Universidad De Granada

Grado ingeniería informática (2017 – 2018)

Aprendizaje Automático

Trabajo 1: Cuestiones de Teoría

Trabajo realizado por Antonio Miguel Morillo Chica.

---

1. **Identificar, para cada una de las siguientes tareas, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) así como los datos de aprendizaje que deberiamos usar en su caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los datos para cada tipo.**

a) **Dada una colección de fotos de caras de personas de distintas razas establecer cuantas razas distintas hay representadas en la colección.**

La detección de caras en fotografías se hace apartir de aprendizaje supervisado para saber que parte de la foto es o no una cara, problema visto en las transparencias de teoría. Tomando esto, supongo que la distinción de las mismas por etnias se ajustaría a un aprendizaje tambien supervisado ya que existen características bien distintivas, tipo de pelo (nivel de rizado), color de piel, anchura de cara, etc.

Las características de los datos podrían ser las ya citadas aunque como son fotos distintas una posible entrada también podría ser colores de los pixeles, ditancias entre los elementos de la cara (nariz, ojos etc.) en mi caso:

$$X = \{(x_1, x_2, \ldots, x_{10}) \in \mathbb{Z}_2^{10}\}$$

$$Y = \{Raza_1, Raza_2, Raza_3, Raza_4\}$$

Debido al tema controversial de la división por razas o etnias de los seres humanos por parte de biologos, antropólogos y otros expertos no defino la salida como caucasico, asiatico etc.

b) **Clasificación automática de cartas por distrito postal.**

Tomando en cuenta que la lectura del digito postal consiste en la digitalización del número ya que este estaría escrito a mano, el problema se transforma en el reconocimiento de números que, como vimos en teoría, era un ejemplo de aprendizaje supervisado.

Los datos para este problema serían unas imágenes que representan a cada dígito individualmente de un tamaño de 16x16 pixeles lo que supone

un vector con 1024 caracteristicas que definen intensidad o contraste del pixel así pues tendríamos:

$$X = \{(x_1, x_{2,} \dots, x_{256}) \in \mathbb{Z}_{255}^{256}\}$$

$$Y = \{1,2,3,4,5,6,7,8,9\}$$

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

El aprendizaje supervisado se ajusta perfectamente a este tipo de problemas ya que lo que buscamos es una pronosticación futura, además según un artículo de la Unicersidad de Medellín, Colombia los modelos no lineales se ajustan mejor a este tipo de problemas, aquí puede encontrar el artículo como dato interesante. Para mi sería abordado por un aprendizaje por refuerzo, el aprendizaje supervisado necesitaría una cantidad increíble de características ya que muchos factores influyen en la economía, y del mismo modo, para el no supervidado necesitaríamos una ingente cantidad de datos, por ello creo que la mejor linea de aprendizaje sería basarse en el dato numerico que representa el valor del mercado durante un periodo de tiempo de tal forma que:

$$X = \{(x_1, x_{2,} \dots, x_{90}) \in \mathbb{R}^{90} \times \mathbb{R}\}$$

$$Y = [0,1]$$

Donde los datos vienen dados por los valores en los 90 días anteriores y una predicción dónde Y indica el valor de éxito de la predicción.

d) **Aprender un algoritmo que permita a un robot rodear un obstaculo.**

Creo que el aprendizaje por refuerzo puede ser muy util para crear trayectorias “globales”, es decir, para planificadores globales vistos en Tecnologías y Sistemas Inteligentes. Imaginemos que lo que queremos es generar una trayectoria aproximada lo más optima posible, los datos de entrada puede ser la propia trayectoria un vector de posiciones, posteriormente haremos simulaciones para retroalimentar al algoritmo. De forma que si trabajamos en un mapa de $n \times n$:

$$X = \{(x_1, x_{2,} \ldots, x_n) \in \mathbb{N}^n \times \mathbb{N}\}$$

$$Y = [0,1]$$

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) **Agrupar los animales vertebrados en mamíferos, reptiles, aves, anfibios y peces.**

Aproximación por aprendizaje. Tenemos unos una serie de caracteristicas distinguibles que puedes ser usados para diferenciar a los vertebrados, como por ejemplo, tener pico, plumas, pelaje, numero de patas, etc. Es un típico problema de aprendizaje supervisado donde:

$$X = \{(x_1, x_{2,} \ldots, x_n) \in \mathbb{N}_2^n\}$$

$$Y = \{Mamifero, Reptiles, Aves, Anfibios, Aves, Peces\}$$

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

Aproximación por aprendizaje por diseño ya que el posible desarrollo de una enfermedad en base acriterios como el contagio puede ser pronosticada, o simulada por procedimientos de algoritmos clásicos.

