Aprendizaje automático. Cuestionario 1

Antonio Jesús Heredia Castillo
76069518P

---

1. Identificar, para cada una de las siguientes tareas, cual es el prioblema, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje (X , f, Ym) que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los elementos para cada tipo.

   a) Clasificación automática de cartas por distrito postal.

   Respuesta:
   Esto es un caso típico de aprendizaje supervisado. En este caso se creara un modelo capaz de reconocer los diferentes números y así poder saber el distrito postal al que pertenece la carta. En este caso los valores de entrada serán datos de la imagen del dígito. Estos valores puede ser como hemos visto en practicas la intensidad promedio y la simetría de la misma. Por otro lado la salida seria cual es el dígito que estamos clasificando.

   - X = simetría, intensidad promedio
   - f = función desconocida
   - Y = 0,1,2,3,4,5,6,7,8,9

   b) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.

   Respuesta:
   Aunque a priori puede parecer fácilmente abordable por aprendizaje supervisado, creo que seria mas conveniente abordarlo por aprendizaje reforzado. Ya que en la bolsa influye demasiadas variables y seria difícil tenerlas todas en cuenta por nosotros. En cambio para un modelo de aprendizaje por refuerzo seria mas fácil encontrar los patrones que se dan cuando recibe un "refuerzo" positivo o negativo(que suba o baje el valor) .

   c) Hacer que un dron sea capaz de rodear un obstáculo

   Respuesta:
   En este caso usaría aprendizaje por refuerzo. Usaría un simulador de drones en una computadora, para que el dron real no tuviera daños. La "recompensa" recibida seria superar el obstáculo. De esta forma no necesitaríamos ningún conjunto de datos anterior para poder realizar el entrenamiento del dron.

   d) Dada una colección de fotos de perros, posiblemente de distintas razas, establecer cuantas razas distintas hay representadas en la colección.

Respuesta:
Para este supuesto intentaría utilizar aprendizaje no supervisado ya que no nos importa que raza tiene cada perro, si no la cantidad de razas. De esta forma la maquina se encargara de encontrar semejanzas entre las diferentes fotografiás y las agrupara de manera natural según el perro que aparece, evitando tener datos etiquetados.

2. ¿Cuales de los siguientes problemas son más adecuados para una aproximación por aprendizaje y cuales más adecuados para una aproximación por diseño? Justificar la decisión

   a) Determinar si un vertebrado es mamífero, reptil, ave, anfibio o pez.

   Respuesta:
   En este acaso usaría aproximación por diseño, ya que cada clase de vertebrados tienen características comunes que la diferencia de otras clases. Estas caracterizaras de cada clase son ya muy conocidas (por ejemplo, los mamíferos tienen glándulas mamarias) y seria fácil crear un algoritmo que lo resolviera.

   b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.

   Respuesta:
   Este problema intentaría abordarlo con una aproximación por diseño. Los virus pueden cambiar de un año para otro e incluso aparecer enfermedades nuevas o que no tengamos suficientes datos estadísticos como para entrenar a un modelo de aprendizaje. Por tanto usando el conocimiento de un infectologo podríamos crear un sistema algorítmico que prediga cuando seria necesario realizar una campaña de evacuación.

   c) Determinar perfiles de consumidor en una cadena de supermercados.

   Respuesta:
   Aquí, al tener que hacer "grupos" de consumidores, podríamos usar una aproximación por aprendizaje. Este se podría encargar de buscar semejanzas entre los diferentes consumidores y agrupar los que mas se parezcan entre si.

   d) Determinar el estado anímico de una persona a partir de una foto de su cara.

   Respuesta:
   También elegiría una aproximación por aprendizaje ya que, podemos proveer con un gran conjunto de datos etiquetados al sistema y que este se encargara de analizar las diferentes variables que tuviera las imágenes proporcionadas y poder predecir con una imagen nueva el estado animico de la persona que aparece.

   e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.

   Respuesta:
   Esto seria un problema típico para resolver con aproximación por diseño. Sabemos bien como afecta las distintas variables al problema (si hay coches pasando, si hay alguien esperando para cruzar la calle, etc), con estos datos seria fácil ajustar un algoritmo que satisfaga las necesidades de los usuarios.

