Cuestionario 1
Aprendizaje Automático

Alumno: Christian Vigil Zamora
DNI:
Curso: 3º

27 de agosto de 2019

1. Identificar, para cada una de las siguientes tareas, cual es el problema, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje $(\mathcal{X}, f, \mathcal{Y})$ que deberiamos 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.
b) Decidir si un determinado indice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.
c) Hacer que un dron sea capaz de rodear un obstaculo.
d) Dada una colección de fotos de perros, posiblemente de distintas razas, establecer cuantas razas distintas hay representadas en la colección.

Solución:
a) En éste caso, sería un problema de Clasificación. El tipo de aprendizaje más adecuado considero que es el Aprendizaje Supervisado. Los elementos de aprendizaje serían: $\mathcal{X}$ = datos propios de la carta, tales como Dirección, Localidad... $\mathcal{Y}$ = distrito postal al que pertenece. La función será de la forma $f : \mathcal{X} \mapsto \mathcal{Y}$ es decir, $f(\mathbf{x}_n) = \mathbf{y}_n$.

b) Nuevamente aquí, el tipo de Aprendizaje sería Supervisado puesto que a partir de indices pasados del mercado de valores, podemos aprender de ellos para poder predecir nuevos índices. El problema podría ser de Regresión o clasificación, según lo que se considere como salida, es decir, si la tarea consiste en decir el índice si subirá o bajará a secas, la salida sería una variable discreta y por tanto un problema de Clasificación, mientras que si la tarea consiste en determinar el índice que lo indica, la salida sería Continua, y por tanto un problema de Regresión. Yo lo voy a considerar como un problema con salida Discreta, así que los elementos de aprendizaje serían: $\mathcal{X}$ = índices del mercado de valores pasados. $\mathcal{Y}$ = Sube o Baja. La función será de la forma $f : \mathcal{X} \mapsto \mathcal{Y}$ es decir, cada índice del mercado de valores pasado tiene asociado la etiqueta de si subió o bajó.

c) Nos encontramos ante un problema de Decisión, en el que el aprendizaje más adecuado sería el Aprendizaje por Refuerzo. Sería el más adecuado puesto que no tenemos ni $\mathcal{X}$ ni $\mathcal{Y}$ ni f, sino que se sigue un Proceso de Decisión de Markov en el que el Dron percibe un conjunto finito de estados en su entorno, para que el tiene un conjunto posible de acciones a realizar. Conforme el Dron va tomando decisiones, se le recompensa en caso de realizar la acción correcta. De ésta forma, conseguimos que el Dron con la experiencia se esfuerce por tomas aquellas decisiones que le llevan a la recompensa. Entendiendo por recompensa una variable númerica positiva.

d) Para éste el apartado, el problema sería de Clustering pues lo más lógico es

agrupar los datos, en éste caso, fotografías, según la similitud que presenten y de ahí poder deducir las distintas razas. Claramente estamos ante un tipo de Aprendizaje No Supervisado, pues no tenemos etiqueta alguna de los datos. Los elementos de aprendizaje serían: $\mathcal{X}$ = Fotografías de las que se obtener características de los perros, tales cómo color de pelo, manchas, orejas... $\mathcal{Y}$ = No hay en éste tipo de Aprendizaje. En cuanto a la función, dado que no tenemos etiquetas, nuestra función sería $f : \mathcal{X} \mapsto \mathcal{R}_i$, siendo $\mathcal{R}_n$ el número total de razas.

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

a) Determinar si un vertebrado es mamifero, reptil, ave, anfibio o pez.
b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.
c) Determinar perfiles de consumidor en una cadena de supermercados.
d) Determinar el estado anímico de una persona a partir de una foto de su cara.
e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.

