Aprendizaje Automático (2017-2018)
Doble Grado en Ingeniería Informática y Matemáticas
Universidad de Granada

Cuestionario 1

Simón López Vico

23 de abril 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.
   No supervisado: La máquina se encargará de buscar patrones que distingan cada uno de los rasgos asociados a cada raza, como puede ser, el color de piel, la forma de los ojos...

   b) Clasificación automática de cartas por distrito postal.
   Supervisado: La máquina necesita información inicial sobre cuál es la manera de clasificar las distintas cartas por código postal, pues si no se le da esa información, clasificará las cartas en distintos patrones como ella decida.

   c) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un período de tiempo fijado.
   Para éste tengo dos respuestas:
   - **Supervisado:** Damos información de cuál es el funcionamiento normal de los índices de mercado, para así aprender qué condiciones se dan antes de una subida o una bajada de valores.
   - **No supervisado:** Dejamos que la máquina busque estructuras y patrones por su cuenta de los requisitos que se dan cuando una subida o una bajada de valores está al llegar.

   d) Aprender un algoritmo que permita a un robot rodear un obstáculo.
   Por refuerzo: la máquina intentará rodear el obstáculo como pueda, y nosotros nos encargaremos de indicarle si la conducta que está teniendo para realizar el trabajo es más o menos correcta.

2. ¿Cuáles 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 diseño: Estudiaremos los datos que diferencian cada uno de los vertebrados y estableceremos diferentes reglas para clasificarlos.

   b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.
   Aproximación por diseño: Podemos realizar una comparación con las anteriores campañas

de vacunación, estableciendo un modelo sobre ellas y generando una distribución de probabilidad para designar si se debe aplicar o no la vacuna.

c) Determinar si un correo electrónico es de propaganda o no.
Aproximación por aprendizaje: Establecemos distintas etiquetas a cada tipo de correo electrónico, mientras que el algoritmo busca una hipótesis que clasifique bien estos datos, procesando cada uno de los correos entrantes y marcándolos como spam o como fiables.

d) Determinar el estado de ánimo de una persona a partir de una foto de su cara.
Aproximación por aprendizaje: Para hacer una aproximación por diseño, el problema ha de estar bien definido, y este problema no lo está por lo que realizaremos una aproximación por aprendizaje.

e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.
Aproximación por aprendizaje: Al igual que el anterior, dicho problema no está bien definido, y mediante el aprendizaje habrá que denotar cuál es dicho ciclo óptimo.

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 $\mathcal{X}, \mathcal{Y}, \mathcal{D}$, 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.

   - $\mathcal{X}$: Dominio del problema, el cual contendrá (según las características elegidas por mi) valores que representarán el peso de la fruta, el volumen y el color de ésta, teniendo así un conjunto que contendrá vectores de tres dimensiones, $(x_1, x_2, x_3) \in \mathcal{X}$.

   - $\mathcal{Y}$: Conjunto que contendrá las etiquetas asignadas a cada fruta, por ejemplo, un 1 para el mango, un 2 para la papaya o un 3 para la guayaba, $\mathcal{Y} = \{1, 2, 3\}$.

   - $\mathcal{D} = \{(x_i, y_i / i = 1...N, x_i \in \mathcal{X}, y_i \in \mathcal{Y})\}$, conjunto de datos para el entrenamiento del algoritmo de aprendizaje, contendrá un vector x en la posición iésima que contendrá la información sobre la fruta etiquetada con el valor de y en la posición iésima.

   - $f : \mathcal{X} \to \mathcal{Y}, f \in \mathcal{H}$, función sobre la que debemos de establecer los parámetros para obtener el menor error durante el aprendizaje. Si los datos se diferencian bien y se juntan en tres distintos cúmulos de datos en $\mathbb{R}^3$, podremos separarlo con una función que genere dos planos, separando así el espacio para cada conjunto de datos.

   Respecto al ruido en las etiquetas, hay que tener en cuenta que las tres frutas son parecidas en varios aspectos, por lo que si queremos diferenciar bien cada una de las frutas tendremos que escoger características de ellas fácilmente medibles y que generen una diferencia mayor entre las tres.