3. Construir un problema de aprendizaje desde datos para un problema de clasificación de fruta en una explotación agraria que produce mangos, papayas y guayabas. Identificar los

siguientes elementos formales X , Y, D, f del problema. Dar una descripción de los mismos que pueda ser usada por un computador. ¿Considera que en este problema estamos ante un caso de etiquetas con ruido o sin ruido? Justificar las respuestas.

Respuesta:

a) **X** : las variables $\{x_1, x_2, ..., x_n\}$ con las que podemos definir las distintas frutas. En este caso puede ser color, tamaño, simetría, etc.

b) **Y**: Al ser un problema de clasificación discreto he elegido esta representación. $\mathcal{Y} = y \in \{mango, papaya, guayaba\}$

c) El conjunto de datos seria $D = \{(x_1, y_1), ..., (x_N, y_N) y_i = f(x_i), i = 1, 2, 3, ..., N / x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$

d) La función f es desconocida y es la que intentamos buscar a partir del aprendizaje de los datos. $f : \mathcal{X} \mapsto \mathcal{Y}$

4. Suponga una matriz cuadrada A que admita la descomposición $A = X^T X$ para alguna matriz X de números reales. Establezca una relación entre los valores singulares de las matriz A y los valores singulares de X.

Respuesta:
Consideraremos la SVD de X como $X = UDV^T$, ademas sabemos según las diapositivas (pagina 15/Sesión 2 Modelos lineales), $X^T X = UDDV^T$. Por lo tanto como tenemos que $A = X^T X = UDDV^T$, los valores singulares de la matriz A serán el cuadrado de los valores singulares de los de X, ya que D es una matriz diagonal de $n \times n$ y la multiplicación de D por ella misma es el cuadrado de cada valor de su diagonal. Esto es una propiedad de la matrices diagonales.

5. Sean x e y dos vectores de características de dimensión $M \times 1$. La expresión $\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} (x_i - \overline{x})(y_i - \overline{y})$ define la covarianza entre dichos vectores, donde $\overline{z}$ representa el valor medio de los elementos de $\mathbf{z}$. Considere ahora una matriz X cuyas columnas representan vectores de características. La matriz de covarianzas asociada a la matriz $X = (x_1, x_2, , x_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

$$\text{cov(X)} = \begin{pmatrix} \text{cov}(\mathbf{x}_1, \mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \dots & \text{cov}(\mathbf{x}_1, \mathbf{x}_N) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_2, \mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(\mathbf{x}_N, \mathbf{x}_1) & \text{cov}(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}$$

Sea $1_M^T = (1, 1, \cdots, 1)$ un vector M de unos. Mostrar que representan las siguientes: expresiones

a) $E1 = 11^T \text{X}$

Respuesta:

$$E_1 = \begin{pmatrix} 1_1 \\ 1_2 \\ \cdots \\ 1_m \end{pmatrix} \times \begin{pmatrix} 1_1 & 1_2 & \cdots & 1_m \end{pmatrix} \times \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix}$$

$$= \begin{pmatrix} 1_{11} & 1_{21} & \cdots & 1_{m1} \\ 1_{12} & 1_{22} & \cdots & 1_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ 1_{1m} & 1_{2m} & \cdots & 1_{mm} \end{pmatrix} \times \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{m} x_{1i} & \sum_{i=1}^{m} x_{2i} & \cdots & \sum_{i=1}^{m} x_{ni} \\ \sum_{i=1}^{m} x_{1i} & \sum_{i=1}^{m} x_{2i} & \cdots & \sum_{i=1}^{m} x_{ni} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{m} x_{1i} & \sum_{i=1}^{m} x_{2i} & \cdots & \sum_{i=1}^{m} x_{ni} \end{pmatrix}$$

b) $E2 = \left(\mathrm{X} - \frac{1}{M}E1\right)^T \left(\mathrm{X} - \frac{1}{M}E1\right)$

