Aprendizaje Automático - Cuestionario 1

David Gil Bautista

Preguntas

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 ra`

Para este problema tenemos una serie de datos y debemos categorizarlos en distintos tipos utilizando sus características. Se trata de un problema de aprendizaje no supervisado, ya que, a partir de unos datos de entrada establecemos una serie de relaciones entre ellos para determinar el número de razas distintas que tenemos. Se podría tener un vector de características (vector de flotantes) en las que se presentaran los distintos rasgos; color de ojos, anchura de nariz, grosor de labios... etc. Las mismas razas tendrán valores similares y al representar los datos veremos que cada raza se agrupará en un espacio.

`b) Clasificación automática de cartas por distrito postal`

Si tratamos este problema como uno de predicción podríamos usar aprendizaje supervisado y a partir de una serie de datos de entrenamiento calcular una función que nos permita clasificar las nuevas cartas. Para este caso necesitaríamos una serie de datos para calcular nuestra función, las cartas con sus distritos postales y su etiqueta. Es cierto, también, que podríamos tratar este problema con aprendizaje no supervisado, ya que, como en el apartado anterior, podríamos representar los datos para ver como se estructuran al representarlos y ver las relaciones que hay entre ellos.

`c) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un peri`

Este problema podríamos tratarlo aplicando la teoría de juegos y el aprendizaje por refuerzo. Para estimar si un índice del mercado de valores subirá o bajará deberíamos estudiar todas las posibilidades que tiene y elegir la más probable.

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

Aprendizaje por refuerzo. Para este problema supondré que el robot tiene una serie de sensores y memoria. Sabiendo esto se me ocurriría generar un árbol de decisión que se generará mediante los datos de entrada del robot y el objetivo (que es rodear el objeto). Al ver un obstáculo el robot solo tomará una serie de decisiones (secuencia de pasos), lo que equivale a aprender un algoritmo.

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.`

Para agrupar un grupo de datos en distintas categorías usaría una aproximación por aprendizaje, ya que, a partir de una sería de datos iniciales con sus etiquetas calcularía una función que me permitiera predecir a que familia de animales pertenecen una serie de datos y agruparlos.

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

Para este problema sería mejor aplicar una aproximación por diseño. Digamos que tenemos una enfermedad y una serie de datos probabilísticos asociados a dicha enfermedad que se recogen cada año. Con esos datos podríamos construir una distribución de probabilidad que nos permita secidir si se aplica una campaña de vacunación o no.

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

Debemos predecir si un correo electrónico es SPAM o no. Podríamos obtener una función a partir de un conjunto de datos de entrenamiento y sus etiquetas que leyera un vector con las palabras y a partir de este determinara si es SPAM o no.

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

En este caso contamos con una serie de datos iniciales, un vector con las facciones faciales, y sus etiquetas. A partir de esos datos usando aproximación por aprendizaje tendríamos una función que nos permita determinar el estado de ánimo de una persona a partir de esos datos.

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

Al no tratarse de de un problema mediante el cual tengamos que estimar una salida a partir de unos datos iniciales podemos deducir que lo mejor sería aplicar una aproximación 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 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.

Nuestros elementos formales serán los siguientes:

X : Vector de flotantes en los que representamos las siguientes características: Carbohidratos, Grasas, Proteínas y Agua. Todos estos datos medidos por cada 100 gramos de la pieza.

Y : Nuestra etiqueta sería el tipo de fruta {mango, papaya, guayaba} identificados por una etiqueta única.

D : Conjunto de datos de entrada para entrenar.

f : Función tal que dado un vector con esos 4 elementos determine a que tipo se ajusta mejor un dato.

Nos encontramos ante un problema con etiquetas con ruido puesto que para una determinada pieza de fruta puede que sus características determinen que se trate de una distinta a la que es.

4) Sea X una matriz de números reales de dimensiones NÖ$d, N > d$. Sea $X = UDV^T$ su descomposición en valores singulares (SVD). Calcular la SVD de X^TX y XX^T 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?

$$X^TX = VDU^TUDV^T = VD^2V^TXX^T = UDV^TVDU^T = UD^2U^T$$

Cuando tenemos un producto de matrices y necesitamos calcular su inversa es el mismo producto en orden invertido en el que cada matriz es su traspuesta. De esta forma obtenemos X^T.

Al hacer X^TX tenemos VDU^TUDV^T donde podemos aplicar una propiedad de las matrices ortogonales en la que al multiplicar una matriz por su traspuesta obtenemos la matriz identidad. Obviando esta matriz, puesto que su producto por cualquier otra genera esa misma matriz, obtenemos VD^2V^T.

1) La matriz U es una matriz ortogonal y la matriz D es una matriz diagonal.

2) El orden de la multiplicación supone que obtengamos la descomposición en valores singulares en función de U o de V.

3) Sea $(XX^T)^n$ con $n >= 1$ tenemos que su descomposición en valores singulares es UD^NU^T.

4) Sea $(X^TX)^n$ con $n >= 1$ tenemos que su descomposición en valores singulares es VD^NV^T

5) Las matrices X^TX y XX^T son simétricas.