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 estas nuevas matrices que no tiene X. ¿Qué valor representa la suma de la diagonal principal de cada una de las matrices producto?

Para empezar, notemos que $U \in \mathcal{M}_{N\times d}$ es una matriz con columnas ortogonales, $D \in \mathcal{M}_{d\times d}$ es una matriz diagonal y $V \in \mathcal{M}_{d\times d}$ es una matriz ortogonal. Por tanto, $U^TU = Id$, $D^T = D$ y $V^TV = Id$.

Dicho esto, calculemos X^TX:

$$X^TX = (UDV^T)^T(UDV^T) = VDU^TUDV^T = VDDV^T = VD^2V^T$$

donde D^2 será igual a la matriz con el cuadrado de los elementos de la diagonal de D.

Ahora hacemos lo mismo con XX^T:

$$XX^T = (UDV^T)(UDV^T)^T = UDV^TVDU^T = UDDU^T = UD^2U^T$$

con D^2 igual que en el anterior.

Por definición, sabemos que X^TX y XX^T son simétricas, por tanto las dos son diagonalizables con autovalores reales y no negativos. Además, los valores singulares de X serán las raíces cuadradas de los autovalores de X^TX y XX^T.

Por tanto, la suma de la diagonal principal de las matrices producto será igual a la traza de X, es decir, $traza(X) = traza(X^TX) = traza(XX^T) =$*(suma de valores propios)*.

5. Sean $\mathbf{x}$ e $\mathbf{y}$ dos vectores de características de dimensión $M \times 1$. La expresión

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

define la covarianza entre dichos vectores, donde $\bar{z}$ representa el valor medio de los elementos de $\mathbf{z}$. Considere ahora una matriz X cuyas columnas representan vectores de características. La matriz de covarianzas asociada a la matriz $\text{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

$$\text{cov(X)} = \begin{pmatrix} \text{cov}(\mathbf{x}_1, \mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_1, \mathbf{x}_N) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_2, \mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(\mathbf{x}_N, \mathbf{x}_1) & \text{cov}(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}$$

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

a) $E1 = 11^TX$

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

Para empezar, calculemos $E1 = 11^TX$, donde entendemos como 11^T la matriz con solo unos:

$$E1 = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & & \vdots \\ x_{M1} & \cdots & x_{MM} \end{pmatrix} = \begin{pmatrix} x_{11} + x_{21} + ... + x_{M1} & \cdots & x_{1M} + x_{2M} + ... + x_{MM} \\ \vdots & & \vdots \\ x_{11} + x_{21} + ... + x_{M1} & \cdots & x_{1M} + x_{2M} + ... + x_{MM} \end{pmatrix}$$

$$E1 = \begin{pmatrix} \sum_{i=1}^{M} x_{i1} & \cdots & \sum_{i=1}^{M} x_{iM} \\ \vdots & & \vdots \\ \sum_{i=1}^{M} x_{i1} & \cdots & \sum_{i=1}^{M} x_{iM} \end{pmatrix}$$

A continuación, calculemos $E2 = (X - \frac{1}{M}E1)^T(X - \frac{1}{M}E1)$. Para ello, empecemos calculando $(X - \frac{1}{M}E1)$.

$$(X - \frac{1}{M}E1) = \begin{pmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & & \vdots \\ x_{M1} & \cdots & x_{MM} \end{pmatrix} - \begin{pmatrix} \frac{1}{M}\sum_{i=1}^{M} x_{i1} & \cdots & \frac{1}{M}\sum_{i=1}^{M} x_{iM} \\ \vdots & & \vdots \\ \frac{1}{M}\sum_{i=1}^{M} x_{i1} & \cdots & \frac{1}{M}\sum_{i=1}^{M} x_{iM} \end{pmatrix}$$

donde todos los elementos de cada fila en la segunda matriz son la media de la columna iésima de la matriz X, es decir:

$$(X - \frac{1}{M}E1) = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1M} - \bar{x}_M \\ \vdots & & \vdots \\ x_{M1} - \bar{x}_1 & \cdots & x_{MM} - \bar{x}_M \end{pmatrix}$$

