Aprendizaje Automático
Cuestionario de teoría 1

Alfonso García Martínez
alfonsogmw@correo.ugr.es

Abril de 2019

1. Identificar, para cada una de las siguientes tareas, cuál es el problema, qué tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje $(\mathcal{X}, f, \mathcal{Y})$ que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar cómo y describir los elementos para cada tipo.

   a) Clasificación automática de cartas por distrito postal.

   Solución: Suponiendo que no disponemos de los distritos postales de las cartas que queremos clasificar (si no, obviamente no sería necesario usar técnicas de aprendizaje), estaríamos ante un caso de aprendizaje supervisado. Para el entrenamiento, se usarían cartas cuyos códigos postales (las etiquetas, elementos de $\mathcal{Y}$) sí conocemos, y nuestro clasificador debería ser capaz de aprender a predecirlos en función de otras características de la carta, presumiblemente de la localidad y la dirección, que son las características que conformarían $\mathcal{X}$. Y como en todo problema de aprendizaje supervisado, en principio no conocemos la verdadera relación f entre las características de la carta y el verdadero código postal que le corresponde.

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

   Solución: Puede enfocarse como un problema de aprendizaje supervisado: en base a datos como los valores actuales de las acciones y midiendo, por ejemplo, la repercusión en los medios de comunicación y en las redes sociales de las compañías que aglutina el índice, podría predecirse su valor en ese periodo fijado de tiempo.

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

   Solución: Aquí, el dron debe ser capaz de saber en todo momento cuál es el mejor movimiento posible dependiendo de en qué posición se sitúe y/o a qué distancia del obstáculo se encuentra. No obstante, no existe un único movimiento *correcto* (o por lo menos no es una tarea trivial definirlo), sino que existen unos movimientos o direcciones a seguir mejores que otros, con distintos grado de bondad en función de cuánto ayuden al dron a avanzar sin que choque con el obstáculo. Para poder entrenar al dron, habría que

ponerlo en marcha sucesivas veces para que intente hacer su recorrido, ver cuánto éxito ha tenido cada vez intentando esquivar el obstáculo y reforzar aquellas situaciones en las que lo consigue. Por ello, diría que el aprendizaje por refuerzo es el adecuado.

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

Solución: En este caso no se nos pide clasificar las fotos de perros según la raza, sino encontrar posibles similitudes entre los perros que aparecen en las fotos atendiento a ciertas características (tamaño, forma de la cabeza, forma de las orejas, tipo y color de pelaje, etc.) y agruparlos en distintas razas. Como a priori no conocemos las posibles razas que conforman $\mathcal{Y}$, sino que nuestra tarea es encontrarlas, estamos ante un buen ejemplo para el uso de aprendizaje no supervisado.

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.

      Solución: Es más fácil clasificar vertebrados a partir de ejemplos (fotografías, tablas de características de una base de datos, etc.) que describir formalmente cada uno de los tipos de vertebrados existentes y a partir de ahí determinar la clase a la que pertenece un ejemplo según esa misma descripción. Por ese motivo esta tarea se aborda mejor con una aproximación por aprendizaje.

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

      Solución: Es posible el modelaje matemático (aproximación por diseño) de epidemias, lo que en teoría nos permitiría saber cuándo es el mejor momento para iniciar una campaña de vacunación.

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

      Solución: A partir de los recibos de los consumidores, o mejor aún, haciéndolos socios de la cadena de supermercados para que faciliten más información, se podrían buscar similitudes entre consumidores para definir agrupaciones según el perfil al que pertenezcan. Justamente esto supone un enfoque de aprendizaje no supervisado (*clustering*). Dicho esto, intentar lo mismo con una aproximación por diseño no parece ser la mejor opción.

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

      Solución: No conocemos fórmulas que describan el estado de ánimo de una persona según parámetros que describan su cara. En cambio, el problema es factible si utilizamos ejemplos con fotos de distintas caras para las cuales conocemos su estado de ánimo, y a partir de ellas intentásemos adivinar el estado de ánimo de otras

caras distintas. En definitiva, es mejor un enfoque por aprendizaje.