5) Sean x e y dos vectores de características de dimensión MÖ1. La expresión

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

define la covarianza a 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 $X = (x_1, x_2, \text{ůůů}, x_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir

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

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

1. $E1 = 11^T X$

   Si tenemos que 1^T es un vector de M unos, al multiplicar por 1 obtenemos una matriz 1_{MxM}, es decir, una matriz cuadrada de orden M rellena de unos.

   Al multiplicar dicha matriz por X tenemos un producto matricial en el que sumamos todas las columnas de X, es decir, tendremos el sumatorio de la columna X,

   Sabiendo que E1 es la sumatoria de las columnas, al dividir entre M tenemos una matriz en la que tenemos la media de las columnas de X.

   Al multiplicar $(X - \bar{X})^T$ por $(X - \bar{X})$ obtenemos una matriz con la sumatoria del producto de las columnas, lo que podemos representar como $\sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y})$.

   De este modo hemos demostrado que $E2 = \sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y}) = M\ cov(x, y)$

6) Considerar la matriz *hat* definida en regresión, $H = X(X^T X)^1 X^T$, donde X es una matriz

$N\ddot{O}(d + 1)$, y $X^T X$ **es invertible.**

a) Mostrar que H es simétrica.

Si H es simétrica podemos decir que $H = H^T$

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

Según las propiedades de las matrices al hacer la traspuesta de un producto de matrices se colocan en orden inverso y haciendo su traspuesta mediante lo que obtenemos $X((X^T X)^{-1})^T X^T$
. Una vez hecho esto intercambiamos el orden de las potencias en $((X^T X)^{-1})^T$

para obtener $((X^TX)^T)^{-1}$ y aplicando la misma propiedad del producto de traspuestas obtenemos que $(X^TX)^T$ es $(X^TX$. Al final obtenemos que $H^T = X(X^TX)^{-1}X^T$ por lo que demostramos que H es simétrica.

b) Mostrar que es idempotente $H^2 = H$

$$H^1 = HH^2 = HH^2 = HH = HH = X(X^TX)^{-1}X^THH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^THH = X(X^TX)^{-1}$$

Al hacer H^2 podemos ver que obtenemos $X^TX(X^TX)^{-1}$ y según las propiedadesde las matrices; una matriz por su inversa es la identidad. Haciendo esto podemos ver que obtenemos la matriz original. De este modo demostramos que H es una matriz idempotente.

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

La matriz H es la matriz de derivadas de segundo orden que indica el avance en la dirección del gradiente.

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** $\boldsymbol{w}$ **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** $x(t)$**.**

Tenemos que demostrar que si en una iteración hemos clasificado mal un dato en la siguiente iteración el perceptron se moverá de forma que lo clasifique bien. Para ello tenemos lo siguiente:

$$y(t)(w^T(t)x(t)) < 0y(t)(w^T(t+1)x(t)) > 0y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)y(t)(w(t)+y(t)x(t))^Tx(t) > y(t)w^T($$

Si está mal clasificado la y será distinta de w^Tx, por lo que al ser uno negativo siempre será menot que 0. Si está bien clasificado puede que ambos sean negativos o ambos positivos, por lo que la multiplicación siempre tendrá signo positivo. Si $y(t)(w^T(t)x(t)) < 0$, está mal clasificado, en la siguiente iteración $y(t)(w^T(t+1)x(t)) > 0$, estará bien clasificado, por lo que podemos decir que $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$.

Desarrollando esta expresión y sabiendo que $w_{new} = w_{old} + yx$ tenemos que $y(t)(w^T(t) + y(t)x(t))x(t) > y(t)w^T(t)x(t)$. Al multiplicar la etiqueta por la etiqueta vamos a obtener un 1 multiplicado por la traspuesta de x, y al multiplicar dicha traspuesta por x obtenemos x^2 lo que garantiza que vaya a ser positivo y por tanto mayor que en la iteración anterior.

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

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

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*

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

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

d) Para el caso $h(x) = (w^T x)$ 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) Mostrar que en regresión logística se verifica:

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

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

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

Despejando la fracción obtenemos $\frac{1}{1+e^{y_n w^T x_n}}$ y como ya hemos visto en la expresión de sigma, esto equivale a $\sigma(y_n w^T x_n)$. Con esto ya podemos demostrar que ambas expresiones son equivalentes.

Cuando tenemos un ejemplo mal clasificado, la potencia de $e^{y_n w^T x_n}$ será negativa puesto que al ser de distinto signo la etiqueta del vector de pesos por el de características el signo será negativo. Cuando la potencia de e es un número negativo, e tiende a ser un número pequeño mientras que si la potencia es positiva, e tiende a crecer. Si la potencia de e es un número negativo, la fracción $\frac{1}{1+e^{y_n w^T x_n}}$ será menor que si la potencia de e es un número positivo.

De esta forma demostramos que si un ejemplo está mal clasificado su error será más significativo y por tanto más contribuyente que si está 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 e_n con tasa de aprendizaje $= 1$.