Aprendizaje Automático (2018-2019)
Grado en Ingeniería Informática
Universidad de Granada

Trabajo 1: Cuestiones de teoría

Montserrat Rodríguez Zamorano

5 de abril de 2019

Índice

1. Preguntas obligatorias **1**

1. Preguntas obligatorias

1. Identificar, para cada una de las siguientes tareas, cual es el problema, qué tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje (X, f, Y) que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar cómo y describir los elementos para cada tipo.

   a) Clasificación automática de cartas por distrito postal.

   El tipo de aprendizaje adecuado sería *aprendizaje supervisado*. Para resolver el problema tendremos que identificar los caracteres que forman el distrito postal.

   Los datos de entrada X pueden ser características útiles en la identificación de caracteres (ejemplo: intensidad, simetría) junto con el dígito al que se asocia, es decir, su etiqueta. El conjunto de etiquetas sería Y. Con estos datos tendremos nuestra función objetivo f que será capaz de clasificar números.

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

   Se trata de *aprendizaje supervisado*. Podemos tener como conjunto de entrada X los datos de períodos anteriores identificando características que nos resulten útiles (por ejemplo, mes del año en el que nos encontramos, nivel de paro, entre otros) etiquetados con el resultado que haya tenido: subida o bajada del índice (estas etiquetas forman Y). A partir de estos datos podremos tener una función objetivo f que sea capaz de precedir la subida o bajada del índice.

   c) Hacer que un dron sea capaz de rodear un obstáculo.

   El tipo de aprendizaje que escogería para esta tarea sería *aprendizaje por refuerzo*, ya que el dron interactúa con su entorno y toma decisiones a partir de la ganancia o la pérdida que supondría escoger una u otra.

   Por tanto la entrada X podría ser el estado o posición en el que se encuentra el dron en un determinado momento, la función objetivo f se encargaría de evaluar la mejor alternativa y la salida Y sería el siguiente movimiento que tendría que hacer el dron para rodear el obstáculo.

   d) Dada una colección de fotos de perros, posiblemente de distintas razas, establecer cuantas razas distintas hay representadas en la colección.

   En este caso se usaría *aprendizaje no supervisado* ya que buscaríamos patrones.

   La entrada X serán características de los perros que que nos permitan identificar su raza. A partir de estas características similares, la función objetivo f buscaría distinguir grupos en función de características similares (forma de las orejas, longitud de las patas, etc). El número de grupos (razas) diferentes encontrados sería nuestra salida Y.

2. ¿Cuáles de los siguientes problemas son más adecuados para una aproximación por aprendizaje y cuáles más adecuados para una aproximación por diseño? Justificar la decisión.

   a) Determinar si un vertebrado es mamífero, reptil, anfibio o pez.

   Podemos encontrar una serie de características concretas, como estructura ósea, forma de reproducción, hábitat,... por lo que este problema es adecuado para una *aproximación*

por diseño. Utilizaremos la distribución de probabilidad en estas características para clasificar un vertebrado sin la necesidad de un aprendizaje.

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

Este problema es adecuado para una *aproximación por aprendizaje*. No se pueden encontrar características concretas, por lo que sería más adecuado un aprendizaje a partir de datos de enfermedades anteriores, como número de contagios, duración de la enfermedad, gravedad, ... que nos permita determinar si llevar a cabo las vacunaciones o no.

c) Determinar perfiles de consumidor en una cadena de supermercados.

En los consumidores podemos encontrar una serie de características que nos permitan hacer una *aproximación por diseño*. Por ejemplo: estado civil, sexo, número de hijos, ciudad,... En función de estas características podemos realizar una clasificación de los clientes, sin la necesidad de un aprendizaje.

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

Este problema es adecuado para una *aproximación por aprendizaje*. Aunque cada estado anímico tenga unas características concretas, por ejemplo, ojos llorosos, ceño fruncido, posición de la boca,... y podría abordarse de esa forma, el estado anímico es más complejo y depende de muchos factores, por lo que creo que una función que aprenda a partir de imágenes con su estado de ánimo correspondiente sería más adecuada para hacer una predicción acertada.

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

Este problema es adecuado para una *aproximación por aprendizaje*, ya que no es un problema con características concretas. Podemos recoger datos que resulten de interés en este problema como cuál es la hora punta, número de peatones,...para aprender una función que nos permita determinar el ciclo óptimo de las luces del semáforo.

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.