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

Solución: Existe toda una rama dentro del campo de la ingeniería civil llamada ingeniería de tráfico, en la cual hay modelos matemáticos para calcular los ciclos óptimos de los semáforos y así poder establecer el mejor plan semafórico para una intersección. En este caso, como podemos encontrar la solución analíticamente ajustando parámetros, el problema es adecuado para un enfoque por diseño.

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 $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{D}$, 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: Para este problema de clasificación, podemos distinguir los siguientes elementos:

   - $\mathcal{X}$: el conjunto de características de las frutas que tendremos en cuenta al clasificar: tamaño, forma, color, textura, sabor, etc.

     Aquellas características que sean variables numéricas pueden codificarse sin problema en el ordenador como tipos de datos reales/flotantes.

     Sin embargo, para las variables categóricas la cosa no será tan fácil: si dicha variable categórica puede valer n valores distintos, y directamente asignamos un natural desde 1 hasta n a cada posible valor categórico, estaremos estableciendo sin quererlo un orden en los valores de esa variable, pues estamos trasladando el orden presente en los números naturales a una variable categórica cuyos valores no tienen por qué tenerlo.

     Para estos casos, es mejor codificar las variables categóricas con vectores o *arrays* de dígitos binarios, de forma que si la variable categórica tiene n valores distintos y una instancia tiene el valor i-ésimo, se codificará con un vector de n binarios, de los cuales todos valdrían 0 menos el i-ésimo, que vale 1.

     Por ejemplo: si la variable categórica *color* está definida en el conjunto $\{verde, amarillo, naranja\}$, y para una instancia $color = amarillo$ (segundo valor), lo codificamos con el vector $(0, 1, 0)$.

   - $\mathcal{Y}$: las posibles frutas (etiquetas) en las que se un elemento puede clasificar: *mango*, *papaya* o *guayaba*. Podemos codificarlos con números enteros, de forma que $\mathcal{Y} = \{1, 2, 3\}$

   - $\mathcal{D}$: la muestra de frutas que recogemos y cuyas características medimos y codificamos para conformar el *dataset* con el que entrenaremos a nuestro clasificador. Debe ser recolectada aleatoriamente de distintos lugares repartidos por toda la explotación agraria, y no en una sola zona, para evitar sesgos en la muestra que más

adelante induzcan a error a nuestro clasificador.

- f: la verdadera relación existente entre las características de las frutas y la clase de fruta a la que pertenezcan, $f : \mathcal{X} \to \mathcal{Y}$.

  Alguien que no haya visto nunca estas frutas, o que no sepa distinguirlas, o que en general no esté familiarizado con el dominio de este problema no sepa cómo las caracterísiticas que tenemos en cuenta determinan qué clase de fruta tiene entre manos. Podríamos intentar conocer mejor esta función objetivo consultando a un agricultor experto en estas tres frutas, no solo para que nos ayude a entrenar el clasificador sino para que nos dé alguna indicación sobre qué hay que tener más en cuenta al identificar una fruta.

En este problema puede darse el caso de que aparezca ruido, o bien por fallos o errores de precisión al medir las características de la fruta, o bien porque en la fase de entrenamiento no hemos consultado al experto y por desconocimiento hemos introducido ejemplos mal clasificados (es perfectamente posible confundir dos frutos muy parecidos que en realidad no son los mismos).

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.

   Solución: Dada una matriz X de dimensiones $N \times d$ con rango ρ, su descomposición en valores singulares (SGV) es la siguiente:

   $$X = UDV^T$$

   Donde las matrices $U \in \mathbb{R}^{N \times \rho}$ y $V \in \mathbb{R}^{d \times \rho}$ son ortogonales:

   $$U^T U = U^{-1} U = Id$$
   $$V^T V = V^{-1} V = Id$$

   y la matriz $D \in \mathbb{R}^{\rho \times \rho}$ es diagonal y positiva, y contiene en su diagonal los valores singulares de X.