c) **Determinar si un correo electrónico es de propaganda o no.**

Aproximación por aprendizaje. Es además un tipo clásico de clasificación de spam que se usa mucho como ejemplo en clase o en otras clases de la rama. Poseemos un conjunto finito de respuestas, es o no spam, las caractierísticas puede ser la presencia o no de ciertas palabras además de poder dar ejemplos claros de lo que es o no un correo de spam para realizar un conjunto de entrenamiento. De forma que:

$$X = \{(x_1, x_{2,} \ldots, x_n) \in \mathbb{N}_2^n\}$$

$$Y = \{Spam, No\ Spam\}$$

d) **Determinar el estado de ánimo de una persona a partir de una**

foto de su cara.

Aproximación por aprendizaje. Es un problema muy típico al de reconocimiento de caras, una vez reconocidas el proceso sería parecido un aprendizaje supersado de clasificación para distinguir otro tipos de caracteristicas a las del problema de detectar que es y no una cara.

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

Aproximación por aprendizaje ya que es inasumuble desde el punto de vista del diseño, existen demasiadas variables y demasiados posibles estados finales para la solución.

Además según la Pontífica Universidad Católica de Chile según la tesis sobre "Aplicación de algoritmos de aprendizaje estadístico para predecir velocidades de buses con información en tiempo real" este problema es un porblema por aprendizaje.

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.**

La entrada sería una matriz de dos componentes de color y tamaño, donde el color viene definido por la composición de tres colores RGB por ello tiene tres dimensiones, además el tamaño sería siempre de tipo real el cual realmente representaría el radio para determinar el tamaño.

- $X = \{(Colorr, Radio)\} / colorr_n \in \mathbb{R}^3_{255}, radio_n \in \mathbb{R}^{+¿}$

La salida sería la clasificación en las etiquetas de las distintas frutas del problema.

- $Y = \{Mangos, Papayas, Guayabas\}$

El conjunto de entrenamiento viene dado por los primeros 50 ejemplos de las

distintas 50 primeras características.

- $D=\{(x_i, y_i) i=1,2,...,50 / x_i \in X, y_i \in Y\}$

La función f será la que tendremos que encontrar para poder categorizar los ejemplos y predecir futuros.

- $f: X \rightarrow Y$

Evidentemente existirá ruido ya que el tamaño por el radio no siempre será perfecto por muy preciso que sea además que de por si ninguna fruta será totalmente esferica.

4. **Sea X una matriz de números reales de dimensiones N × d, N > d. Sea X = UDV^T su descomposición en valores singulares (SVD). Calcular la SVD de X^{T}X y XXT en función de la SVD de X. Identifique dos propiedades de estás nuevas matrices que no tiene X. ¿Qué valor representa la suma de la diagonal principal de cada una de las matrices producto?**

1. **Para X^{T}X**

$X^TX = (UDV^T)^T\ UDV^T$, por propiedad de matrices sabemos que:

$A = UEV^T = (UEV^T)^T = VE^TU^T$, entonces:

$X^TX = (UDV^T)^T\ UDV^T = (VD^TU^T)\ UDV^T$, también sabemos que $I = U^TU$

$X^TX = (UDV^T)^T\ UDV^T = (VD^TU^T)\ UDV^T = VD^TDV^T$, como D es una matriz con solo diagonal por propiedad de matrices $D^TD = D^2$, entonces:

$X^TX = (UDV^T)^T\ UDV^T = (VD^TU^T)\ UDV^T = VD^TDV^T = VD^2V^T$

Propiedades: Siendo U y D matrices cuadradas donde U es una matriz triangularizable y D la submatriz con los valores propios, entonces UD^T y DU^T serían dos matrices simétricas.

2. **Para XX^T**

$XX^T = (UDV^T)(UDV^T)^T = (UDV^T)(VD^TV^T) = UDD^TU^T$

Propiedades: Siendo U y D matrices cuadradas donde U es una matriz

triangularizable y D la submatriz con los valores propios, entonces UD es la traspuesta de DU es $D^T U^T$.