Solución:
a) Éste apartado sería adecuado para una aproximación por aprendizaje, puesto que podríamos tomar como datos de entrenamiento las características propias de los vertebrados, y como etiquetas el grupo a que pertenecen; mamíferos, reptiles, aves, anfibios o peces. A partir de ahí, se podría llevar a cabo el aprendizaje para poder predecir.

b) Éste, se considera más adecuado para una aproximación por diseño puesto que se recopila información acerca de los periodos en los que más aparece dicha enfermad y su duración. Se construye un modelo físico para el periodo y la duración, y se prueban diferentes variaciones y se van almacenando los errores obtenidos. Finalmente, se construye una distribución de probabilidad en función del periodo y su duración, la cual se usa para determinar si se debe aplicar una campaña o no.

c) Nuevamente éste apartado es más adecuado para una aproximación por aprendizaje, ya que los elementos del aprendizaje se deducen fácilmente. A partir de datos de consumidores tales como frecuencia de compra, dinero gastado, productos adquiridos... se podrían determinar mediante Clustering, perfiles de consumidor.

d) Determinar el estado anímico de una persona a partir de una foto de su cara sería mas adecuando para una aproximación por aprendizaje, concretamente Aprendizaje No Supervisado, ya que el aprendizaje puede llevarse a cabo a partir de fotografías, en las que se van identificando patrones en la cara según su estado,

y así poder determinar en un futuro cuando no se conozca el estado.

e) Éste caso lo considero mejor para una aproximación por aprendizaje, puesto que es fácil reconocer patrones asociados a la cantidad de tráfico que pasa por ese cruce y a la cantidad de tiempo que debe pasar ese tráfico en el cruce por las luces de los semáforos. La recopilación de datos podría llevarse a cabo a partir de cámaras instaladas en el cruce, e ir entrenando con esos datos, para poder llegar a determinar por ejemplo que si en el cruce hay mucha densidad de tráfico, lo lógico es que el semáforo esté mas tiempo con luz verde, para disminuir esa densidad.

3. Construir un problema de aprendizaje desde datos para un problema de clasificación de fruta en una explotación agraria que produce mangos y guayabas. Identificar los siguientes elementos formales $\mathcal{X}, \mathcal{Y}, \mathcal{D}, f$ 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.

Solución: En primer lugar, se buscan características que definen a los ambos tipos de frutas, por lo tanto nuestra $\mathcal{X}$ podría ser: Color, Calorías, Fibra, Grasa, Azúcares... La $\mathcal{Y}$ se corresponde con las etiquetas, en éste caso tenemos Mango o Guayaba. El dataset $\mathcal{D}$ sigue la distribución $\mathcal{D} = \{(\mathcal{X}i, \mathcal{Y}i), i = 1, ..., N\}$, siendo un ejemplo éste: (Color: 0.2,X1:Calorías=0.3,Fibra=5.0,Grasa=0.7,Azúcares=0.2, Y1=Mango). La función será de la forma $f : \mathcal{X} \mapsto \mathcal{Y}$, correspondiendo una etiqueta para cada conjunto de características. En cuanto a la descripción de los mismos para poder ser usada por un computador, las características de la $\mathcal{X}$ tomarían valores flotantes y el Color se representaría mediante el intervalo $[0, 1]$ pudiendo tomar valores flotantes intermedios y siendo 0.0 el color rojo más puro y 1.0 el color verde más puro. La $\mathcal{Y}$ tomaría los valores $\{0, 1\}$,, siendo Mango si es 0 y Guayaba si es 1. El dataset $\mathcal{D}$ sería una matriz en la que cada fila sería un conjunto de características de $\mathcal{X}$ y a su vez, $\mathcal{Y}$ sería un vector con las etiquetas correspondientes a cada fila de la matriz del dataset. En éste problema estamos ante un caso de etiquetas con ruido puesto que es muy probable que haya clasificaciones erróneas, es decir, las características con las que estamos clasificando no son lo suficientemente discriminatorias, ya que puede haber tanto mangos como guayabas que posean características muy parecidas entre sí y por tanto sean mal clasificados.

