Aprendizaje automático
Cuestionario 1

Alejandro García Montoro
agarciamontoro@correo.ugr.es

29 de marzo de 2016

1. Ejercicios

Ejercicio 1. *Identificar, para cada una de las siguientes tareas, qué tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los datos de aprendizaje que deberíamos usar. Si una tarea se ajusta a más de un tipo, explicar cómo y describir los datos para cada tipo.*

- *Categorizar un grupo de animales vertebrados en pajaros, mamíferos, reptiles, aves y anfibios.*
- *Clasificación automática de cartas por distrito postal.*
- *Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.*

Ejercicio 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.*

- *Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.*
- *Determinar los ingresos medios de una persona a partir de sus datos de nivel de educación, edad, experiencia y estatus social.*
- *Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.*

Ejercicio 3. *Construir un problema de aprendizaje desde datos para un problema de selección de fruta en una explotación agraria (ver transparencias de clase). Identificar y describir cada uno de sus elementos formales. Justificar las decisiones.*

Ejercicio 4. *Suponga un modelo PLA y un dato $x(t)$ mal clasificado respecto de dicho modelo. Probar que la regla de adaptación de pesos del PLA es un movimiento en la dirección correcta para clasificar bien $x(t)$.*

Solución. Sea $(x(t), y(t))$, con $y(t) \in \{-1, +1\}$ la muestra mal clasificada respecto del modelo PLA. La siguiente iteración del algoritmo nos dará un vector de pesos

$$w(t+1) = w(t) + y(t)x(t)$$

Como el dato $x(t)$ está mal etiquetado, tenemos que $sign(w^T(t)x(t)) \neq y(t)$, luego podemos concluir que

$$y(t)sign(w^T(t)x(t)) < 0$$

De hecho, como $sign(x) \in \{-1, 1\} \forall x \in \mathbb{R}$, tenemos que

$$y(t)sign(w^T(t)x(t)) = -1$$

Por otro lado, tenemos que

$$\begin{aligned} y(t)w^T(t+1)x(t) &= y(t)(w^T(t) + y(t)x(t))x(t) = \\ &= (y(t)w^T(t) + x(t))x(t) = \\ &= y(t)w^T(t)x(t) + x^2(t) > y(t)w^T(t)x(t) \end{aligned}$$

donde hemos usado que $y(t)^2 = 1$, ya que $y(t) \in \{-1, +1\}$ y que

Tomando signos, tenemos la siguiente desigualdad:

$$sign(y(t)w^T(t+1)x(t)) > sign(y(t)w^T(t)x(t)) = sign(-1) = -1$$

Como la anterior es una desigualdad estricta y, de nuevo, $sign(x) \in \{-1, 1\} \forall x \in \mathbb{R}$, podemos concluir que $sign(y(t)w^T(t+1)x(t)) = 1$, luego necesariamente $y(t)$ y $w^T(t+1)x(t)$ tienen el mismo signo. Como $y(t) = sign(y(t))$, concluimos que

$$y(t) = sign(w^T(t+1)x(t))$$

es decir, la muestra $(x(t), y(t))$ está ahora bien etiquetada.

Ejercicio 5. *Considere el enunciado del ejercicio 2 de la sección FACTIBILIDAD DEL APRENDIZAJE de la relación apoyo.*

- *Si $p = 0{,}9$, ¿cuál es la probabilidad de que S produzca una hipótesis mejor que C?*
- *¿Existe un valor de p para el cual es más probable que C produzca una hipótesis mejor que S?*

Ejercicio 6. *La desigualdad de Hoeffding modificada nos da una forma de caracterizar el error de generalización con una cota probabilística*

$$\mathbb{P}[|E_{out}(g) - E_{in}(g)| > \varepsilon] \leq 2Me^{-2N^2\varepsilon}$$

para cualquier $\varepsilon > 0$*. Si fijamos* $\varepsilon = 0{,}05$ *y queremos que la cota probabilística* $2Me^{-2N^2\varepsilon}$ *sea como máximo* 0,03*, ¿cuál será el valor más pequeño de* N *que verifique estas condiciones si* $M = 1$*? Repetir para* $M = 10$ *y para* $M = 100$*.*

Solución. Si imponemos que la cota probabilística sea menor o igual que un valor k, basta despejar N de la desigualdad

$$2Me^{-2N^2\varepsilon} \geq k$$

y estudiar lo que se nos pide. Tenemos entonces:

$$\begin{aligned} 2Me^{-2N^2\varepsilon} &\leq k \\ e^{-2N^2\varepsilon} &\leq \frac{k}{2M} \\ -2N^2\varepsilon &\leq ln(\frac{k}{2M}) \\ N^2 &\geq \frac{1}{-2\varepsilon} ln(\frac{k}{2M}) \\ N &\geq \sqrt{\frac{1}{-2\varepsilon} ln(\frac{k}{2M})} \end{aligned}$$

