Universidad de Granada

Ingeniería Informática

Computación y Sistemas Inteligentes

Practica 1

Autor: José Antonio Ruiz Millán
Asignatura: Aprendizaje Automático
23 de marzo de 2018

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.**

      Este problema lo clasificaría como un problema de aprendizaje no supervisado, teniendo como datos de entrada los distintos rasgos de la cara como pueden ser el tamaño de la nariz, de los labios, la largura y anchura de la cara, el color de piel, entre otros. Gracias a estos datos, conseguiríamos crear "grupos" entre las distintas caras a los que finalmente denominaríamos razas. Por lo que al final, podríamos diferencias las distintas razas entre todas las personas que hayan sido procesadas.

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

      En este problema utilizaría aprendizaje no supervisado. En este caso únicamente necesitamos agrupar los datos por distrito postal, que es el dato que utilizaríamos como dato de entrada, y simplemente con eso podríamos clasificarlas por grupos.

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

      Para este problema utilizaría aprendizaje por refuerzo. Ya que la economía no sigue un crecimiento o unos movimientos constantes, por lo que aprendizaje por refuerzo iría aprendiendo respecto a como la economía está en el momento en el que nos encontramos y así poder estimar sobre lo que tenemos y viendo como ha ido cambiando desde el pasado al día de hoy.

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

      Para este tipo de problema utilizaría aprendizaje por refuerzo. Tendríamos una función que va premiando lo "bueno" y perjudicando lo "malo" por lo que si el robot chocase con algun objeto, ese valor sería muy malo e iría aprendiendo que no chocarse con el y alejarse de él es el mejor resultado posible respecto a la función que tendríamos definida. Finalmente el robot sabría que chocarse es malo y no chocarse es bueno por lo que cuando se enfrentase con un objeto, lo esquivaría y evitaría el choque.

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.**

      Este problema lo clasificaría como una aproximación por diseño. Sabemos qué características cumplen cada especie de los animales vertebrados, por ello, recopilaríamos información sobre cada tipo de animal (mamífero, resptiles...) respecto al número de patas, reproducción, entre otros, que son características que definen unívocamente a cada una de las especies. Con estos datos, construimos un modelo físico teniendo en

cuenta las variaciones de error medidos, y finalmente construimos una distribución de probabilidad que utilizamos para clasificar.

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

En este problema utilizaría aproximación por aprendizaje. Podríamos obtener diversos datos sobre las distintas enfermedades como síntomas, probabilidad de contagio, medios de contagio, gravedad de la enfermedad, etc... y a raíz de estos datos poder decidir si es necesario realizar una campaña o no para esa enfermedad.

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

Para este problema utilizaría aproximación por aprendizaje. Supongamos que utilizamos aprendizaje supervisado, podríamos obtener los datos del texto como datos de entrada para el aprendizaje e indicarle si ese mensaje es spam o no. Con suficiente información y entrenamiento conseguiríamos crear un clasificador que dado un mensaje, nos dijera si ese mensaje es spam o no.

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

Como en el caso anterior, utilizaría aproximación por aprendizaje por la misma razón. Supongamos que utilizamos aprendizaje supervisado, podemos a través de diversas fotografías de caras, obtener información sobre ellas y utilizar esa información como datos de entrada, indicando el estado de ánimo de esa persona. Finalmente despues de un buen entrenamiento tendríamos un clasificador que dada una imagen reconocería el estado de ánimo de la misma.

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

Este problema necesita de una exactitud notable ya que es un problema crítico. Por ello, utilizaremos aproximación por diseño ya que estudiando el problema podemos definir una secuencia mejor o al menos mas fiable que aproximacion 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.**

Empezaremos definiendo X, que será el conjunto de todos los x_n donde cada x_n es un elemento a clasificar y que estará compuesto por todo el conjunto de características como pueden ser el diámetro de la fruta, una imagen, el color, textura, etc...

Ahora definimos Y, que será el conjunto de valores que nos permitirán clasificar las distintas frutas. En este caso, al tener 3 frutas, Y sería un valor perteneciente al conjunto $[-1, 1]$ donde -1 indica por ejemplo que es un mango, 0 que es una papaya y 1 que es una guayaba.

Para terminar de definir Y, vamos a definir f que sería la funcion que dado un x_n nos devolvería su correspondiente y_n que como hemos comentado antes, es un valor comprendido

entre $[-1, 1]$. Si ese valor está mas cera de -1 que de 0, se clasificaría como un mango, si el valor esta más cerca de 0 que de -1 y 1, esta fruta se clasificaría como una papaya, y finalmente si el valor está mas cerca de 1 que de 0, sería una guayaba.

Por último definimos D, que no es nada mas que el conjunto de datos con sus etiquetas, por lo que $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$.

4. **Sea X una matriz de números reales de dimensiones $N \times d, N > d$. 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?**

   Tenemos que $X = UDV^T$ por lo que $X^T = (UDV^T)^T = VD^TU^T$

   Pasamos ahora a calcular X^TX usando las fórmulas que acabamos de indicar.