Buscamos algunas características que nos resulten útiles para la clasificación de estas frutas. A partir de imágenes se puede deducir que la principal diferencia está en el color de la piel, en los mangos predominan los tonos naranjas y rojizos, las papayas son principalmente verdes y las guayabas tienen tonos amarillos. Puede tenerse en cuenta también la forma y el tamaño: los mangos y las guayabas son redondeados mientras que la papaya es alargada. Además los mangos y la papaya son más grandes que la guayaba.

Como con una sola de estas características no podemos definir si se trata de una fruta u otra, tenemos que tener en cuenta todas ellas. Se da una descripción de los datos que pueda ser usada por un computador: el vector de entrada será este conjunto de características:

$$X = \{x \in [0,1] \times [0,1] \times [0,1] \times [0,1] \times \mathbb{R}^+\}$$

donde los colores tomarán un valor entre 0 y 1 en función de la cantidad de ese color que se encuentra en la fruta, sumando entre los tres 1, siendo 1 el máximo. El cuarto elemento también será un valor entre 0 y 1 dependiendo de lo redondeada que sea la fruta, siendo 0

muy redonda y 1 muy alargada. El último elemento será el diámetro de la fruta, y por tanto será un valor real positivo.

A estos datos de entrada se les asociará una etiqueta (1,2,3) en función de si se trata de un mango una papaya o una guayaba, respectivamente. Por tanto, se tratará de un conjunto discreto: $Y = \{1, 2, 3\}$.

Por tanto, los datos de entrenamiento serán elementos de X asociados con la etiqueta que corresponda según la fruta de la que se trate $D = \{(x_i, y_i), x_i \in X, y_i \in Y\}$.

Para terminar, determinamos la función objetivo. A partir de los datos de entrenamiento, obtendremos una función que debería ser capaz de hacer una hipótesis sobre el tipo de fruta que es, es decir, devolver la etiqueta correspondiente, a partir de las características que hemos determinado antes. Por tanto, la definición de la función será la siguiente:

$$f : X \to Y$$

Considero que estamos ante un caso de etiquetas con ruido ya que a veces estas frutas son parecidas y por tanto es fácil confundirlas.

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.

   Consideramos la descomposición de las matrices de la forma $A = FD_1G^T$, $X = UD_2V^T$. Tenemos en cuenta que como U, V^T, F, G^T son ortogonales, $(P_1 \cdot P_2)^T = P_2^T \cdot P_1^T$ y que $D = D^T$.

$$A = X^T X$$
$$FD_1G^T = (UDV^T)^T UD_2V^T$$
$$FD_1G^T = VD_2^T U^T UD_2V^T$$
$$FD_1G^T = VD_2D_2V^T$$
$$FD_1G^T = VD_2^2V^T$$

   Por tanto, 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 $M \times 1$. 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 $\hat{z}$ 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 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,

$$cov(X) = \begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & ... & cov(x_1, x_N) \\ cov(x_2, x_1) & cov(x_2, x_2) & ... & cov(x_2, x_N) \\ ... & ... & ... & ... \\ cov(x_N, x_1) & cov(x_N, x_2) & ... & cov(x_N, x_N) \end{pmatrix} \tag{1.1}$$

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

a) $E1 = 11^T X$

$E1$ será una matriz en la que la cada componente de la columna 1 será la suma de los componentes de x_1, la columna 2 la suma de los componentes de x_2 y así sucesivamente. Consideramos que x_{MN} es la componente M del vector x_N. Los cálculos serían los siguientes:

$$\begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & ... & 1 \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & ... & x_{1N} \\ x_{21} & x_{22} & ... & x_{2N} \\ ... & ... & ... & ... \\ x_{M1} & x_{32} & ... & x_{MN} \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{pmatrix} \begin{pmatrix} x_{11} & x_{21} & ... & x_{N1} \\ x_{12} & x_{22} & ... & x_{N2} \\ ... & ... & ... & ... \\ x_{1M} & x_{2M} & ... & x_{NM} \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & ... & \sum_{i=1}^{M} x_{Ni} \\ \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & ... & \sum_{i=1}^{M} x_{Ni} \\ ... & ... & ... & ... \\ \sum_{i=1}^{M} x_1 & \sum_{i=1}^{M} x_2 & ... & \sum_{i=1}^{M} x_{Ni} \end{pmatrix}$$

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

En primer lugar, calculamos la componente $(X - \frac{1}{M} E1)$.