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

Solución: Puesto que la matriz cuadrada A debe admitir la descomposición para una matriz X de números reales, asumimos que la matriz X se trata también de una matriz cuadrada, y por tanto ambas matrices son simétricas. Si descomponemos en valores singulares la matriz X nos queda que $\mathrm{X} = \mathrm{UDV}^{\mathrm{T}}$, siendo D una matriz diagonal. De esa descomposición podemos obtener que $\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{VDDV}^{\mathrm{T}}$. Llegados a éste punto y recordando lo comentado al inicio, de que tanto la matriz A como X son simétricas, podemos admitir la descomposición $\mathrm{A} = \mathrm{X}^{\mathrm{T}}\mathrm{X}$, obteniendo la relación de que los valores singulares de la matriz A son los mismos que los de la matriz X al cuadrado, y eso es debido a dicha descomposición en valores singulares: $\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{VDDV}^{\mathrm{T}}$ en la que nos queda DD, siendo D la matriz diagonal que contiene los valores singulares en su diagonal.

5. Sean x e y dos vectores de características de dimensión $M \times 1$. La expresión

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} (x_i - \overline{x})(y_i - \overline{y})$$

define la covarianza entre dichos vectores, donde $\overline{z}$ representa el valor medio de los elementos de z. Considere ahora una matriz X cuyas columnas representan vectores de caracteristicas. La matriz de covarianzas asociada a la matriz $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

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

Sea $1_M^T = (1, 1, \cdots, 1)$ un vector $M \times 1$ de unos. Mostrar que representan las siguientes expresiones
a) $E1 = 11^T\mathrm{X}$
b) $E2 = \left(\mathrm{X} - \frac{1}{M}E1\right)^T \left(\mathrm{X} - \frac{1}{M}E1\right)$

Solución:
a) Para ejemplificar lo que representan las siguiente expresiones, voy a considerar M = 3.

$$1 = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$$

$$1^T = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Luego 1 * $1^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ y considero X = $\begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}$

Por tanto, la expresión E1 = $\begin{pmatrix} 6 & 15 & 24 \\ 6 & 15 & 24 \\ 6 & 15 & 24 \end{pmatrix}$

De ahí, concluimos que la expresión E1 se corresponde con la sumatoria de cada columna de la matriz X, es decir, el valor de las columnas de E1 viene a ser la sumatoria de esa misma columna en la matriz X, siendo cada columna un vector de características.

b) Considero los siguientes valores:

$\frac{1}{M}E1 = \begin{pmatrix} 2 & 5 & 8 \\ 2 & 5 & 8 \\ 2 & 5 & 8 \end{pmatrix}$

$\left(\mathrm{X} - \frac{1}{M}E1\right)^T = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}$

$\left(\mathrm{X} - \frac{1}{M}E1\right) = \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$.

Por tanto, una vez que tenemos calculadas ambas expresiones, obtenemos que $E2 = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$. Aparentemente podría parecer que dicha expresión no significa nada, en cambio para nada es así. Si calculamos la matriz de covarianzas asociada a la matriz X, obtenemos que cov(X) = $\begin{pmatrix} 0,\widehat{6} & 0,\widehat{6} & 0,\widehat{6} \\ 0,\widehat{6} & 0,\widehat{6} & 0,\widehat{6} \\ 0,\widehat{6} & 0,\widehat{6} & 0,\widehat{6} \end{pmatrix}$ y si recordamos que habíamos considerado M = 3, podemos llevar a cabo la operación M * cov(X) = $\begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$, obteniendo como resultado la expresión E2, por lo tanto podemos concluir que la expresión E2 representa M * cov(X), siendo M la longitud de los vectores de características y cov(X), la matriz de covarianzas asociada a la matriz $\mathcal{X}$.

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 $N \times (d+1)$, y$\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es invertible.
a) ¿Que representa la matriz $\hat{\mathrm{H}}$ en un modelo de regresión?
b) Identifique la propiedad más relevante de dicha matriz en relación con regresión lineal.