Es decir, la cota probabilística será menor o igual que k si y sólo si $N \geq \sqrt{\frac{1}{-2\varepsilon} ln(\frac{k}{2M})}$. Como $N \in \mathbb{N}$, tenemos que el menor N que cumple la condición para un M dado es exactamente

$$N_M = \left\lceil \sqrt{\frac{1}{-2\varepsilon} ln\left(\frac{k}{2M}\right)} \right\rceil \tag{1}$$

donde $\lceil x \rceil$ es el menor entero mayor o igual que x.

Ahora basta tomar $k = 0{,}03$, $\varepsilon = 0{,}05$ y, para $M \in \{1, 10, 100\}$, calcular la expresión obtenida en 1, lo que nos da los siguientes valores:

$$\begin{aligned} N_1 &= 7 \\ N_{10} &= 9 \\ N_{100} &= 10 \end{aligned}$$

Ejercicio 7. *Consideremos el modelo de aprendizaje «M-intervalos »donde* $h\colon \mathbb{R} \to \{1, +1\}$*, y* $h(x) = +1$ *si el punto está dentro de cualquiera de* m *intervalos arbitrariamente elegidos y 1 en otro caso. ¿Cuál es el más pequeño punto de ruptura para este conjunto de hipótesis?*

Solución. Es claro que M intervalos pueden separar cualquier muestra de $2M$ puntos. Imaginemos, por ejemplo, un tal conjunto de tamaño $2M$ en el que las muestras con distintas etiquetas se van alternando:

$$+1 \quad -1 \quad +1 \quad -1 \quad \cdots \quad +1 \quad -1$$

Evidentemente, hay M puntos etiquetados con $+1$, luego basta con *rodear* esos puntos con los M intervalos de los que disponemos para poder separar completamente la muestra.

De hecho, esta dicotomía es la más *difícil* de implementar, en el sentido de que necesitamos todos los intervalos disponibles para separarla. Supongamos ahora que intercambiamos dos puntos con etiquetas diferentes de la disposición anterior. En ese caso necesitaríamos sólo $M-1$ intervalos para implementar la dicotomía. Cualquier otra disposición resulta en un número menor de intervalos necesarios, ya que agrupa puntos con etiquetas iguales bajo un mismo intervalo.

Concluimos así que $m_{\mathcal{H}}(2M) = 2^{2M}$.

Para ver que $2M+1$ es un punto de ruptura basta encontrar una muestra de ese tamaño de manera que $\mathcal{H}$ no sea capaz de conseguir etiquetarla bien.

Consideramos de nuevo la misma muestra anterior, esta vez de tamaño $2M+1$, para lo que añadimos un $+1$ al final; es decir, tenemos $M+1$ puntos etiquetados como $+1$ y una dicotomía como la siguiente:

$$+1 \quad -1 \quad +1 \quad -1 \quad \cdots \quad +1 \quad -1 \quad +1$$

Si nuestro objetivo es separar la muestra, debemos empezar por la izquierda y, cada vez que encontremos un $+1$, usar un intervalo para etiquetarlo bien. Esto no se puede hacer de otra manera, ya que los $+1$ tienen que estar dentro de un intervalo y los -1 fuera. Si seguimos hacia delante, habremos gastado los intervalos al *encerrar* al M-ésimo $+1$, luego todo lo que haya a su derecha será etiquetado como -1, ya que se queda fuera de un intervalo.

El último $+1$, por tanto, será etiquetado incorrectamente. Tenemos así una muestra de $2M+1$ puntos con una dicotomía que $\mathcal{H}$ no puede implementar; es decir, $k = 2M+1$ es un punto de ruptura.

Entonces, como $m_{\mathcal{H}}(2M+1) < 2^{2M}$ y $m_{\mathcal{H}}(2M) = 2^{2M}$, podemos ya afirmar que $k = 2M+1$ es el más pequeño punto de ruptura para este conjunto de hipótesis.