$$X - \frac{1}{M} E1 = \begin{pmatrix} x_{11} & x_{21} & ... & x_{N1} \\ x_{12} & x_{22} & ... & x_{N2} \\ ... & ... & ... & ... \\ x_{1M} & x_{2M} & ... & x_{NM} \end{pmatrix} - \frac{1}{M} \begin{pmatrix} \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & ... & \sum_{i=1}^{M} x_{Ni} \\ \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & ... & \sum_{i=1}^{M} x_{Ni} \\ ... & ... & ... & ... \\ \sum_{i=1}^{M} x_{1i} & \sum_{i=1}^{M} x_{2i} & ... & \sum_{i=1}^{M} x_{Ni} \end{pmatrix} =$$

$$\begin{pmatrix} x_{11} & x_{21} & ... & x_{N1} \\ x_{12} & x_{22} & ... & x_{N2} \\ ... & ... & ... & ... \\ x_{1M} & x_{2M} & ... & x_{NM} \end{pmatrix} - \begin{pmatrix} \frac{\sum_{i=1}^{M} x_{1i}}{M} & \frac{\sum_{i=1}^{M} x_{2i}}{M} & ... & \frac{\sum_{i=1}^{M} x_{Ni}}{M} \\ \frac{\sum_{i=1}^{M} x_{1i}}{M} & \frac{\sum_{i=1}^{M} x_{2i}}{M} & ... & \frac{\sum_{i=1}^{M} x_{Ni}}{M} \\ ... & ... & ... & ... \\ \frac{\sum_{i=1}^{M} x_{1i}}{M} & \frac{\sum_{i=1}^{M} x_{2i}}{M} & ... & \frac{\sum_{i=1}^{M} x_{Ni}}{M} \end{pmatrix}$$

$$\begin{pmatrix} x_{11} & x_{21} & ... & x_{N1} \\ x_{21} & x_{22} & ... & x_{N2} \\ ... & ... & ... & ... \\ x_{1M} & x_{2M} & ... & x_{NM} \end{pmatrix} - \begin{pmatrix} \bar{x_1} & \bar{x_2} & ... & \bar{x_N} \\ \bar{x_1} & \bar{x_2} & ... & \bar{x_N} \\ ... & ... & ... & ... \\ \bar{x_1} & \bar{x_2} & ... & \bar{x_N} \end{pmatrix} =$$

$$\begin{pmatrix} x_{11} - \bar{x_1} & x_{21} - \bar{x_2} & ... & x_{1n} - \bar{x_n} \\ x_{12} - \bar{x_1} & x_{22} - \bar{x_2} & ... & x_{N2} - \bar{x_N} \\ ... & ... & ... & ... \\ x_{1M} - \bar{x_1} & x_{2M} - \bar{x_2} & ... & x_{NM} - \bar{x_N} \end{pmatrix}$$

Sustituimos ahora en la expresión completa:

$$\begin{pmatrix} x_{11} - \bar{x_1} & x_{21} - \bar{x_2} & ... & x_{N1} - \bar{x_N} \\ x_{12} - \bar{x_1} & x_{22} - \bar{x_2} & ... & x_{N2} - \bar{x_N} \\ ... & ... & ... & ... \\ x_{1M} - \bar{x_1} & x_{2M} - \bar{x_2} & ... & x_{NM} - \bar{x_N} \end{pmatrix}^T \begin{pmatrix} x_{11} - \bar{x_1} & x_{21} - \bar{x_2} & ... & x_{N1} - \bar{x_N} \\ x_{12} - \bar{x_1} & x_{22} - \bar{x_2} & ... & x_{N2} - \bar{x_N} \\ ... & ... & ... & ... \\ x_{1M} - \bar{x_1} & x_{2M} - \bar{x_2} & ... & x_{NM} - \bar{x_N} \end{pmatrix} =$$

$$\begin{pmatrix} x_{11}-\bar{x_1} & x_{12}-\bar{x_1} & ... & x_{1M}-\bar{x_1} \\ x_{21}-\bar{x_2} & x_{22}-\bar{x_2} & ... & x_{2M}-\bar{x_2} \\ ... & ... & ... & ... \\ x_{N1}-\bar{x_n} & x_{N2}-\bar{x_N} & ... & x_{NM}-\bar{x_N} \end{pmatrix} \begin{pmatrix} x_{11}-\bar{x_1} & x_{21}-\bar{x_2} & ... & x_{N1}-\bar{x_N} \\ x_{12}-\bar{x_1} & x_{22}-\bar{x_2} & ... & x_{N2}-\bar{x_N} \\ ... & ... & ... & ... \\ x_{1M}-\bar{x_1} & x_{2M}-\bar{x_2} & ... & x_{NM}-\bar{x_N} \end{pmatrix} =$$