Solución:
a) En un modelo de regresión, la matriz $\hat{\mathrm{H}}$ representa el peso que tienen las etiquetas de aprendizaje para un conjunto de datos. Concretamente, la matriz $\hat{\mathrm{H}}$ proyecta el vector de etiquetas $\mathcal{Y}$ sobre el vector de etiquetas de aprendizaje $\hat{\mathcal{Y}}$.

b) La propiedad más relevante de dicha matriz en relación con regresión lineal sería la Idempotencia, ésto quiere decir que $\mathrm{H}^2 = \mathrm{H}$. Ésto se puede comprobar de la siguiente forma: $(\mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}})(\mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}) = \mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}} = \mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}$. Es la más relevante puesto que si se realizan dos observaciones iguales con la matriz $\hat{\mathrm{H}}$, la predicción sobre las etiquetas debe ser la misma.

7. La regla de adaptación de los pesos del Perceptron ($\mathbf{w}_{\text{ new}} = \mathbf{w}_{\text{old}} + y\mathbf{x}$) tiene la interesante propiedad de que mueve el vector de pesos en la dirección adecuada para clasificar x de forma correcta. Suponga el vector de pesos 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 $\mathbf{x}(t)$.

Solución:
Para probar matemáticamente lo pedido, voy a considerar los siguientes datos iniciales: $\mathbf{x} = 0{,}5$ $\mathbf{w} = -1{,}2$ e $\mathbf{y} = 1$, siendo ésta la etiqueta bien clasificada de x. El Perceptrón viene dado por la función $\mathbf{h}(x) = sign(\mathbf{w}^T\mathbf{x})$, siendo $sign(x) = \begin{cases} 1 & \text{si} \quad x > 0 \\ -1 & \text{si} \quad x < 0 \end{cases}$ además $\mathbf{h}(x)$ toma los valores $\{-1, 1\}$, . Como asumimos que x está mal clasificado, el Perceptrón va a iterar hasta clasificar bien x. En la primera iteración, $\mathbf{h}(x) = sign(-\mathbf{1}{,}\mathbf{2}^T\mathbf{0}{,}\mathbf{5}) = -0{,}6$, es decir, $\mathbf{h}(x) = -1$ distinto de $\mathbf{y}$, por lo que la predicción sigue siendo errónea, así que se actualiza w de la forma: $\mathbf{W}_{new} = \mathbf{w} + \mathbf{yx}$; $\mathbf{W}_{new} = -1{,}2 + 0{,}5 = -0{,}7$. En la segunda iteración, se realiza el mismo procedimiento, obteniendo: $\mathbf{h}(x) = sign(-\mathbf{0}{,}\mathbf{7}^T\mathbf{0}{,}\mathbf{5}) = -0{,}35$, siendo $\mathbf{h}(x) = -1$, distinto de $\mathbf{y}$, por lo que la predicción sigue siendo errónea, así que se actualiza w: $\mathbf{W}_{new} = -0{,}7 + 0{,}5 = -0{,}2$. En la tercera iteración, se realiza el mismo procedimiento, obteniendo: $\mathbf{h}(x) = sign(-\mathbf{0}{,}\mathbf{2}^T\mathbf{0}{,}\mathbf{5}) = -0{,}1$, siendo $\mathbf{h}(x) = -1$, distinto de $\mathbf{y}$, por lo que la predicción sigue siendo errónea, así que se actualiza w: $\mathbf{W}_{new} = -0{,}2 + 0{,}5 = 0{,}3$. En la cuarta iteración, se realiza el

mismo procedimiento, obteniendo: $\mathbf{h}(x) = sign(\mathbf{0},\mathbf{3}^T\mathbf{0},\mathbf{5}) = 0{,}15$, siendo $\mathbf{h}(x) = 1$, concluyendo en la cuarta iteración con una predicción correcta tras mover w en la dirección correcta.