Por tanto, continuamos para calcular $E2$ por completo:

$$E2 = (X - \frac{1}{M}E1)^T(X - \frac{1}{M}E1) = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1M} - \bar{x}_M \\ \vdots & & \vdots \\ x_{M1} - \bar{x}_1 & \cdots & x_{MM} - \bar{x}_M \end{pmatrix}^T \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1M} - \bar{x}_M \\ \vdots & & \vdots \\ x_{M1} - \bar{x}_1 & \cdots & x_{MM} - \bar{x}_M \end{pmatrix} =$$

$$= \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{M1} - \bar{x}_1 \\ \vdots & & \vdots \\ x_{1M} - \bar{x}_M & \cdots & x_{MM} - \bar{x}_M \end{pmatrix} \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1M} - \bar{x}_M \\ \vdots & & \vdots \\ x_{M1} - \bar{x}_1 & \cdots & x_{MM} - \bar{x}_M \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^{M}(x_{i1} - \bar{x}_1)(x_{i1} - \bar{x}_1) & \cdots & \sum_{i=1}^{M}(x_{i1} - \bar{x}_1)(x_{iM} - \bar{x}_M) \\ \vdots & & \vdots \\ \sum_{i=1}^{M}(x_{iM} - \bar{x}_M)(x_{i1} - \bar{x}_1) & \cdots & \sum_{i=1}^{M}(x_{iM} - \bar{x}_M)(x_{iM} - \bar{x}_M) \end{pmatrix} = M \begin{pmatrix} cov(x_1, x_1) & \cdots & cov(x_1, x_M) \\ \vdots & & \vdots \\ cov(x_M, x_1) & \cdots & cov(x_M, x_M) \end{pmatrix}$$

Por tanto, podremos concluir que $E2 = M * cov(X)$.

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

   a) Mostrar que H es simétrica.
   Si H es simétrica significa que $H^T = H$; comprobémoslo:

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

   Por tanto, H es simétrica.

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

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

   Demostrando así que H es idempotente.

c) ¿Qué representa la matriz H en un modelo de regresión?
Podemos ver que la matriz $H = X(X^TX)^{-1}X^T = XX^\dagger$, donde $X^\dagger$ es la pseudo-inversa de X; por tanto, la matriz puede definir los niveles de aprendizaje usando solamente los valores en X, al igual que la pseudo-inversa.

7. La regla de adaptación de los pesos del Perceptron ($\mathbf{w}_{new} = \mathbf{w}_old + y\mathbf{x}$) tiene la interesante propiedad de que los mueve en la dirección adecuada para clasificar $\mathbf{x}$ de forma correcta. Suponga el vector de pesos $\mathbf{w}$ de un modelo y un dato $\mathbf{x}(t)$ 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 $\mathbf{x}(t)$.
La prueba es simple, solo tendremos que tener en cuenta que la regla de adaptación de pesos del Perceptron modificará el valor de w siempre que haya un mínimo error, y por tanto, aunque los datos no estén dispuestos en su sitio, con una cantidad de iteraciones necesaria acabarán moviéndose en la dirección correcta para clasificar bien $\mathbf{x}(t)$.

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.
   Entre otras opciones, podremos usar $P(Y) = (1 - h(x))(1 - Y) + h(x)Y$.

   b) Considere una muestra N v.a. independientes. Escribir la función de Máxima Verosimilitud para dicha muestra.
   La función de verosimilitud será el producto de las N variables independientes sobre la aplicación $P(Y)$, es decir, $\mathcal{L}(Y|w_1, ..., w_N) = \prod_{i=1}^{N} P(Y_i) = \prod_{i=1}^{N}((1 - h(x))(1 - Y_i) + h(x)Y_i)$.

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

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

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

   [...]

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

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

   [...]

9. Mostrar que en regresión logística se verifica:

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

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

   [...]

10. Definamos el error en un punto (x_n, y_n) por

$$\mathbf{e}_n(\mathbf{w}) = max(0, -y_n \mathbf{w}^T \mathbf{x}_n)$$

Argumentar si con esta función de error el algoritmo PLA puede interpretarse como SGD sobre $\mathbf{e}_n$ con tasa de aprendizaje $\eta = 1$.

[...]