   Gracias a la propiedad de la traspuesta de un producto de matrices, según la cual:

   $$(AB)^T = B^T A^T$$

   Sabemos que

   $$X^T = (UDV^T)^T = (V(UD)^T) = VD^T U^T$$

   Por lo que, descomponiendo A:

   $$A = X^T X = VD^T U^T UDV^T = VD^T DV^T$$

   En esta descomposición de A, que además coincide con su diagonalización (recordemos que V es ortogonal, luego $V^T = V^{-1}$), la matriz $D^T D$ contiene en su diagonal los elementos de la diagonal de D al cuadrado, puesto que D es diagonal y cuadrada.

   Por tanto, los valores singulares de A (que además son sus valores propios) equivalen a los correspondientes valores singulares de X al cuadrado.

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

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

define la covarianza entre dichos vectores, donde $\bar{z}$ 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 $\text{X} = (x_1, x_2, \ldots, x_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

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

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

a) $E1 = 11^T X$

Solución:

$$E1 = 11^T X = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{N1} \\ x_{12} & x_{22} & \ldots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \ldots & x_{NM} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{N1} \\ x_{12} & x_{22} & \ldots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \ldots & x_{NM} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & \ldots & \sum_{i=1}^{M} x_{Ni} \\ \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & \ldots & \sum_{i=1}^{M} x_{Ni} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & \ldots & \sum_{i=1}^{M} x_{Ni} \end{pmatrix}$$

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

Solución:

$X - \frac{1}{M}E1 =$

$$= \begin{pmatrix} x_{11} & x_{21} & \dots & x_{N1} \\ x_{12} & x_{22} & \dots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \dots & x_{NM} \end{pmatrix} - \begin{pmatrix} \frac{1}{M}\sum_{i=1}^{M} x_{1i} & \frac{1}{M}\sum_{i=1}^{M} x_{2i} & \dots & \frac{1}{M}\sum_{i=1}^{M} x_{Ni} \\ \frac{1}{M}\sum_{i=1}^{M} x_{1i} & \frac{1}{M}\sum_{i=1}^{M} x_{2i} & \dots & \frac{1}{M}\sum_{i=1}^{M} x_{Ni} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{M}\sum_{i=1}^{M} x_{1i} & \frac{1}{M}\sum_{i=1}^{M} x_{2i} & \dots & \frac{1}{M}\sum_{i=1}^{M} x_{Ni} \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11} & x_{21} & \dots & x_{N1} \\ x_{12} & x_{22} & \dots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \dots & x_{NM} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_N \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_N \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \dots & x_{N1} - \bar{x}_N \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{N2} - \bar{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} - \bar{x}_1 & x_{2M} - \bar{x}_2 & \dots & x_{NM} - \bar{x}_N \end{pmatrix}$$

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

$$= \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \dots & x_{N1} - \bar{x}_N \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{N2} - \bar{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} - \bar{x}_1 & x_{2M} - \bar{x}_2 & \dots & x_{NM} - \bar{x}_N \end{pmatrix}^T \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \dots & x_{N1} - \bar{x}_N \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{N2} - \bar{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} - \bar{x}_1 & x_{2M} - \bar{x}_2 & \dots & x_{NM} - \bar{x}_N \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_1 & \dots & x_{1M} - \bar{x}_1 \\ x_{21} - \bar{x}_2 & x_{22} - \bar{x}_2 & \dots & x_{2M} - \bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} - \bar{x}_N & x_{N2} - \bar{x}_N & \dots & x_{NM} - \bar{x}_N \end{pmatrix} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \dots & x_{N1} - \bar{x}_N \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{N2} - \bar{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} - \bar{x}_1 & x_{2M} - \bar{x}_2 & \dots & x_{NM} - \bar{x}_N \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^{M}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1) & \sum_{i=1}^{M}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) & \cdots \sum_{i=1}^{M}(x_{1i} - \bar{x}_1)(x_{Ni} - \bar{x}_N) \\ \sum_{i=1}^{M}(x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1) & \sum_{i=1}^{M}(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2) & \cdots \sum_{i=1}^{M}(x_{2i} - \bar{x}_2)(x_{Ni} - \bar{x}_N) \\ \vdots & \vdots & \ddots \quad \vdots \\ \sum_{i=1}^{M}(x_{Ni} - \bar{x}_N)(x_{1i} - \bar{x}_1) & \sum_{i=1}^{M}(x_{Ni} - \bar{x}_N)(x_{2i} - \bar{x}_2) & \cdots \sum_{i=1}^{M}(x_{Ni} - \bar{x}_N)(x_{Ni} - \bar{x}_N) \end{pmatrix} =$$