Ejercicio 8. *Suponga un conjunto de k^* puntos $x_1, x_2, \ldots, x_{k^*}$ sobre los cuales la clase $\mathcal{H}$ implementa $< 2^{k^*}$ dicotomías. ¿Cuáles de las siguientes afirmaciones son correctas?*

- *k^* es un punto de ruptura.*
- *k^* no es un punto de ruptura.*
- *Todos los puntos de ruptura son estrictamente mayores que k^*.*

- *Todos los puntos de ruptura son menores o iguales a k^**
- *No conocemos nada acerca del punto de ruptura.*

Solución. w

- Si $\mathcal{H}$ implementa menos de 2^{k^*} dicotomías, tenemos que $m_{\mathcal{H}}(k^*) < 2^{k^*}$. Esta es exactamente la definición de punto de ruptura, luego podemos afirmar que k^* lo es.
- No podemos conocer nada acerca de cualquier otro punto de ruptura k. Supongamos el caso anterior de M intervalos y consideremos $k^* = 2M + 2$. Hemos probado en el ejercicio anterior que $k = 2M + 1$ es un punto de ruptura, luego existen puntos de ruptura menores o iguales que k^*. Pero sabemos que dado k^*, cualquier $k > k^*$ es un punto de ruptura —si $\mathcal{H}$ no puede separar k^* puntos, evidentemente tampoco puede separar $k^* + 1$—, luego existen también puntos de ruptura mayores estrictos que k^*.

Ejercicio 9. *Para todo conjunto de k^* puntos, $\mathcal{H}$ implementa $< 2^{k^*}$ dicotomías. ¿Cuáles de las siguientes afirmaciones son correctas?*

- *k^* es un punto de ruptura.*
- *k^* no es un punto de ruptura.*
- *Todos los $k \geq k^*$ son puntos de ruptura.*
- *Todos los $k < k^*$ son puntos de ruptura.*
- *No conocemos nada acerca del punto de ruptura.*

Ejercicio 10. *Si queremos mostrar que k^* es un punto de ruptura, ¿cuáles de las siguientes afirmaciones nos servirían para ello?:*

- *Mostrar que existe un conjunto de k^* puntos $x_1, x_2, \ldots, x_{k^*}$ que $\mathcal{H}$ puede separar* —shatter—.
- *Mostrar que $\mathcal{H}$ puede separar cualquier conjunto de k^* puntos.*
- *Mostrar un conjunto de k^* puntos $x_1, x_2, \ldots, x_{k^*}$ que $\mathcal{H}$ no puede separar.*
- *Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de k^* puntos.*
- *Mostrar que $m_{\mathcal{H}}(k) = 2^{k^*}$*

Ejercicio 11. *Para un conjunto $\mathcal{H}$ con $d_{VC} = 10$, ¿qué tamaño muestral se necesita —según la cota de generalización— para tener un 95 % de confianza de que el error de generalización sea como mucho* 0,05*?*

Ejercicio 12. *Consideremos un escenario de aprendizaje simple. Supongamos que la dimensión de entrada es uno. Supongamos que la variable de entrada x está uniformemente distribuida en el intervalo $[-1, 1]$ y el conjunto de datos consiste en 2 puntos $\{x_1, x_2\}$ y que la función objetivo es $f(x) = x^2$. Por tanto el conjunto de datos completo es $\mathcal{D} = \{(x_1, x_1^2), (x_2, x_2^2)\}$. El algoritmo de aprendizaje devuelve la línea que ajusta estos dos puntos como g; es decir, $\mathcal{H}$ consiste en funciones de la forma $h(x) = ax + b$.*

- *Dar una expresión analítica para la función promedio $\bar{g}(x)$.*
- *Calcular analiticamente los valores de E_{out},* bias *y* var.

2. Bonus

Bonus 1. *Considere el enunciado del ejercicio 2 de la sección ERROR Y RUIDO de la relación de apoyo.*

- *Si su algoritmo busca la hipótesis h que minimiza la suma de los valores absolutos de los errores de la muestra,*

$$E_{in}(h) = \sum_{n=1}^{N} |h - y_n|$$

  entonces mostrar que la estimación será la mediana de la muestra, h_{med} —cualquier valor que deje la mitad de la muestra a su derecha y la mitad a su izquierda—.

- *Suponga que y_N es modificado como $y_N + \varepsilon$, donde $\varepsilon \to \infty$. Obviamente el valor de y_N se convierte en un punto muy alejado de su valor original. ¿Cómo afecta esto a los estimadores dados por h_{mean} y h_{med}?*

Bonus 2. *Considere el ejercicio 12.*

- *Describir un experimento que podamos ejecutar para determinar —numéricamente— $\bar{g}(x)$, E_{out},* bias *y* var.
- *Ejecutar el experimento y dar los resultados. Comparar E_{out} con bias + var. Dibujar en unos mismos ejes $\bar{g}(x)$, E_{out} y $f(x)$.*