Respuesta:

$$\left(\frac{1}{m}E1\right) = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} x_{1i} & \frac{1}{m}\sum_{i=1}^{m} x_{2i} & \cdots & \frac{1}{m}\sum_{i=1}^{m} x_{ni} \\ \frac{1}{m}\sum_{i=1}^{m} x_{1i} & \frac{1}{m}\sum_{i=1}^{m} x_{2i} & \cdots & \frac{1}{m}\sum_{i=1}^{m} x_{ni} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{m}\sum_{i=1}^{m} x_{1i} & \frac{1}{m}\sum_{i=1}^{m} x_{2i} & \cdots & \frac{1}{m}\sum_{i=1}^{m} x_{ni} \end{pmatrix} = \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_N \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_N \\ \cdots & \cdots & \cdots & \cdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_N \end{pmatrix}$$

$$\left(X - \frac{1}{m}E1\right) = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix} - \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \\ \cdots & \cdots & \cdots & \cdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{21} - \overline{x}_2 & \cdots & x_{n1} - \overline{x}_n \\ x_{12} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{m2} - \overline{x}_n \\ \cdots & \cdots & \cdots & \cdots \\ x_{1m} - \overline{x}_1 & x_{2m} - \overline{x}_2 & \cdots & x_{nm} - \overline{x}_n \end{pmatrix}$$

$$\left(X - \frac{1}{m}E1\right)^T = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_1 & \cdots & x_{1m} - \overline{x}_1 \\ x_{21} - \overline{x}_2 & x_{22} - \overline{x}_2 & \cdots & x_{2m} - \overline{x}_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} - \overline{x}_n & x_{n2} - \overline{x}_n & \cdots & x_{nm} - \overline{x}_n \end{pmatrix}$$

$$E2 = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_1 & \cdots & x_{1m} - \overline{x}_1 \\ x_{21} - \overline{x}_2 & x_{22} - \overline{x}_2 & \cdots & x_{2m} - \overline{x}_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} - \overline{x}_n & x_{n2} - \overline{x}_n & \cdots & x_{nm} - \overline{x}_n \end{pmatrix} \begin{pmatrix} x_{11} - \overline{x}_1 & x_{21} - \overline{x}_2 & \cdots & x_{n1} - \overline{x}_n \\ x_{12} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{m2} - \overline{x}_n \\ \cdots & \cdots & \cdots & \cdots \\ x_{1m} - \overline{x}_1 & x_{2m} - \overline{x}_2 & \cdots & x_{nm} - \overline{x}_n \end{pmatrix}$$

$$= m \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m}(x_{1i} - \overline{x}_1)(x_{1i} - \overline{x}_1) & \frac{1}{m}\sum_{i=1}^{m}(x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2) & \cdots & \frac{1}{m}\sum_{i=1}^{m}(x_{1i} - \overline{x}_1)(x_{ni} - \overline{x}_n) \\ \frac{1}{m}\sum_{i=1}^{m}(x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2) & \frac{1}{m}\sum_{i=1}^{m}(x_{2i} - \overline{x}_2)(x_{2i} - \overline{x}_2) & \cdots & \frac{1}{m}\sum_{i=1}^{m}(x_{2i} - \overline{x}_2)(x_{ni} - \overline{x}_m) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{m}\sum_{i=1}^{m}(x_{ni} - \overline{x}_n)(x_{1i} - \overline{x}_1) & \frac{1}{m}\sum_{i=1}^{m}(x_{ni} - \overline{x}_n)(x_{2i} - \overline{x}_2) & \cdots & \frac{1}{m}\sum_{i=1}^{m}(x_{ni} - \overline{x}_n)(x_{ni} - \overline{x}_n) \end{pmatrix}$$

$$= m \begin{pmatrix} \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_n) \\ \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_n) \\ \mathrm{cov}(\mathbf{x}_n, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_n, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}$$

$$= m \cdot \mathrm{cov}(X)$$

$$E2 = m \cdot \mathrm{cov}(X)$$

6. Considerar la matriz **hat** definida en regresión, $\hat{\mathrm{H}} = \mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}$, donde X es la matriz de observaciones de dimensión $N \times (d+1), \mathrm{yX^TX}$, y X^TX es invertible.

   a) ¿Que representa la matriz $\hat{\mathrm{H}}$ en un modelo de regresión?