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

$$= M\text{cov}(\text{X})$$

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

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

   Solución: En el contexto de la regresión lineal, cuando igualamos a cero la derivada (gradiente) del error dentro de la muestra usando notación matricial, obtenemos que:

$$\nabla E_{in}(\mathbf{w}) = \frac{2}{N}(X^TX\mathbf{w} - X^Ty) = 0$$
$$\rightarrow \; X^TX\mathbf{w} - X^Ty = 0$$

   Luego

$$X^TX\mathbf{w} = X^Ty$$

   Y si X^TX es invertible, finalmente

$$\mathbf{w} = (X^TX)^{-1}X^Ty = X^\dagger y$$

   Con la $\mathbf{w}$ obtenida, es posible estimar las etiquetas, pues ahora tenemos el vector de pesos que minimiza E_{in}:

$$\hat{y} = X\mathbf{w} = X(X^TX)^{-1}X^Ty = \hat{H}y$$

   Podemos decir que $\hat{H}$ define una aplicación lineal que transforma la verdadera etiqueta y en nuestra estimación $\hat{y}$, es decir, transforma los datos observados en estimaciones utilizando únicamente la matriz X.

   b) Identifique la propiedad más relevante de dicha matriz en relación con regresión lineal. Justificar las respuestas.

   Solución: $\hat{H}$ tiene la particularidad de que es idempotente, es decir, $\hat{H}^2 = \hat{H}$

$$\hat{H}^2 = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) =$$
$$= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T =$$
$$= X(X^TX)^{-1}X^T = \hat{H}$$

También es fácil ver que $\hat{H}^k = \hat{H}$ para cualquier entero positivo k

$$\hat{H}^k = \hat{H}^{k-2}\hat{H}^2 = \hat{H}^{k-2}\hat{H} =$$
$$= \hat{H}^{k-1} = ... = \hat{H}$$

La idempotencia de $\hat{H}$ es muy relevante, porque implica que la transformación puede aplicarse numerosas veces sin que afecte al resultado:

$$\hat{H}\hat{H}\hat{H}\ldots\hat{H}y = \hat{H}y = \hat{y}$$

Con lo cual **si se intenta predecir la etiqueta numerosas veces, siempre tendremos el mismo resultado**.

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 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 w en la dirección correcta para clasificar bien $\mathbf{x}(t)$.

   Solución: La regla de adaptación del PLA, en términos de t, es la siguiente:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y(t)\mathbf{x}(t)$$

   Veamos la casuística para el caso en que $x(t)$ esté mal clasificado:

   - Si $y(t) = 1$ pero la estimación $\text{sign}(\mathbf{w}(t)^T\mathbf{x}(t)) = -1$, entonces a $\mathbf{w}(t)$ estaremos **sumándole** $\mathbf{x}(t)$:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \mathbf{x}(t)$$

     con lo cual, al hacer la nueva predicción

$$\begin{aligned}\text{sign}(\mathbf{w}(t+1)^T\mathbf{x}(t)) &= \text{sign}((\mathbf{w}(t) + \mathbf{x}(t))^T\mathbf{x}(t)) = \\ &= \text{sign}((\mathbf{w}(t)^T + \mathbf{x}(t)^T)\mathbf{x}(t)) = \\ &= \text{sign}(\mathbf{w}(t)^T\mathbf{x}(t) + \mathbf{x}(t)^T\mathbf{x}(t))\end{aligned}$$

     Como a $\mathbf{w}(t)^T\mathbf{x}(t)$, que en este caso tiene signo negativo, le estamos sumando $\mathbf{x}(t)^T\mathbf{x}(t) = ||\mathbf{x}(t)||^2$, que sabemos que será siempre positivo, luego estaremos contribuyendo a que $\mathbf{w}(t)^T\mathbf{x}(t)$ sea *menos negativo*, o sea, que lo desplazamos en sentido positivo para que se acerque más a 0 o que incluso lo sobrepase y pase a ser positivo (en cuyo caso $\mathbf{x}(t)$ estaría ahora incluso bien clasificado).

