TRABAJO 1: Cuestiones de Teoría

Autor : Sergio Aguilera Ramírez

1- Identificar, para cada uno 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 $(X, £, Y)$que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los elementos por cada tipo.

a) Clasificación automática de cartas por distrito postal.

   Aprendizaje supervisado x={conjunto de cartas de entrenamiento}
   y= {etiquetas de salida (Dirección)}

   £ = $x \rightarrow y$

b) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.
   - Aprendizaje no supervisado x = {diferentes precios en distintas fechas} , y = {salida del problema (baja , sube)} , £ = $x \rightarrow y$

c) Hacer que un dron sea capaz de rodear un obstáculo.
   - Aprendizaje por refuerzo -> x = {movimientos posibles} ,
     y = {positivo, negativo} , £ = $x \rightarrow y$

d) Dada una colección de fotos de perro, posiblemente de distintas razas, establecer cuantas razas distintas hay representadas en la colección.
   - Aprendizaje no supervisado -> x = {imágenes de perros},
     y ={diferentes razas} , £ = $x \rightarrow y$

2- ¿Cuáles 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.
   - Aproximación por aprendizaje, ya que mediante un conjunto de datos de entrenamiento que contenga diversidad de clases de mamíferos, un algoritmo podría aprender sobre las diferentes clases y clasificarlos de manera correcta.

b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad
   - Aproximación por aprendizaje, podemos determinar la época en las cuales las personas tiene una alta probabilidad de contraer alguna enfermedad o estimar un tanto por ciento de enfermos en un área determinada, en caso de ocurrir alguno de estos casos aplicar dicha campaña de vacunación.

c) Determinar perfiles de consumidor en una cadena de supermercados.
   - Aproximación por aprendizaje, dados un conjunto de consumidores con sus respectivas características, (nº compras, tipo de porductos, dinero gastado etc) podemos aprender a clasificar estos mediante un entrenamiento previo.

d) Determinar el estado anímico de una persona a partir de una foto de su cara.
   - Aproximación por aprendizaje, ya que podemos enseñar a un algoritmo mediante un conjunto de entrada de diferentes imágenes de caras y sus respectivas salidas, a clasificar o determinar el estado anímico de una "cara".

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

   - Aproximación por diseño, podríamos ver las horas del día con mayor tráfico en dicho cruce y aplicar un periodo de cambio de luces en un determinado intervalo del día dependiendo de la cantidad de tráfico. Considerando una gran parte de combinaciones de estas podríamos conseguir el ciclo óptimo.

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?

X = Conjunto de frutas de diferente clase

Y = {salida de la clasificación de las frutas {1(mango),2(papaya), 3(guayabas)}}

D = $(x_1, y_1), \dots , (x_2, y_2)$

£ = X -> Y

El conjunto de frutas X, estaría compuestos por un número determinado de ejemplos (frutas) en los cuales cada uno de ellos estaría constituido por una serie de características, la Y contendría los valores de las respectivas clases de cada ejemplo en X. Además, D se compone de tuplas (x_n, y_n) , que representarían la relación de cada ejemplo con su clase.

Al computador le pasaríamos un subconjunto de X con sus respectivas etiquetas para que este aprenda a clasificar las distintas frutas. En consecuencia, cuando le pasamos un conjunto de ejemplos de entrada sin sus etiquetas, este nos devolvería la salida correspondiente (conjunto de etiquetas asociadas con su x_n).

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 la matriz A y los valores singulares de X.

X = $\begin{pmatrix} 0 & 0 \\ 0 & t \\ r & 0 \end{pmatrix}$

X = UDV^T

U = $\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ V = $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ D = $\begin{pmatrix} t & 0 \\ 0 & r \\ 0 & 0 \end{pmatrix}$

X = $\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} t & 0 \\ 0 & r \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & t \\ r & 0 \end{pmatrix}$

X = => $X^T X = \begin{pmatrix} 0 & 0 & r \\ 0 & t & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & t \\ r & 0 \end{pmatrix} = \begin{pmatrix} r^2 & 0 \\ 0 & t^2 \end{pmatrix}$ => A = $\begin{pmatrix} r^2 & 0 \\ 0 & t^2 \end{pmatrix}$

Solución: Por lo tanto, concluimos que los valores singulares de A son el cuadrado de los valores singulares de X.

5- Sean **x** e **y** dos vectores de características de dimensión M x 1. La expresión

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

Define la covarianza entre dichos vectores, donde $\bar{z}$ representa el valor medio de los elementos de z. Considere ahora una matriz X cuyas columnas representan vectores de características. La matriz de covarianzas asociada a la matriz X = $(x_1, x_2, \dots, x_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores de columnas. Es decir,

$$cov(X) = \begin{pmatrix} cov(x_1,x_1) & cov(x_1,x_2) & \dots & cov(x_1,x_N) \\ cov(x_2,x_1) & cov(x_2,x_2) & \dots & cov(x_2,x_N) \\ \dots & \dots & \dots & \dots \\ cov(x_N,x_1) & cov(x_N,x_2) & \dots & cov(x_N,x_N) \end{pmatrix}$$

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

a) $E1 = 11^T X$
b) $E2 = \left(X - \frac{1}{M}E1\right)^T (X - \frac{1}{M}E1)$

--Para el problema X = $\begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 3 & 5 & 6 \end{pmatrix}$

a) E1 = $11^T X$ = 1 X = $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 3 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 6 & 12 & 9 \\ 6 & 12 & 9 \\ 6 & 12 & 9 \end{pmatrix}$