8. Sea un problema probabilistico de clasificación binaria con etiquetas $\{0, 1\}$, es decir $\mathrm{P(Y} = 1) = h(\mathrm{x})$ y $\mathrm{P(Y} = 0) = 1 - h(\mathrm{x})$, para una función $h($ 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 función h que maximiza la verosimilitud de la muestra es la misma que minimiza

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

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

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

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

Solución:

a) Conocemos que la función h es $\mathbf{h}(x) = (\mathbf{w}^T\mathbf{x})$ y con la finalidad de abreviar $\mathrm{P(Y} = 1) = h(\mathrm{x})$ y $\mathrm{P(Y} = 0) = 1 - h(\mathrm{x})$, consideramos que: $P(y|x) = \begin{cases} h(x) & \text{si} \quad Y = 1 \\ 1 - h(x) & \text{si} \quad Y = 0 \end{cases}$. Además, conociendo h(x) podemos concluir que $P(y|x) = (\mathbf{y}\mathbf{w}^T\mathbf{x})$. Una vez sabemos ésto, obtenemos la verosimilitud de la muestra que viene dada por la expresión $L(\mathbf{w}) = \prod_{n=1}^{N} P\left(y_i|\mathbf{x}_i\right) = \prod_{n=1}^{N} \sigma\left(y_n \mathbf{w}^{\mathrm{T}} \mathbf{x}_n\right)$, de la que h forma parte. Ahora vamos a mostrar que dicha función h es la misma que minimiza el error en la muestra, ya que a partir de la expresión que calcula la verosimilitud de la muestra, podemos obtener el error en la muestra: $E_{\text{in}}(\mathbf{w}) = -\frac{1}{N}\ln(\mathrm{L}(\mathbf{w})) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{P(y_n|\mathbf{x}_n)}\right)$, teniendo en cuenta la definición de $P(y|x)$ dada arriba, en la que se encuentra h.

b) En éste apartado partimos con que $h(x) = \sigma\left(\mathbf{w}^T\mathbf{x}\right)$, que a su vez podemos expresar como $h(x) = \frac{e^{w^T x}}{1+e^{w^T x}}$. Ésta última expresión es la que vamos

a usar para minimizar el error usando la fórmula del apartado anterior. Para mostrar la equivalencia entre ambas, voy tomar como datos $\mathbf{x} = 2$, $\mathbf{w} = 0{,}3$ e $\mathbf{y} = 1$. Primero voy a calcular el error mediante la fórmula del apartado anterior: $h(x) = \frac{e^{0,3}2}{1+e^{0,3}2} = 0{,}65$. Puesto que la etiqueta es 1, me quedo sólo con ésta parte de la fórmula: $E_{\text{in}}(\mathbf{w}) = \sum_{n=1}^{N} [y_n = 1] \ln \frac{1}{h(\mathbf{x}_n)}$ ya que la segunda parte se anula por valer 0 y como ya tengo calculada $h(x)$ obtengo que $E_{\text{in}}(\mathbf{w}) = 0{,}437$ en la fórmula del apartado anterior. Ahora procedo a minimizar el error con la fórmula de éste apartado, la cuál puedo calcular directamente considerando los datos iniciales anteriores. $E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-0,6}) = 0{,}437$

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.