   - Si $y(t) = -1$ pero la estimación $\text{sign}(\mathbf{w}(t)^T\mathbf{x}(t)) = 1$, , entonces a $\mathbf{w}(t)$ estaremos **restándole** $\mathbf{x}(t)$. Siguiendo un razonamiento similar al anterior:

$$\begin{aligned}\text{sign}(\mathbf{w}(t+1)^T\mathbf{x}(t)) &= \text{sign}((\mathbf{w}(t) - \mathbf{x}(t))^T\mathbf{x}(t)) = \\ &= \text{sign}((\mathbf{w}(t)^T - \mathbf{x}(t)^T)\mathbf{x}(t)) =\end{aligned}$$

$$= \text{sign}(\mathbf{w}(t)^T\mathbf{x}(t) - \mathbf{x}(t)^T\mathbf{x}(t))$$

Como a $\mathbf{w}(t)^T\mathbf{x}(t)$, que ahora tiene signo positivo, le estamos restando $\mathbf{x}(t)^T\mathbf{x}(t) = ||\mathbf{x}(t)||^2$, que sabemos que será siempre positivo, luego estaremos contribuyendo a que $\mathbf{w}(t)^T\mathbf{x}(t)$ sea *menos positivo*, o sea, que lo desplazamos en sentido negativo para que se acerque más a 0 o que incluso lo sobrepase y pase a ser negativo.

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

   a) Considere una muestra i.i.d. de tamaño N $(x_1, ..., x_N)$. Mostrar que la función h que maximiza la verosimilitud de la muestra es la misma que minimiza

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

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

   Solución: En un problema de clasificación probabilística (regresión lineal), el concepto de verosimilitud (*likelihood*) sería cómo de probable es que obtengamos una etiqueta y dado un dato $\mathbf{x}$ si nuestra hipótesis h describe correctamente la distribución condicionada $P(y|\mathbf{x})$:

$$P(y|\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{para } y = 1 \\ 1 - h(\mathbf{x}) & \text{para } y = 0 \end{cases}$$

   Cabe destacar que $P(y|\mathbf{x})$ depende de $h(\mathbf{x})$, y $h(\mathbf{x})$ a su vez está en función de $\mathbf{w}$, luego $P(y|\mathbf{x})$ depende de $\mathbf{w}$.

   Puesto que nuestra muestra de tamaño N es i.i.d., todos y cada uno de los N puntos de la muestra son independientes entre sí, luego la probabilidad de obtener todas las etiquetas es

$$\mathcal{L}(\mathbf{w}) = \prod_{n=1}^{N} P(y_n|\mathbf{x}_n)$$

   Nuestra tarea, por tanto, es encontrar la hipótesis h que maximice nuestra verosimilitud o *likelihood* $\mathcal{L}(\mathbf{w})$, lo que se conoce como el *Criterio de Máxima Similitud*. Para hacerlo más sencillo, podemos asumir que **el error que queremos minimizar equivale al opuesto del neperiano de la verosimilitud**, con lo cual, en vez de maximizar $\mathcal{L}(\mathbf{w})$, ahora tendremos que minimizar una función extrictamente decreciente:

$$E_{in}(\mathbf{w}) = -\ln(\mathcal{L}(\mathbf{w})) = -\ln(\prod_{n=1}^{N} P(y_n|\mathbf{x}_n))$$

Y por las propiedades de los logaritmos:

$$E_{in}(\mathbf{w}) = -\sum_{n=1}^{N} \ln(P(y_n|\mathbf{x}_n)) = \sum_{n=1}^{N} \ln(\frac{1}{P(y_n|\mathbf{x}_n)})$$

Finalmente, podemos expandir esta expresión de acuerdo a la definición de $P(y_n|\mathbf{x}_n)$ que hemos dado al comienzo de este apartado:

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

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

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

Solución: Para el caso particular en que $h(x) = \sigma(\mathbf{w}^T\mathbf{x})$, y cambiando las etiquetas, tenemos que:

$$P(y|\mathbf{x}) = \begin{cases} \sigma(\mathbf{w}^T\mathbf{x}) & \text{para } y = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x}) & \text{para } y = -1 \end{cases}$$

Si ahora tenemos en cuenta que

$$\sigma(y_n \mathbf{w}^T \mathbf{x}_n) = \frac{e^{y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}$$

Entonces $P(y|\mathbf{x})$ puede expresarse como $\sigma(y\mathbf{w}^T\mathbf{x})$

Partiendo de la anterior expresión

$$E_{in}(\mathbf{w}) = \sum_{n=1}^{N} \ln(\frac{1}{P(y_n|\mathbf{x}_n)})$$

Vemos que

$$E_{in}(\mathbf{w}) = \sum_{n=1}^{N} \ln(\frac{1}{\sigma(y\mathbf{w}^T\mathbf{x})}) =$$

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

$$\sum_{n=1}^{N} \ln(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n})$$

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

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

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

Solución: En regresión logística tenemos que:

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

Luego el gradiente es igual a:

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

$$= \frac{1}{N}\sum_{n=1}^{N}-y_n\mathbf{x}_n\frac{(e^{-y_n\mathbf{w}^T\mathbf{x}_n})(e^{y_n\mathbf{w}^T\mathbf{x}_n})}{(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n})(e^{y_n\mathbf{w}^T\mathbf{x}_n})} =$$

$$= \frac{1}{N}\sum_{n=1}^{N}-y_n\mathbf{x}_n\frac{1}{e^{y_n\mathbf{w}^T\mathbf{x}_n}+1} = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{x}_n}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}$$

Si ahora tenemos en cuenta que

$$\sigma(y_n\mathbf{w}^T\mathbf{x}_n) = \frac{e^{y_n\mathbf{w}^T\mathbf{x}_n}}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}$$

y que

$$\sigma(-y_n\mathbf{w}^T\mathbf{x}_n) = \frac{1}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}$$

nos queda que

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

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

$$\mathbf{e}_n(\mathbf{w}) = max(0, -y_n \mathbf{w}\mathbf{x}_n)$$

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

Solución: Definimos la función del error a trozos:

$$\mathbf{e}_n(\mathbf{w}) = \begin{cases} 0 & \text{para } -y_n \mathbf{w}\mathbf{x}_n \leq 0 \\ -y_n \mathbf{w}\mathbf{x}_n & \text{para } -y_n \mathbf{w}\mathbf{x}_n > 0 \end{cases}$$

El gradiente del error en un punto, cuando no vale cero, sería:

$$\frac{\partial}{\partial w_j}(-y_n \mathbf{w}\mathbf{x}_n) = -y_n \mathbf{x}_n$$

Tal y como hemos dicho en la definición a trozos, el error es distinto de 0 cuando $-y_n \mathbf{w}\mathbf{x}_n > 0$, o sea, cuando $y_n \mathbf{w}\mathbf{x}_n < 0$. Esta condición se da justamente cuando y_n y $\mathbf{w}\mathbf{x}_n$ tienen distinto signo, es decir, cuando $\mathbf{x}_n$ **está mal clasificado**. En caso contrario, que es cuando esté bien clasificado, el error y su gradiente valdrán 0.

Dicho esto, la regla de actualización del SGD (evaluando un solo punto), sería:

$$w_j \leftarrow w_j - \nu \frac{\partial \mathbf{e}_n(\mathbf{w})}{\partial w_j}$$

En los casos en los que el punto está bien clasificado, el error y el gradiente son nulos, y por tanto la actualización no es efectiva. Sin embargo, para los ejemplos mal clasificados, tendremos que:

$$w_j \leftarrow w_j - \nu(-y_n \mathbf{x}_n)$$

Y como $\nu = 1$:

$$w_j \leftarrow w_j + y_n \mathbf{x}_n$$

Que es justamente la regla de actualización del PLA.