   - $X^TX = VD^TU^TUDV^T$

     Sabemos que U es una matriz ortogonal, por lo que $U^T = U^{-1}$. Gracias a esta propiedad conseguimos eliminar U de la fórmula ya que tenemos $U^TU = U^{-1}U$.

   - $X^TX = VD^TDV^T$

     Tenemos que D es la diagonal. Por lo que tenemos una propiedad que nos dice que si D es diagonal, $D^T = D$. Con este paso sustituimos D^T por D

   - $X^TX = VDDV^T$

     Finalmente hemos obtenido este valor y podemos dar por finalizado el calculo de X^TX

   Ahora pasamos a calcular XX^T de la siguiente manera:

   - $XX^T = UDV^TVD^TU^T$

     Al igual que los apartados anteriores, sabemos que V es una matriz ortogonal, por lo que $V^T = V^{-1}$ asique podemos sustituir en la formula y eliminar V de la formula ya que $V^TV = V^{-1}V$

   - $XX^T = UDD^TU^T$

     En este caso al igual que el anterior, tenemos D^T que como sabemos que es la diagonal, tenemos que $D^T = D$

   - $XX^T = UDDU^T$

   Podemos observar que $(X^TX)^T = X^TX$, recalco esta propiedad porque nos será útil en un ejercicio posterior.

5. **Sean x e y dos vectores de características de dimensión $M \times 1$. La expresión:**

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

   define la covarianza entre dichos vectores, donde $\overline{z}$ 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 $X =$

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

$$cov(x, y) = \begin{pmatrix} cov(x_1, y_1) & cov(x_1, y_2) & ... & cov(x_1, y_N) \\ cov(x_2, y_1) & cov(x_2, y_2) & ... & cov(x_2, y_N) \\ ... & ... & ... & ... \\ cov(x_N, y_1) & cov(x_N, y_2) & ... & cov(x_N, y_N) \end{pmatrix}$$

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

a) $E1 = 11^T X$

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

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

   a) **Mostrar que H es simétrica**

   Primero definiremos que significa que H sea simétrica. Esto se cumple cuando $H^T = H$ por lo que:

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

     Para este paso he utilizado la siguiente propiedad: $((X)^{-1})^T = ((X)^T)^{-1}$
   - $H^T = X(X^T X)^{-1} X^T = H$

     Para este paso he utilizado la siguiente propiedad:(4) $(X^T X)^T = X^T X$.

   Por lo que finalmente podemos afirmar que H **es simétrica**.

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

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

     En este paso hemos utilizado la siguiente propiedad: $(X^T X)^{-1}(X^T X) = I$ en el que estamos multiplicando un valor por su inversa, por lo que podemos eliminarlo.

     Podemos ver y comprobar que efectivamente H **es idempotente**

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

   Tenemos que $X\dagger = (X^T X)^{-1} X^T$. En nuestro caso, tenemos que $H = X(X^T X)^{-1} X^T$ que no es mas que $XX\dagger$. Por lo que tenemos la matriz X multiplicando a la pseudoinversa.

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).

En este caso, sabemos que tanto y como x, son variables "constantes", estas variables indican la etiqueta del valor (y) y las características de un determinado valor (x). Por lo que aunque el programa avance, estos valores no cambian.

Con esto sólo nos queda ver que si un x está mal claseificado, al sumar su valor al peso actual, esto hará que la recta se acerque hacia su posición para conseguir meterlo en su posicion correcta. Sin embargo, si x está bien clasificado, al sumar sus valores a los pesos, la recta se ajusta y se aleja de el punto hacia otra posicion conservando el punto en su sitio. Por lo que finalmente, siempre tenemos que la recta se va a mover hacia la direccion correcta.

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

   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 $h(x) = \sigma(w^T x)$ 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 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.

Para ello, partimos de que en regresión logística[1],

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

También podemos asumir que[1]

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

Por último, para poder hacer la demostración necesitamos saber una característica básica de la regresión logística[1]

$$\sigma(s) = \frac{e^s}{1+e^s}$$

Pasamos ahora a demostrar lo que el enunciado nos pregunta, partiremos de la segunda forma de la fórmula hasta llegar a la primera para demostrar que son iguales.

- El primer paso a realizar será sustituir $\sigma(-y_n w^T x_n)$ usando la propiedad anterior.

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

- El siguiente paso será bajar $e^{-y_n w^T x_n}$ del numerador al denominador, cambiando el signo del exponente.

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

- Realizamos la multiplicacion que tenemos en el denominador

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

- Por último, sacamos el "-" de la sumatoria y ponemos $y_n x_n$ como numerador.

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

Por lo que finalmente hemos demostrado que

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

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 $v = 1$.

Referencias

[1] Learning from Data Short Course, Yser S. Abu-Mostafa; Malik Magdon-Ismail; Hsuan-Tien Lin, `https://libsvm.com/data/learn.from.data.pdf`, Accedido el 23 de marzo de 2018.