Aprendizaje Automático: Cuestionario 1

Anabel Gómez Ríos

3 de abril de 2016

1. Cuestiones

Pregunta 1. Identificar, para cada una de las siguiente tareas, qué tipo de aprendizaje automático 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.

a) Categorizar un grupo de animales vertebrados en pájaros, mamíferos, reptiles, aves y anfibios.

b) Clasificación automática de cartas por distrito postal.

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

a) Con los siguientes datos de aprendizaje: (tamaño, número de patas, tipo de piel, pelo, alas, tipo) lo adecuado sería un aprendizaje supervisado, ya que tendríamos un conjunto de entrenamiento del que conocemos una serie de características y a qué tipo de animales pertenecen para cada uno de los datos en ese conjunto. A partir de aquí podemos aprender y si tenemos un nuevo grupo de animales del que sólo conocemos las características, obtener el tipo de animal que es.

b) Si nos llega una carta lo que tenemos que hacer es leer de forma digital el código postal, de forma que lo obtengamos como una foto y el problema se transforma en reconocimiento de dígitos, que es algo parecido a lo que hemos estado haciendo en las prácticas, necesitaríamos también un conjunto de prueba lo suficientemente grande en el que para cada dígito

tengamos el número que representa, de forma que si nos llega uno nuevo podamos (en base por ejemplo a intensidad y simetría, como hacíamos en la prácticas) distinguir qué número es. Una vez el sistema sabe qué números hay, sólo tiene que distinguir las cartas agrupando aquellas que tengan el mismo código postal, y tendríamos por tanto un aprendizaje supervisado.

c) Para este problema yo creo que sería más adecuado un aprendizaje por refuerzo, ya que la subida o bajada de valores en un periodo de tiempo no es algo seguro dadas unas determinadas características, es decir, nuestro conjunto de prueba serían las características que habría que tener en cuenta para cada valor junto con una probabilidad de que ese valor suba en el mercado: (características, subida, probabilidad de subida) o bien (características, bajada, probabilidad de bajada). Cuando en el conjunto de entrenamiento tenemos posibles salidas pero no una segura, el aprendizaje es por refuerzo.

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

b) Determinar los ingresos medios de una persona a partir de sus datos de nivel de educación, edad, experiencia y estatus social.

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

a) Creo que este problema sería más adecuado para diseño, ya que habría que informarse bien con alguien que supiera de los parámetros que influyen a la hora de cambiar las luces de los semáforos (que haya otros semáforos alrededor o que no, que sea entrada o salida de rotondas, etc.) y para poder abordarlo por aprendizaje haría falta tener una gran cantidad de datos de prueba y aun así no tiene por qué haber dos cruces iguales ya que con la mínima variación puede cambiar el momento de cambiar las luces, y es algo delicado como para equivocarse, aunque sea en un porcentaje mínimo de las veces.

b) Este sí podría ser bueno para tratarlo por aprendizaje: si tenemos un conjunto de características y un conjunto de datos para aprender en base a estas características tenemos nuestro conjunto de entrenamiento y podemos dar una buena aproximación de los ingresos si nos llegan nuevos datos que clasificar (otro conjunto de características de las que no sabemos los ingresos).

c) Este también se puede hacer por aprendizaje si por ejemplo tenemos datos de entrenamiento que son otras enfermedades y a partir de qué momento se ha aplicado una campaña de vacunación: por ejemplo a partir de qué número de afectados o la gravedad de la enfermedad. Si aparece una nueva enfermedad podemos fácilmente sacar estos datos y elegir en base a lo que hemos aprendido.

Pregunta 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.

Cada dato x serán las características de la fruta específica que nos interesan para cada pieza de fruta: $x = (color, tamaño, textura, ...)$ y por tanto el conjunto X será el conjunto de todos los posibles x y habrá tantos elementos como piezas de la misma fruta tengamos. El conjunto Y de etiquetas que tenemos será discreto: $Y = \{-1, +1\}$ y la función $f : X \to Y$ será la que para cada pieza de fruta nos dice si debemos clasificarla para la venta $(+1)$ o no (-1). El conjunto D que tiene los ejemplos de prueba será de la forma $(x_1, y_1), ..., (x_N, y_N)$ donde N es el número de ejemplos e $y_i = f(x_i)$.

Pregunta 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 $x(t)$.

