Primer cuestionario de teoría de Aprendizaje Automático

UNIVERSIDAD DE GRANADA

Antonio Manuel Fresneda Rodríguez

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 deberíamos usar en su caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los datos
para cada tipo.

1) Dada una colección de fotos de caras de personas de distintas razas establecer cuantas razas distintas hay representadas en la colección.
2) Clasificación automática de cartas por distrito postal
3) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.
4) Aprender un algoritmo que permita a un robot rodear un obstáculo.

1. Si queremos diseñar un algoritmo para distinguir las razas de personas hay en una foto podríamos usar aprendizaje supervisado si tenemos un conjunto de muestras analizadas previamente. Las características que identificaría de las distintas razas seria forma de la cabeza, color, y rasgos faciales. Como etiquetas usaría: caucásico, asiático y africano.
   En el caso de que no tuviésemos muestras para analizar, podríamos usar un aprendizaje no supervisado y veriamos que según las características mencionadas anteriormente, el conjunto de las muestras que se separaría.
2. Podríamos usar aprendizaje no supervisado. Usando este tipo de aprendizaje, el algoritmo se encargaría de identificar que códigos postales son los que se van agrupando según la localización y a partir de esa información poder predecir nuevas muestras.
3. Este problema podríamos abordarlo desde el aprendizaje supervisado. Podríamos usar las características de dicho valor en momentos anteriores y el estado en el que estaba (subida, bajada, estable) y así poder estimar si subirá o bajará
4. Podríamos diseñarlo mediante aprendizaje por refuerzo. En el caso de que choque con el obstáculo, este marcaría en un mapa que por esa zona a chocado y evitaría pasar por esa zona.

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

1. Agrupar los animales vertebrados en mamíferos, reptiles, aves, anfibios y peces.
2. Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.
3. Determinar si un correo electrónico es de propaganda o no.
4. Determinar el estado de ánimo de una persona a partir de una foto de su cara.
5. Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.

1. La agrupación de animales es bastante adecuado para un problema de diseño. Podríamos usar distintas características (vuela, tiene pelo, tiene plumas, vive en el agua, etc)
2. Para este problema yo usaría una aproximación por diseño, puesto que podríamos usar características de enfermedades donde si ha sido necesaria la vacuna. Es un problema parecido al de determinar si el correo electrónico es spam.
3. Este tipo de problema tendría una aproximación por diseño, puesto que la mayoría de los correos de propaganda suelen ser de empresas conocidas o en

paginas webs donde el usuario se ha registrado. No seria difícil tener una serie de remitentes que sabemos que son spam (han sido reportados por otros usuarios, provienen de empresas o de alguna pagina donde nos hayamos registrado) y diseñar un algoritmo que los separe.

4. Este problema seria difícil de realizar mediante diseño puesto que las muestras tendrían bastante ruido (con ruido me refiero a las caras de la gente cuando esta feliz, triste, etc puede variar mucho). Con un modelo de aprendizaje y las suficientes muestras podríamos generar una muy buena solución a este problema.
5. Este problema no seria apto para un enfoque por aprendizaje, puesto que por aprendizaje siempre vamos a tener una probabilidad de error, y puesto que es un un problema que cualquier fallo puede provocar un accidente, no usaría un algoritmo de 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.

X = Vector de características: [Tamaño, peso, color, forma]

1. Tamaño: Medida del tamaño medio de cada fruta.
2. Peso: Peso medio de cada fruta.
3. Color: Distintos colores posibles para cada fruta.
4. Forma: Redonda, ovalada, etc

Y = Etiquetas: [Guayaba, mangos, papaya]

D =Tupla que asigna a cada vector de características una posible etiqueta

[Grande, 200gr, verde-rojo, ovalada]→ Mango

F = F definida como $F: X \rightarrow Y$, es la función óptima que clasificaría perfectamente cada una de las muestras (es este caso fruta). Esta función es desconocida y la vamos a intentar estimar con nuestro modelo usando la función h.

