.9468$
- PLA-POCKET: $E_{out} \leq 0.9908$

lo cual no aporta gran información (entre 1 % y 5.4 % de precisión garantizada).

3.2.4.2 | Desigualdad de Hoeffding

Para $\delta > 0$,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \log \left(\frac{2|\mathcal{H}|}{\delta} \right)} \quad (3.2)$$

con probabilidad $1 - \delta$. Siendo N el tamaño de \mathcal{D}_{test} y $|\mathcal{H}|$ el tamaño del conjunto de hipótesis. Podemos calcular la cota a partir de E_{test} (en lugar de E_{in}) teniendo en cuenta que esta se verifica para g prefijado. Así $|\mathcal{H}| = 1$ y obtenemos:

- Pseudoinversa: $E_{out} \leq 0.3220$
- PLA (w_{lin}): $E_{out} \leq 0.3720$
- RL: $E_{out} \leq E_{in} \leq 0.3220$
- PLA-POCKET: $E_{out} \leq 0.3520$

Concluimos por tanto, que las cotas obtenidas a partir de la desigualdad de Hoeffding mediante E_{test} son mejores que las de la dimensión VC a pesar de ser E_{test} superior a E_{in} . Estas cotas nos ofrecen un porcentaje de precisión mínimo del 68 % fuera de la muestra.

Bibliografía

Hsuan-Tien Lin Yaser S. Abu-Mostafa, Malik Magdon-Ismail. *Learning From Data. A Short Course*. AMLbook, 2012. URL AMLbook.com.