Vamos a seguir el esquema del ejercicio 3 del apartado PLA de la relación de problemas, que como aparece también en el libro de la bibliografía [1] nos puede servir de ayuda. Recordemos que según nos dice el enunciado $x(t)$ está mal clasificado en este momento, es decir, $w(t)$ está clasificando mal $x(t)$. Recordemos también que de una iteración a otra, w se actualiza de la siguiente forma: $w(t + 1) = w(t) + y(t)x(t)$, que es lo que queremos probar

que va en la dirección correcta. En concreto dicho ejercicio pide que se pruebe lo siguiente:

a) $y(t)w^T(t)x(t) < 0$. Como $x(t)$ está mal clasificado por $w(t)$, entonces el signo de $w^T(t)x(t)$ tiene el signo contrario a $y(t)$, que es la verdadera etiqueta de $x(t)$, por lo que la multiplicación $y(t)w^T(t)x(t)$ tiene que ser negativa.
Tengamos en cuenta también que si $w(t)$ llega a clasificar bien a $x(t)$, entonces ambos tendrían el mismo signo y la multiplicación sería positiva. Por lo tanto, para que vaya en la dirección correcta lo que tenemos que hacer es irnos acercando con ese producto cada vez más a un número positivo.

b) $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$. Ver esto es muy fácil sin más que utilizar la regla de adaptación $w(t+1) = w(t) + y(t)x(t)$. La trasponemos para poder sustituirla en la expresión $y(t)w^T(t+1)x(t)$ y sustituimos. Entonces: $w^T(t+1) = x^T(t)y(t)$ ($y(t)$ es un número escalar) y sustituyendo, $y(t)w^T(t+1)x(t) = y(t)(w^T(t) + x^T(t)y(t))x(t) = y(t)w^T(t)x(t) + y(t)^2x^T(t)x(t)$. Ahora, como $y(t)^2$ es siempre positivo y $x^T(t)x(t)$ es la norma al cuadrado de $x(t)$ que es por definición positiva, tenemos que $y(t)^2x^T(t)x(t)$ es un término positivo que está sumando, luego $y(t)W^T(t)x(t) + y(t)^2x^T(t)x(t) > y(t)w^T(t)x(t)$, con lo que demostramos lo que queríamos.

Nos queda ya probar por tanto que en efecto la regla de adaptación lleva w en la dirección correcta. Como hemos comentado, ahora mismo $w(t)$ está clasificando mal $x(t)$. Por el apartado b), $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$, que por el apartado a) es negativo. Por lo tanto, como $y(t)w^T(t+1)x(t)$ es estrictamente mayor que $y(t)w^T(t)x(t)$ se está quedando más cerca de ser un número positivo que $y(t)w^T(t)x(t)$, que es la iteración anterior, y como hemos dicho antes, esto se queda más cerca de clasificar bien $x(t)$, con lo que w se está moviendo en la dirección correcta, como queríamos probar.

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

a) Si $p = 0,9$, ¿cuál es la probabilidad de que S produzca una hipótesis mejor que C?

b) ¿Existe un valor de p para el cual es más probable que C produzca una hipótesis mejor que S?

En dicho ejercicio a partir del apartado b) se da una hipótesis adicional que vamos a utilizar en este ejercicio, que todos los ejemplos en D (que son 25) tienen $y_n = +1$. También nos dice que para el caso probabilístico supongamos que hay una distribución de probabilidad sobre X y que $P[f(x) = +1] = p$