Solución: Partimos de la fórmula inicial del error:
$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right)$. A continuación, derivamos la expresión anterior, obteniendo que:
$\nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + e^{-y_n \mathbf{w}^{\mathrm{T}} \mathbf{x}_n}\right)\right) = \frac{1}{N} \sum_{n=1}^{N} -y_n \mathbf{x}_n \frac{e^{-y_n \mathbf{w}^{\mathrm{T}} \mathbf{x}_n}}{1+e^{-y_n \mathbf{w}^{\mathrm{T}} \mathbf{x}_n}}$.
Una vez que tenemos la fórmula del error derivada, observando el resultado nos damos cuenta de que podemos realizar la equivalencia de la función logística, propia de una Regresión logística: $\frac{e^{-\mathbf{y_n} \mathbf{w}^{\mathrm{T}} \mathbf{x}_n}}{1+e^{-y_n \mathbf{w}^{\mathrm{T}} \mathbf{x}_n}} = \sigma\left(-\mathbf{y_n} \mathbf{w}^{\mathrm{T}} \mathbf{x_n}\right))$, obteniendo por tanto $\nabla E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} -y_n \mathbf{x}_n \sigma\left(-y_n \mathbf{w}^{\mathrm{T}} \mathbf{x}_n\right)$ y verificando la expresión del enunciado.
Efectivamente, un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado. Para demostrarlo, voy a utilizar la expresión $\nabla E_{\text{in}}(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1+e^{y_n \mathbf{w}^T \mathbf{x}_n}}$ quedándome sólo con la parte de la división. Antes de continuar, aclarar que los valores posibles de las etiquetas son $\{-1, 1\}$. Si tenemos un ejemplo mal clasificado, el exponente de la exponencial del denominador será negativo, pues si el ejemplo está mal clasificado la etiqueta mal clasificada es la contraria a la buena, por tanto si la etiqueta bien clasificada es 1, mal clasificada sería -1 y puesto que en el exponente de la exponencial se multiplica tanto la etiqueta bien clasificada como la mal clasificada, el exponente de la exponencial será negativo, lo que da lugar a valores más pequeños. El problema viene a la hora de realizar la división, ya que el hecho de que el denominador sea pequeño en una división hace que el resultado de ella sea mayor. En cambio, cuando un ejemplo está bien clasificado tanto su etiqueta como la predicción coinciden, de forma que el exponente de la exponencial nunca

es negativo, y por tanto da lugar a valores más grandes que cuando se efectúa la división, el resultado se hace más pequeño. De esa forma, el resultado de un ejemplo mal clasificado será mayor que el de un ejemplo bien clasificado, por tanto, un ejemplo mal clasificado contribuye más al gradiente.

10. Definamos el error en un punto $(\mathbf{x}_n, y_n)$ 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 $\mathbf{e}_n$ con tasa de aprendizaje $\nu = 1$

Solución: La respuesta es que sí, su argumento el siguiente: Sabemos que el algoritmo PLA cuando produce una clasificación correcta, $\mathbf{w}$ se queda tal cual está, mientras que cuando produce una predicción incorrecta, mueve $\mathbf{w}$ hacia la clasificación correcta así: $\mathbf{w}(t+1) = \mathbf{w}(t) + \mathbf{y}_i \mathbf{x}_i$. Ahora pasamos al SGD, en el que de forma parecida al PLA mueve la posición de w, en cada punto, sin influir los demás puntos. En SGD, $\mathbf{w}$ se mueve así: $\mathbf{w}_j := \mathbf{w}_j - \eta \frac{\partial E_{in}(\mathbf{w})}{\partial w_j}$, siendo el cociente la expresión del Gradiente. Como el ejercicio nos propone usar la función de error $\mathbf{e}_n(\mathbf{w})$, para el SGD habrá que usar su derivada y sustituirla por la expresión del Gradiente, por lo que la función $\mathbf{e}_n(\mathbf{w})$ derivada nos queda $\mathbf{e}_n(\mathbf{w}) = \text{máx}\,(0, -y_n \mathbf{x}_n)$ y completando la sustitución obtenemos que $\mathbf{w}_j := \mathbf{w}_j - \eta(-y_n \mathbf{x}_n)$. Esa expresión puede ser simplificada puesto que conocemos que la tasa de aprendizaje es 1, de forma que: $\mathbf{w}_j := \mathbf{w}_j + \mathbf{y}_n \mathbf{x}_n$, obteniendo la misma expresión que con el algoritmo PLA. Llegados aquí queda argumentado que el algoritmo PLA puede interpretarse como SGD sobre $\mathbf{e}_n(\mathbf{w})$ ya que cuando la clasificación sea correcta, la $\mathbf{w}$ se queda tal cual está y aplicándolo a SGD, cuando la clasificación sea correcta, el valor de la derivada será 0 por lo que $\mathbf{w}$ se queda igualmente como estaba, mientras que cuando la clasificación sea incorrecta, tal y como he demostrado anteriormente, la fórmula para actualizar el movimiento de $\mathbf{w}$ acaba siendo la misma tanto en PLA como en SGD sobre $\mathbf{e}_n(\mathbf{w})$.