$$\begin{pmatrix} \sum_{i=1}^{M}(x_{1i}-\bar{x}_1)(x_{1i}-\bar{x}_1) & \sum_{i=1}^{M}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2) & ... & \sum_{i=1}^{M}(x_{1i}-\bar{x}_1)(x_{Ni}-\bar{x}_N) \\ \sum_{i=1}^{M}(x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1) & \sum_{i=1}^{M}(x_{2i}-\bar{x}_2)(x_{2i}-\bar{x}_2) & ... & \sum_{i=1}^{M}(x_{2i}-\bar{x}_2)(x_{Ni}-\bar{x}_N) \\ ... & ... & ... & ... \\ \sum_{i=1}^{M}(x_{Ni}-\bar{x}_N)(x_{1i}-\bar{x}_1) & \sum_{i=1}^{M}(x_{Ni}-\bar{x}_N)(x_{2i}-\bar{x}_2) & ... & \sum_{i=1}^{M}(x_{Ni}-\bar{x}_N)(x_{Ni}-\bar{x}_N) \end{pmatrix} =$$

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

Y por tanto, $E2 = M \cdot cov(X)$

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

   a) ¿Qué representa la matriz $\hat{H}$ en un modelo de regresión?

   Se trata de la matriz de proyección que nos da la predicción del modelo para un vector de entrada.

   $$\hat{y} = \hat{H}y$$

   b) Identifique la propiedad más relevante de dicha matriz en relación con regresión lineal.

   La propiedad es que la matriz es idempotente, esto es $\hat{H}^2 = \hat{H}$. Por esta propiedad, si intentamos predecir un valor ya precedido por el modelo, nos dará el mismo valor.

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 la adaptación de pesos siempre produce un movimiento de w en la dirección correcta para clasificar $x(t)$.

   Tenemos que comprobar que w_{new} predice mejor que w_{old}.

   $$w_{new}^T x = (w_{old} + yx)^T x$$
   $$w_{new}^T x = (w_{old}^T + (yx)^T)x$$
   $$w_{new}^T x = (w_{old}^T + x^T y^T)x$$
   $$w_{new}^T x = w_{old}^T x + x^T y^T x$$

   El dato está mal clasificado si $sign(w^Tx) \neq y$. Tenemos dos posibilidades: $w^Tx > 0$ e $y = -1$ o $w^Tx < 0$ e $y = 1$.

   En el primer caso,

   $$w_{new}^T x = w_{old}^T x - x^T x$$

donde $w_{old}^T x > 0$, por hipótesis y $x^T x > 0$. Por tanto, se le está restando a un número positivo otro positivo y w_{new} hace que la nueva predicción se acerque más a 0, y por tanto a cambiar de signo, que la anterior, estando un paso más cerca de la clasificación correcta. En el segundo caso se razona de manera análoga.

$$w_{new}^T x = w_{old}^T x + x^T x$$

donde $w_{old}^T x < 0$ y $x^T x > 0$, de forma que el número negativo se acerca más a 0.

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 muestre i.i.d de tamaño $N(x_1, ..., x_N)$. 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 qué sea verdad o falso respectivamente la expresión en su interior.

   b) 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

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

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.

   Para demostrar que se verifica la expresión tan sólo tenemos que derivar:

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

   Por tanto el gradiente es:

   $$\nabla E_{in}(w) = \frac{1}{N} \sum_{n=0}^{N} \frac{1}{1 + e^{-y_n w^T x_n}} \cdot -y_n x_n e^{-y_n w^T x_n} =$$

   $$-\frac{1}{N} \sum_{n=0}^{N} \frac{y_n x_n e^{-y_n w^T x_n}}{1 + e^{-y_n w^T x_n}} =$$

   $$-\frac{1}{N} \sum_{n=0}^{N} \frac{y_n x_n e^{-y_n w^T x_n}}{1 + e^{-y_n w^T x_n}} \cdot \frac{e^{y_n w^T x_n}}{e^{y_n w^T x_n}} =$$

   $$-\frac{1}{N} \sum_{n=0}^{N} \frac{y_n x_n}{e^{y_n w^T x_n} + 1}$$

Obtenemos así la primera igualdad. La segunda es inmediata usando el cambio:

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

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 $\nu = 1$.