   Respuesta: Según Learning from Data[1], la matriz $\hat{H}$ representa cuando $\nabla E_{\mathrm{in}}(\mathbf{w}) = 0$. Ademas como el enunciado nos indica que X^TX es invertible, tenemos que $w = y\hat{H}$ es una solución única y optima. También podemos ver $\hat{H}$, como la matriz proyección que expresa los valores de las observaciones en la variable independiente y, en términos de combinación linear de ls columnas vector de la matriz modelo, X, que contiene las observaciones para cada una de las múltiples variables de las que estamos haciendo la regresión.[2]

   b) Identifique la propiedad mas relevante de dicha matriz en relación con regresión lineal.

   Respuesta:
   La propiedad mas relevante de la matriz es la idempotencia ($H = H$).Ya que facilita el analisisde errores de la regresión lineal dentro y fuera de la muestra.[1]

7. La regla de adaptación de los pesos del Perceptron ($\mathbf{w}_{new} = \mathbf{w}_{old} + y\mathbf{x}$) tiene la interesante propiedad de que mueve el vector de pesos en la dirección adecuada para clasificar $\mathbf{x}$ de forma correcta. Suponga el vector de pesos $\mathbf{w}$ de un modelo y un dato $\mathbf{x(t)}$ mal clasificado respecto de dicho modelo. Probar matemáticamente que el movimiento de la regla de adaptación de pesos siempre produce un movimiento de $\mathbf{w}$ en la dirección correcta para clasificar bien $\mathbf{x(t)}$.

8. Sea un problema probabilistico de clasificación binaria con etiquetas $\{0, 1\}$, es decir $\mathrm{P(Y} = 1) = h(\mathrm{x}) \mathrm{yP(Y} = 0) = 1 - h(\mathrm{x})$, para una funcion $h()$ dependiente de la muestra.

   a) Considere una muestra i.i.d. de tamaño $\mathrm{N}\left(\mathrm{x}_1, \cdots, \mathrm{x}_N\right)$. Mostrar que la funcion h que maximiza la verosimilitud de la muestra es la misma que minimiza

$$E_{\mathrm{in}}(\mathbf{w}) = \sum_{n=1}^{N} [y_n = 1] \ln \frac{1}{h\left(\mathbf{x}_n\right)} + [y_n = 0] \ln \frac{1}{1 - h\left(\mathbf{x}_n\right)}$$

   donde $[\cdot]$ vale 1 o 0 según que sea verdad o falso respectivamente la expresión en su interior.

   Respuesta:
   Partiéremos de la función **likelihood** para ir transformándola hasta donde queremos llegar.

$$L(w) = \prod_{i=1}^{N} P\left(y_i|\mathbf{x}_i\right) \quad = \prod_{i=1}^{N} [y_n = 1]\, h\left(\mathbf{x}_n\right) + [y_n = 0]\, 1 - h\left(\mathbf{x}_n\right)$$

Ahora aplico le aplico el menos logaritmo para cambiar el productorio por un sumatorio y ademas se queden la inversa de $h(x)$ y $1 - h(x)$.

$$
\begin{aligned}
&- \ln(\prod_{i=1}^{N} [y_n = 1]\, h\left(\mathbf{x}_n\right) + [y_n = 0]\, 1 - h\left(\mathbf{x}_n\right)) \\
&= \sum_{i=1}^{N} [y_n = 1] - \ln(h\left(\mathbf{x}_n\right)) + [y_n = 0] - \ln(1 - h\left(\mathbf{x}_n\right)) \\
&= \sum_{i=1}^{N} [y_n = 1] \ln(\frac{1}{h\left(\mathbf{x}_n\right)}) + [y_n = 0] \ln(\frac{1}{1 - h\left(\mathbf{x}_n\right)})
\end{aligned}
$$

b) Para el caso $h(x) = \sigma\left(\mathbf{w}^T\mathbf{x}\right)$ mostrar que minimizar el error de la muestra en el apartado anterior es equivalente a minimizar el error muestral.