5. **Sean x e y dos vectores de características de dimensión M × 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** $X=(x_1,x_{2,}\ldots,x_N)$ **es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,**

$$cov(X)=\begin{pmatrix} conv(x_1,x_1)\ conv(x_1,x_2)\ldots conv(x_1,x_N) \\ \ldots \quad \ldots \quad \ldots \quad \ldots \\ conv(x_N,x_1)\,conv(x_1,x_N)\ldots conv(x_N,x_N) \end{pmatrix}$$

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

a)

$$E1=11^{(T)}X=\begin{pmatrix}1_1\\1_2\\\ldots\\1_M\end{pmatrix}\times(1_{1,}1_{2,}\ldots,1_{M0})\times(w_{1,}w_{2,}\ldots,w_N)=\begin{pmatrix}1_11_1\ 1_11_2\ldots 1_11_M\\1_21_1\ 1_21_2\ldots 1_21_M\\\ldots\quad\ldots\quad\ldots\quad\ldots\\1_M1_11_M1_2\cdots1_M1_M\end{pmatrix}\times(w_{1,}w_{2,}\ldots,w_N)$$

$$=\begin{pmatrix}\sum_{i=0}^{C}w1_i\sum_{i=0}^{C}w2_i\ \ldots\ \sum_{i=0}^{C}w_{ni}\\\sum_{i=0}^{C}w1_i\sum_{i=0}^{C}w2_i\ldots\sum_{i=0}^{C}w_{ni}\\\ldots\quad\ldots\quad\ldots\quad\ldots\\\sum_{i=0}^{C}w1_i\sum_{i=0}^{C}w2_i\ \sum_{i=0}^{C}w_{ni}\end{pmatrix}=E1$$

Es decir, E1 es una matriz donde en cada columna está en las sumatorias de sus caracterías.

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

$$\left((w_1, w_2, \ldots, w_N)-\begin{pmatrix} \frac{1}{M}\sum_{i=0}^{C} w_{1i} & \frac{1}{M}\sum_{i=0}^{C} w_{2i} & \cdots & \frac{1}{M}\sum_{i=0}^{C} w_{ni} \\ \frac{1}{M}\sum_{i=0}^{C} w_{1i} & \frac{1}{M}\sum_{i=0}^{C} w_{2i} & \cdots & \frac{1}{M}\sum_{i=0}^{C} w_{ni} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{M}\sum_{i=0}^{C} w_{1i} & \frac{1}{M}\sum_{i=0}^{C} w_{2i} & \cdots & \frac{1}{M}\sum_{i=0}^{C} w_{ni} \end{pmatrix}\right)^{T} \times$$

$$\left((w_1, w_2, \ldots, w_N)-\begin{pmatrix} \frac{1}{M}\sum_{i=0}^{C} w_{1i} & \frac{1}{M}\sum_{i=0}^{C} w_{2i} & \cdots & \frac{1}{M}\sum_{i=0}^{C} w_{ni} \\ \frac{1}{M}\sum_{i=0}^{C} w_{1i} & \frac{1}{M}\sum_{i=0}^{C} w_{2i} & \cdots & \frac{1}{M}\sum_{i=0}^{C} w_{ni} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{M}\sum_{i=0}^{C} w_{1i} & \frac{1}{M}\sum_{i=0}^{C} w_{2i} & \cdots & \frac{1}{M}\sum_{i=0}^{C} w_{ni} \end{pmatrix}\right)$$

Si usamos $\frac{1}{M}\sum_{i=0}^{C} w_{1i}$ como una constante y la llamamos $\bar{w}_N$ y tras hacer las restas de los vectores de caracteristicas en cada paréntisis:

$$\begin{pmatrix} w_{11}-\bar{w}_1 & w_{12}-\bar{w}_1 & w_{11}-\bar{w}_1 & \cdots & w_{1N}-\bar{w}_1 \\ w_{21}-\bar{w}_2 & w_{22}-\bar{w}_2 & w_{21}-\bar{w}_2 & \cdots & w_{2N}-\bar{w}_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ w_{N1}-\bar{w}_N & w_{N2}-\bar{w}_N & w_{N1}-\bar{w}_N & \cdots & w_{NN}-\bar{w}_N \end{pmatrix} \times \begin{pmatrix} w_{11}-\bar{w}_1 & w_{21}-\bar{w}_1 & w_{11}-\bar{w}_1 & \cdots & w_{1N}-\bar{w}_1 \\ w_{12}-\bar{w}_2 & w_{21}-\bar{w}_2 & w_{21}-\bar{w}_2 & \cdots & w_{2N}-\bar{w}_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ w_{1N}-\bar{w}_N & w_{N2}-\bar{w}_N & w_{N1}-\bar{w}_N & \cdots & w_{NN}-\bar{w}_N \end{pmatrix} =$$

$$\begin{pmatrix} cov(x_1,x_1) & cov(x_1,x_2) & \cdots & cov(x_1,x_N) \\ cov(x_2,x_1) & cov(x_2,x_2) & \cdots & cov(x_2,x_1) \\ \vdots & \vdots & \vdots & \vdots \\ cov(x_N,x_1) & cov(x_N,x_2) & \cdots & cov(x_N,x_N) \end{pmatrix} = E2$$

Es decir, E2 es la matriz de covarianzas del vector de características, E2 = cov(X).