a) Tenemos que $P[f(x) = +1] = 0{,}9$. Dado que todos los puntos en D tienen etiqueta positiva, S va a elegir la hipótesis h_1, que es la función constante igual a 1. Vamos a ver cuál es la distribución de probabilidad que hay sobre X. Sabemos que la probabilidad de que cada $x \in X$ sea +1 o -1 es independiente para cada x y además cada etiqueta puede tomar sólo dos valores, +1 ó -1 (lo que se puede tomar como éxito/fracaso), por lo que estamos ante una distribución binomial, que es la que nos dice la probabilidad de que de n sucesos x de ellos sean un éxito (en nuestro caso, +1), donde un éxito tiene una probabilidad p de suceder. La función de probabilidad de la binomial es $q(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$, donde x es el número de éxitos. La probabilidad entonces de que la hipótesis de S sea mejor que la de C depende del número de datos fuera de la muestra. En concreto, para que esto sea así lo que tendría que pasar es que más de la mitad de los datos fuera de la muestra tengan etiqueta positiva. De hecho, basta con que uno más de la mitad tengan etiqueta positiva. Supongamos que fuera de la muestra hay un dato. Como la probabilidad de que ese dato sea 1 es 0.9, la probabilidad de que la hipótesis de S sea mejor es 0.9. Si hay, por ejemplo, otros 25 datos fuera de la muestra, $x = E[25/2] + 1 = 13$ como mínimo, donde $E[]$ es la función parte entera. Por lo tanto, nos sirve que x sea 13, 14, 15,... hasta el total, 25, puesto que en todos estos casos S producirá mejor hipótesis que C al haber más puntos etiquetados con +1 que con -1. Por lo tanto la probabilidad de que esto ocurra es $\sum_{i=13}^{25} q(i)$. Esta cuenta la he hecho haciendo una pequeña función en R que calcula q y el resultado es: $\sum_{i=13}^{25} q(i) = 0{,}9999998$. Por lo tanto tenemos que con 25 datos fuera de la muestra, la probabilidad sube a 0.9999998. Según subimos el número de datos, la probabilidad va aumentando hasta 1 (de hecho, en R para 50 datos ya sale probabilidad 1 por el redondeo que tiene). Esto tiene sentido porque cuando aumentamos el número de datos aumentamos no sólo la suma de las q

que estamos haciendo, si no el número de 1's que tiene que haber en la muestra para que la hipótesis de S sea mejor. Es decir, por ejemplo para 25 estábamos sumando 13 probabilidades, para 50 estaríamos sumando 24 (todas positivas, luego a mayor número de datos, más se acerca la probabilidad a 1), pero además para 25 estábamos sumando la probabilidad de que hubiera 13, 14, 15,..., 25 1's, mientras que para 50 tenemos que sumar la probabilidad de que haya al menos 26,27,...,50, que es mayor cuando mayor es el número de 1's puesto que los 1 tienen probabilidad 0.9 de aparecer y los -1 tienen probabilidad 0.1.

Es decir, de todo esto se desprende que a mayor número de datos fuera de la muestra (y hemos visto que ese número no tiene que ser muy grande) la probabilidad de que la hipótesis de S sea mayor que la de C es 1.

b) Para que C produzca una mejor hipótesis que la de S hace falta que la probabilidad que acabamos de calcular sea menor que 0.5. Si $p > 0{,}5$ sigue siendo más probable que la hipótesis de S sea mejor. Para $p = 0{,}5$ la probabilidad anterior vale exactamente 0.5, con lo que no podemos decir tampoco que sea más probable que la mejor sea la de C. Sin embargo para $p = 0{,}4$ esta probabilidad ya vale menos que 0.5, por ejemplo vale 0.1537678 para 25 datos , con lo que para $p \leq 0{,}4$ es más probable que C produzca mejor hipótesis que S.

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

$$P[|E_{out}(g) - E_{in}(g)| > \epsilon] \leqslant 2Me^{-2N\epsilon^2} \tag{1}$$

para cualquier $\epsilon > 0$. Si fijamos $\epsilon = 0,05$ y queremos que la cota probabilística $2Me^{-2N\epsilon^2}$ 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 $M = 100$.

Queremos que $2Me^{-2N\epsilon^2} \leq 0{,}03$. Despejamos de esta expresión N.

Para $M = 1$, $2e^{-2N0{,}05^2} \leq 0{,}03 \Rightarrow e^{-2N0{,}05^2} \leq 0{,}015 \Rightarrow -2N0{,}05^2 \leq ln(0{,}015) = -4{,}199705$. Si multiplicamos por -1 y despejamos N, $N \geq$

$\frac{4,199705}{2*0,05^2} = 839,941$. Como N tiene que ser entero, entonces el N más pequeño es 840.

Para $M = 10$, $e^{-2N0,05^2} \leq 0,0015 \Rightarrow -2N0,05^2 \leq ln(0,0015) = -6,50229$, luego $N \geq \frac{6,50229}{2*0,05^2} = 1300,458$ y por tanto el valor más pequeño de N es 1301.

Para $M = 100$, $e^{-2N0,05^2} \leq 0,00015 \Rightarrow -2N0,05^2 \leq ln(0,0015) = -8,804875$, luego $N \geq \frac{8,804875}{2*0,05^2} = 1760,975$ y por tanto el valor más pequeño de N es 1761.

Pregunta 7. Consideremos el modelo de aprendizaje "M-intervalos" donde $h : \mathbb{R} \to -1, +1$ 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?