BONUS

1. En regresión lineal con ruido en las etiquetas, el *error fuera de la muestra para una h dada* puede expresarse como

$$E_{\text{out}}(h) = \mathbb{E}_{x,y}[(h(\mathbf{x}) - y)^2] = \int\int (h(\mathbf{x}) - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy$$

   a) Desarrollar la expresión y mostrar que

$$E_{\text{out}}(h) = \int \left(h(\mathbf{x})^2 \int p(y|\mathbf{x}) dy - 2h(\mathbf{x}) \int y \cdot p(y|\mathbf{x}) dy + \int y^2 p(y|\mathbf{x}) dy \right) p(\mathbf{x}) d\mathbf{x}$$

   Solución:

$$E_{\text{out}}(h) = \int\int (h(\mathbf{x}) - y)^2 p(\mathbf{x}, y) = \int\int (h(\mathbf{x})^2 - 2h(\mathbf{x})y + y^2) p(\mathbf{x}, y) d\mathbf{x} dy =$$

$$= \int\int (h(\mathbf{x})^2 - 2h(\mathbf{x})y + y^2) p(y|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dy =$$

$$= \int \left(\int h(\mathbf{x})^2 p(y|\mathbf{x}) dy + \int -2h(\mathbf{x}) y \cdot p(y|\mathbf{x}) dy + \int y^2 p(y|\mathbf{x}) dy \right) p(\mathbf{x}) d\mathbf{x} dy =$$

$$\int \left(h(\mathbf{x})^2 \int p(y|\mathbf{x}) dy - 2h(\mathbf{x}) \int y \cdot p(y|\mathbf{x}) dy + \int y^2 p(y|\mathbf{x}) dy \right) p(\mathbf{x}) d\mathbf{x} dy$$

   b) El término entre paréntesis en E_{out} corresponde al desarrollo de la expresión

$$\int (h(\mathbf{x}) - y)^2 p(y|\mathbf{x}) dy$$

   ¿Qué mide este término para una h dada?

   Solución: En la integral aparece la diferencia entre la predicción y la etiqueta real (error) al cuadrado, multiplicado por la probabilidad de y dado $\mathbf{x}$. La interpretación de esta expresión es la media de errores cuadrados en esa distribución de probabilidad condicional, por lo que se trata de la **varianza**.

c) El objetivo que se persigue en Regression Lineal es encontrar la función $h \in \mathcal{H}$ que minimiza $E_{\text{out}}(h)$. Verificar que si la distribución de probabilida $p(x, y)$ con la que extraemos las muestras es conocida, entonces la hipótesis óptima h^* que minimiza $E_{\text{out}}(h))$ está dada por

$$h^*(\mathbf{x}) = \mathbb{E}_y[y|\mathbf{x}] = y \cdot p(y|\mathbf{x})dy$$

d) ¿Cuál es el valor de $E_{\text{out}}(h^*)$?

e) Dar una interpretación, en términos de una muestra de datos, de la definición de la hipótesis óptima.

2. Una modificación del algoritmo perceptron denominada ADALINE, incorpora en la regla de adaptación una poderación sobre la cantidad de movimiento necesaria. En PLA se aplica w new = w old +y n x n y en ADALINE se aplica la regla $\mathbf{w}_{new} = \mathbf{w}_{old} + \eta(y_n - \mathbf{w}^T\mathbf{x}_n)\mathbf{x}_n$. Considerar la función de error $En(w)) = (max(0, 1 - y_n\mathbf{w}^T\mathbf{x}_n))2$. Argumentar que la regla de adaptación de ADALINE es equivalente a gradiente descendente estocástico (SGD) sobre $\frac{1}{N}\sum_{n=1} NE_n(\mathbf{w})$.

Material citado y/o consultado

- Y.S. Abu-Mustafa, M. Magdom-Ismail, H. Lin, *Learning from Data*, AMLbook.com, 2012
- *A contribution to the mathematical theory of epidemics*, William Ogilvy Kermack and A. G. McKendrick, 1927.

  https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1927.0118
- *Estudio de la optimización del tráfico en un cruce a través del ajuste de los ciclos de los semáforos mediante recocido simulado*

  https://dialnet.unirioja.es/descarga/articulo/6017739.pdf