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

Las demostraciones del ejercición han sido obtenidas de un canal de youtube que se dedica a la divulgación del aprendizaje automático aquí.

a) **Mostrar que H es simétrica**

Suponemos que $H^T = X(X^TX)^{-1}X^T$ ya que si fuesen simetricas H y H^T serían iguales así:

$$\begin{aligned} H^T &= X(X^TX)^{-1}X^T \\ &= X\,[\,(X^TX)^{-1}]^TX^T \\ &= X\,(\,(X^TX)^T)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

b) **Mostrar que es idempotente $\mathbf{H^2 = H}$**

Si H fuese idempotente entonces la matriz al cuidadro sería lo mismo que sin elevarla a nada asi pues:

$$\begin{aligned} H^2 &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-!}X^T) \\ &= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T \\ &= X\,(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

c) **Que representa la matriz H en un modelo de regresión?**

Si H es un modelo de regresión entonces todos sus vecino no son negativos, de esta forma:

$Hw = \lambda w \rightarrow$

$H^2w = \lambda Hw = \lambda(\lambda w)$; pero como H es idepotente: $\rightarrow$

$H^2\, w = Hw = \lambda w \rightarrow$

$\lambda^2\, w = \lambda w \rightarrow$

$\lambda(\lambda\text{-}1)w = 0 \rightarrow \lambda = 0$ o 1

7. **La regla de adaptación de los pesos del Perceptron ($\mathbf{w}_{new} = \mathbf{w}_{old} + \mathbf{yx}$) tiene la interesante propiedad de que los mueve 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 que la regla de adaptación de pesos siempre produce un movimiento en la dirección correcta para clasificar bien x(t).**

x(t) clasifica según $w^t x(t)y(t) > 0$ si está bien clasificada y $w^t x(t)y(t) <= 0$ si no tal y como vemos en teoría si estuviera mal clasificada entonces

y(t) != sing($W^{T(t)}$+x(t)y(t)) en cambio y(t) = sing($W^{T(t)}$+x(t)y(t)) si lo estuviese bien. Teniendo esto en cuenta y siguiendo la regla de adaptción vemos que pase en t+1: $W^{T(t+1)}$x(t)y(t)) donde $W^{T(t+1)} = W^{T(t)}$+x(t)y(t)

De esta forma si $W^{T(t+1)}x(t)y(t) > W^{T(t)}x(t)y(t) \rightarrow (W^{T(t)}+x(t)y(t))$ x(t)y(y) $> W^{T(t)}$x(t)y(t) lo que supone añadir un peso extra x(t)y(t) que hará que el valor de clasificación sea mejor en x(t+1) que en x(t). Si no está bien clasificado lo provocará es que pueda ser clasificado más facilmente en los x(t+n).

8. **Sea un problema probabilistico de clasificación binaria cuyas etiquetas son {0,1}, es decir P(Y=1) = h(x) y P(Y=0) = 1−h(x).**

a) **Dar una expresión para P(Y) que sea válida tanto para Y=1 como para Y=0.**

Al igual que en las transparencias de teoría tenemos que:

$$P(y|x) = \{ h(x) \text{ for } y = +1 \; ; \; 1 - h(x) \text{ for } y = 0$$

b) **Considere una muestra de N v.a. independientes. Escribir la función de Máxima Verosimilitud para dicha muestra.**

c) **Mostrar que la función h que maximiza la verosimilitud de la muestra es la misma que minimiza**

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

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

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

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

9. **Mostrar que la 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.

10. **Definimos 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 $\nu = 1$.**

Sabemos que la forma de actualizar los pesos en el SGD es en primer lugar es ver el gradiante en el instante t con unos pesos w:

$$gt = \text{gradiante } (E_{in}(w(t))$$

Cambiamos la dirección en la que nos movemos y además actualizamos los pesos como en las transparencias de teoría:

$$v = -gt$$

$$w(t+1) = w(t) + mu*v \rightarrow w(t+1) = w(t) + mu*\text{graciante}(E_{in}(w(t)))$$

Ahota tenemos en cuenta que el rror en un punto en un determinado instante

t es $e_n(w) = \max(0, -y_n w^T x_n)$, derivamos respecto de w^T, obteniendo el graciante de la función $e_n(w)$:

$$f(w) = -y_n w^T x_n \rightarrow f'(w) = -y_n x_n$$

Actualizamos los pesos pata t+1, además hay que tener en cuenta que en el enununciado pone que mu = 1:

$$w(t+1) = w(t) + mu\ (\ -y_n x_n\ ) = \ w(t) - y_n x_n$$

Esta es la misma formula que la que tiene la actualización de los pesos el algortirmo PLA.