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right)$$

Respuesta:

$$
\begin{aligned}
E_{\text{in}}(\mathbf{w}) &= \sum_{n=1}^{N} [y_n = 1] \ln \frac{1}{\sigma\left(\mathbf{w}^T\mathbf{x}\right))} + [y_n = 0] \ln \frac{1}{1 - \sigma\left(\mathbf{w}^T\mathbf{x}\right))} \\
&= \sum_{n=1}^{N} [y_n = 1] \ln\left(1 + e^{\mathbf{w}^T \mathbf{x}_n}\right) - \mathbf{w}^T\mathbf{x}_n + [y_n = 0] \ln\left(1 + e^{\mathbf{w}^T \mathbf{x}_n}\right) \\
&= - \sum_{n=1}^{N} [y_n = 1]\, \mathbf{w}^T\mathbf{x}_n + \underbrace{\sum_{n=1}^{N} [y_n = 0] \ln\left(1 + e^{\mathbf{w}^T \mathbf{x}_n}\right)}
\end{aligned}
$$

9. Derivar el error E_{in} para mostrar que en regresión logística se verifica:

$$\nabla E_{\text{in}}(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} = \frac{1}{N} \sum_{n=1}^{N} -y_n \mathbf{x}_n \sigma\left(-y_n \mathbf{w}^T \mathbf{x}_n\right)$$

Argumentar sobre si un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado.

$$
\begin{aligned}
E_{\text{in}}(\mathbf{w}) &= \frac{1}{N} \sum_{i=0}^{N} \ln\left(1 + e^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}\right) \\
\nabla_w E_{\text{in}}(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{N} \sum_{i=0}^{N} \ln\left(1 + e^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}\right) \right) \\
&= -\frac{1}{N} \sum_{i=0}^{N} y_i \mathbf{x}_i \frac{e^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}}{1 + e^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}} \\
&= -\frac{1}{N} \sum_{i=0}^{N} y_i \mathbf{x}_i \frac{e^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i} e^{y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}}{(1 + e^{-y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}) e^{y_i \mathbf{w}^{\mathrm{T}} \mathbf{x}_i}} \\
&== -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}
\end{aligned}
$$

Si esta mas clasificado significa que la etiqueta y el valor predecido tienen distinto valor y por tanto $y_n \boldsymbol{w}^T \boldsymbol{x}_n < 0$. Y como sabemos que cuando un numero esta exponenciado a un numero negativo tenemos $0 < x1$ y por lo tanto el denominador de la fracción sera menor que si los dos tuvieran el mismo signo y el exponente de la e fuera positivo. Al ser menor el denominador, el resultado de la división sera mayor y por lo tanto ese ejemplo contribuiera mas al gradiente.

10. Definamos el error en un punto (x_n, y_n) por

$$\mathbf{e}_n(\mathbf{w}) = \text{máx}\left(0, -y_n \mathbf{w}^T \mathbf{x}_n\right)$$

Argumentar si con esta función de error el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $v = 1$.

Respuesta:
El gradiente descendente actualiza w a partir de los pesos anteriores restando el gradiente por una tasa de aprendizaje. Es decir :

$$w = w_0 + \eta \frac{\partial E_{in}(\mathbf{w})}{\partial w_j} = w_0 + \eta \frac{\partial \,\text{máx}\left(0, -y_n \mathbf{w}^T \mathbf{x}_n\right)}{\partial w^T}$$

Teniendo en cuanta que el enunciado nos dice que $\eta = 1$ obtenemos la siguiente:

$$w = w_0 - y_n x_n$$

Esto es lo mismo que utiliza PLA para actualizar sus valores.

Referencias

[1] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012.

[2] Antoni Parellada (https://stats.stackexchange.com/users/67822/antoni parellada). Hat matrix and leverages in classical multiple regression. Cross Validated. URL:https://stats.stackexchange.com/q/208299 (version: 2018-01-29).