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

1^T = (1,1,1)

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

Solución: Por lo tanto concluimos que las componentes por filas de E1 es igual a la suma de las componentes de su columna.

b) $E2 = \left(X - \frac{1}{M}E1\right)^T (X - \frac{1}{M}E1) = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 1 \\ -2 & 0 & 3 \end{pmatrix} \begin{pmatrix} -1 & -2 & -2 \\ 0 & 0 & 0 \\ 1 & 1 & 3 \end{pmatrix} =$

$\begin{pmatrix} 2 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 13 \end{pmatrix}$

$\cdot \left(X - \frac{1}{M}E1\right)^T = \left(\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 5 & 6 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} 6 & 12 & 9 \\ 6 & 12 & 9 \\ 6 & 12 & 9 \end{pmatrix} \right)^T = \left(\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 5 & 6 \end{pmatrix} - \begin{pmatrix} 2 & 4 & 3 \\ 2 & 4 & 3 \\ 2 & 4 & 3 \end{pmatrix} \right)^T$

$= \begin{pmatrix} -1 & -2 & -2 \\ 0 & 0 & 0 \\ 1 & 1 & 3 \end{pmatrix}^T = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 1 \\ -2 & 0 & 3 \end{pmatrix}$

$\cdot \left(X - \frac{1}{M}E1\right) = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 5 & 6 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} 6 & 12 & 9 \\ 6 & 12 & 9 \\ 6 & 12 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 3 \\ 3 & 5 & 6 \end{pmatrix} - \begin{pmatrix} 2 & 4 & 3 \\ 2 & 4 & 3 \\ 2 & 4 & 3 \end{pmatrix} =$

$\begin{pmatrix} -1 & -2 & -2 \\ 0 & 0 & 0 \\ 1 & 1 & 3 \end{pmatrix}$

Por lo que podemos concluir:

$\text{Cov(X)} = \begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & cov(x_1, x_3) \\ cov(x_2, x_1) & cov(x_2, x_2) & cov(x_2, x_3) \\ cov(x_3, x_1) & cov(x_3, x_2) & cov(x_3, x_3) \end{pmatrix} = \begin{pmatrix} 0.66 & 1 & 1.6667 \\ 1 & 1.6667 & 2.3333 \\ 1.6667 & 2.3333 & 4.3333 \end{pmatrix}$

Solución: Como vemos E2 = M * cov(X)

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

a) ¿Qué representa la matriz $\widehat{H}$ en un modelo de regresión?
b) Identifique la propiedad más relevante de dicha matriz en relación con regresión lineal.

Justificar las respuestas.

a) Expresa los valores de las observaciones de la variable **y**,

   La matriz de proyección o hat expresa los valores de las observaciones en la variable y, en términos de las combinaciones lineales de los vectores de columna de la matiz X, que contiene las observaciones para cada una de las múltiples variables en las que está aplicando regresión. Esta matriz mapea el vector **y** al vector de valores de predicción(**Ŷ**), por lo tanto describe la influencia que tiene cada valor de **y** en **Ŷ.**

b) La principal propiedad de la matriz hat o de proyección es la idempotencia, esto significa que dicha matriz multiplicada por si misma da como resultado a sí misma. Veamos que $X\ (X^TX)^{-1}X^T$ a x, donde x ϵ C(X). Si x ya está en el espacio de la columna de X, entonces su proyección en C(X) no tendrá ninguna variación. $\hat{H}^2 = \hat{H}$

7- La regla de adaptación de los pesos del Perceptron $(w_{new} = w_{old} + yx)$ 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 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 x(t).

w = (0.5, 1.5, 1.2)

x_1 = 2

y_1 = - 1

1 - Sign($\begin{pmatrix}0.5\\1.5\\2.5\end{pmatrix}$2) = $1 \neq y_1$ => $w_{new} = w_{current} + y * x$ = (0.5, 1.5, 1.2) + -1*2= (-1.5, -0.5, 0.5)

2 - Sign($\begin{pmatrix}-1{,}5\\-0.5\\0.5\end{pmatrix}$2) = -1 = y_1 => $w_{new} = w_{current}$

Como podemos observar el algoritmo del Perceptron nos redirecciona **w** hacia la correcta clasificación de un dato.

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

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

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

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

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

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

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

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

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

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

$\sigma(y_n w^T x_n) \nabla(\frac{1}{\sigma(y_n w^T x_n)}) = \sigma(y_n w^T x_n) - \frac{-\nabla \sigma(y_n w^T x_n)}{\sigma(y_n w^T x_n)^2} = \frac{-\sigma(y_n x_n)\,(1-\sigma(y_n w^T x_n)\,(y_n x_n)}{\sigma(y_n x_n)^2}$

= -($1 - \sigma(y_n w^T x_n)$) $(y_n x_n)$ =

$= -\big(\sigma(-y_n w^T x_n)\big)(y_n x_n) = \nabla E_{in}(w) = \sum_{n=1}^{N} \frac{1}{N} \sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n)$

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

$$e_n(w) = \max(0, -y_n w^T x_n)$$

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

El algoritmo PLA minimiza el error mediante la regla de actualización $w(t+1) = w(t) + y_i x_i$, si aplicamos un learning rate igual a 1 y el gradiente viene representado por $y_i x_i$, podríamos ver esta regla como una regla de SGD, por lo que podemos verificar que la función $e_n(w) = \max(0, -y_n w^T x_n)$ sea utilizada para la obtención del gradiente.