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 $X X^T$ en función de la SVD de X. Identifique dos propiedades de estas nuevas matrices que no tiene X ¿Qué valor representa la suma de la diagonal principal de cada una de las matrices producto?

Sabemos que V y U son matrices ortogonales y que D es una matriz diagonal cuadrada.

$X^T = (UDV^T)^T \rightarrow (V^T)^T D^T U^T = VDU^T$

$X^T X = (VDU^T UDV^T) \; como \; U^T U = I \rightarrow VD^2 V^T$

$X X^T = UDV^T V D U^T = como \; V^T V = I \rightarrow XX^T = UD^2 U^T$

Estas dos matrices son matrices cuadradas $N \times N$ y son matrices simétricas, es decir,

$$A = A^T = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} = \begin{bmatrix} a^2+d^2 & ab+de & ac+df \\ ba+ed & b^2+e^2 & bc+ef \\ ca+fd & cb+fe & c^2+f^2 \end{bmatrix}$$

(Los valores de la matriz cumplen la propiedad conmutativa para la multiplicación y para la suma)

La suma de la diagonal representa la suma de los cuadrados de los valores propios.

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

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

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 = ($x_1, x_2, \cdots, x_N$) es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

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

Sea 1_M^T = (1, 1, · · ·, 1) un vector M × 1 de unos. Mostrar que representan las siguientes expresiones

a) E1 = $11^T X$

Tenemos que 1 es un vector 1 x M de 1, que multiplicado por 1^T (vector M x 1) obtenemos una matriz M x M cuyos coeficientes son todos 1.
Si esta matriz la multiplicamos por X obtenemos la siguiente matriz:

$$\begin{bmatrix} x_{11}+x_{21}\dots+x_{M1} & x_{12}+x_{22}..+X_{M2} & \cdots & x_{1N}+x_{2N}\dots+x_{MN} \\ \vdots & & & \vdots \\ x_{11}+x_{21}\dots+x_{M1} & x_{12}+x_{22}..+X_{M2} & \cdots & x_{1N}+x_{2N}\dots+x_{MN} \end{bmatrix}$$

Que en cada fila contiene la suma de todas las características de la muestra de la columna que pertenece a la columna i en la matriz X

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

$$\frac{1}{M}E1=\begin{bmatrix} \frac{x_{11}+x_{21}\dots+x_{M1}}{M} & \frac{x_{12}+x_{22}..+X_{M2}}{M} & \cdots & \frac{x_{1N}+x_{2N}\dots+x_{MN}}{M} \\ \vdots & & & \vdots \\ \frac{x_{11}+x_{21}\dots+x_{M1}}{M} & \frac{x_{12}+x_{22}..+X_{M2}}{M} & \cdots & \frac{x_{1N}+x_{2N}\dots+x_{MN}}{M} \end{bmatrix}$$

Si tenemos la sumatoria y la dividimos entre M nos va a dar la media de cada característica de la columna i y si le restamos X obtenemos:

$$X-\frac{1}{M}E1=\begin{bmatrix} x_{11}-\overline{x_1} & x_{12}-\overline{x_2} & \cdots & x_{1N}-\overline{x_4} \\ \vdots & & & \vdots \\ x_{11}-\overline{x_1} & x_{12}-\overline{x_2} & \cdots & x_{1N}-\overline{x_4} \end{bmatrix}$$

$$\left(X-\frac{1}{M}E1\right)^T\left(X-\frac{1}{M}E1\right)=\begin{bmatrix} \sum_{i=1}^{M}(X_{i1}-\overline{x_1})^2 & \cdots & \cdots & \sum_{i=1}^{M}(X_{i1}-\overline{x_1})(X_{Ni}-\overline{x_n}) \\ \vdots & & & \vdots \\ \sum_{i=1}^{M}(X_{Mi}-\overline{x_M})(X_{1i}-\overline{x_1}) & \cdots & \cdots & \sum_{i=1}^{M}(X_{Mi}-\overline{x_M})(X_{NM}-\overline{x_n}) \end{bmatrix}$$

Cuando vemos esta matriz, vemos que tiene cierta simetría con la matriz cov(x), solo que falta que dividamos por M en cada posición de la matriz.

$$\frac{E2}{M}=\begin{pmatrix} cov(x_1,x_1) & cov(x_1,x_2) & \dots & cov(x_1,x_N) \\ cov(x_2,x_1) & cov(x_2,x_2) & \dots & cov(x_2,x_N) \\ \dots & \dots & \dots & \dots \\ cov(x_N,x_1) & cov(x_N,x_2) & \dots & cov(x_N,x_N) \end{pmatrix}$$

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

a) Mostrar que H es simétrica

H es simétrica si y solo si $H=H^T$

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

b) Mostrar que es idempotente H^2 = H

$H^2=(X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)$ → *Podemos agrupar X^T del primer termino y X del segundo* →

$(X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T)$ → $(X^TX)^{-1}(X^TX)=I$ → $(XI(X^TX)^{-1}X^T)=(X(X^TX)^{-1}X^T)=H$

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

En regresión, H representa la matriz de proyección ortogonal en el espacio generado por las variables regresoras.

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

8. Sea un problema probabilístico 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

$y\in[0,1]P(y)=(1-h(x))^{1-y}*h(x)^y$

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

$L(w)=\prod_{i=1}^{N}P(Y_i/x_i)=\prod_{i=1}^{N}\sigma(y_iw^tX_i)$

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

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

¿

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

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

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

9. 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

Partimos de que σ es la función sigmoide: $\sigma(x)=\frac{1}{1-e^{-x}} \rightarrow \sigma(-y_n w^T x_n)=\frac{1}{1-e^{y_n w^T x_n}}$

Sustituimos en la segunda ecuación: $\frac{1}{N}\sum_{n=1}^{N}-y_n x_n \sigma(-y_n w^T x_n)=\frac{-1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1-e^{y_n w^T x_n}}$ y comprobamos que se verifica.

Partimos de $\frac{-1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1-e^{y_n w^T x_n}}$, y de que las etiquetas son {-1,1}, tenemos dos casos:

1. Que la etiqueta x_n está bien clasificada → $y_n w^T x_n>0$ → el denominador de la función crece y por tanto el valor que se añade a la sumatoria decrece (tiende a 0).
2. Que la etiqueta x_n está mal clasificada → $y_n w^T x_n<0$ → el denominador de la función decrece y el valor que se añade a la sumatoria crece.

Por tanto, podemos decir que una etiqueta mal clasificada contribuye al gradiente más que un ejemplo bien clasificado

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 en con tasa de aprendizaje ν = 1.

La fórmula para el cálculo de w en el SGD es $w=w-\mu\nabla E_{in}(w)$
La regla de actualización del algoritmo PLA es: $w(t+1)=w(t)+y_n x_n$
Vamos a sustituir E_{in} por e_n

$$w=w-\mu\nabla e_n=w-\mu\frac{\partial}{\partial w}(max(0,-y_n w^T x_n))$$

$$\frac{\partial}{\partial w}e_n=\frac{\partial}{\partial w}(max(0,-yw^T x_n))=max(0,-y_n x_n)$$

Sustituyendo tenemos que:

$$w=w-\mu\nabla e_n=w-\mu(-y_n x_n)\rightarrow w=w-1(-y_n x_n)=w+y_n x_n$$

Demostramos que usando este e_n sobre el SGD nos da la misma regla que el PLA por tanto podemos ver como el algoritmo PLA se puede interpretar como SGD con tasa de aprendizaje = 1 usando este error.

Bibliografía.

Me he apoyado en las transparencias de teoría, en los libros que hay en la plataforma, en el libro Learning from Data y dudas con algunos compañeros de clase.

https://en.wikipedia.org/wiki/Projection_matrix