Como h sale de $\mathbb{R}$, los puntos tienen que estar en una recta, uno detrás de otro, y sólo existe esta configuración de puntos. Por lo tanto, para encontrar un punto k de ruptura nos basta con encontrar un conjunto de k puntos para el que no se puedan hacer todas las dicotomías para los intervalos. Si el número de puntos es menor o igual que $2m$ es fácil ver que se pueden implementar todas las dicotomías, puesto que tenemos m intervalos disponibles: si hay dos o más puntos seguidos con la etiqueta +1 se meten en el mismo intervalo y se va dejando fuera a los que tienen etiqueta -1. El peor de los casos sería que no hubiera más de un punto seguido con la misma etiqueta, es decir, que las etiquetas fueran +1, -1, +1, -1, etc., pero entonces, como hay $2m$ puntos, habría m de ellos con etiqueta +1, que son lo que meteríamos dentro de los m intervalos, por lo que para $2m$ puntos, se pueden hacer todas las dicotomías. Ahora, para $2m + 1$ puntos ya no se pueden hacer todas las dicotomías, por lo que este será el mínimo punto de ruptura. Supongamos que las etiquetas están puestas de la forma +1, -1, +1, -1,...,-1, +1, es decir, de forma que los dos extremos sean positivos y no haya dos puntos seguidos con la misma etiqueta. Como hay $m+1$ puntos con etiqueta positiva, que son los que tienen que ir dentro de los intervalos, separados por uno con etiqueta negativa, no se puede implementar esta dicotomía con sólo m intervalos y por tanto $2m + 1$ es el mínimo punto de ruptura.

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

a) k^* es un punto de ruptura

b) k^* no es un punto de ruptura

c) todos los puntos de ruptura son estrictamente mayores que k^*

d) todos los puntos de ruptura son menores o iguales a k^*

e) no conocemos nada acerca del punto de ruptura

a) Falso. Para que k^* sea un punto de ruptura hace falta que h implemente menos de 2^{k^*} dicotomías para todo conjunto de k^* puntos y no sólo para uno de ellos, como tenemos aquí, luego no sabemos si es un punto de ruptura o no.

b) Falso. k^* podría ser un punto de ruptura si implementa menos de 2^{k^*} dicotomías para todo conjunto de puntos, pero no sabemos si lo hace o no, sólo sabemos que para uno no las implementa.

c) Falso. De nuevo, no lo sabemos. Puede que H no implemente tampoco 2^{k^*-1} dicotomías para ningún conjunto de $k^* - 1$ puntos, con lo que $k^* - 1$ sería un punto de ruptura.

d) Falso. Esto no puede darse porque si k^* es un punto de ruptura, entonces todos los números mayores que k^* son también puntos de ruptura.

e) Verdadero. Basándonos en las 4 respuestas anteriores, podemos afirmar que no sabemos nada del punto de ruptura.

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

a) k^* es un punto de ruptura

b) k^* no es un punto de ruptura

c) todos los $k \geqslant k^*$ son puntos de ruptura

d) todos los $k < k^*$ son puntos de ruptura

e) no conocemos nada acerca del punto de ruptura

a) Verdadero. Es la definición de punto de ruptura, que se implementen menos (estrictamente) de 2^{k^*} dicotomías para todo conjunto de k^* puntos.

b) Falso. En base a la respuesta anterior.

c) Verdadero. Por a) tenemos que k^* es punto de ruptura, luego todos los números mayores que k^* son puntos de ruptura.

d) Falso. No lo sabemos. Puede que H sí implemente 2^p dicotomías para $p < k^*$, con lo que p no sería punto de ruptura.

e) Falso. Sabemos que k^* y $k > k^*$ son puntos de ruptura.

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

a) Mostrar que existe un conjunto de k^* puntos $x_1, ...x_{k^*}$ que H puede separar ("shatter").

b) Mostrar que H puede separar cualquier conjunto de k^* puntos.

c) Mostrar un conjunto de k^* puntos $x_1, ..., x_{k^*}$ que H no puede separar.

d) Mostrar que H no puede separar ningún conjunto de k^* puntos.

e) Mostrar que $m_H(k) = 2^{k^*}$

a) Esto no nos sirve para probar que k^* es un punto de ruptura, pero sí para probar que no lo es, ya que lo que necesitamos para que sea punto de ruptura es que H no pueda separar ningún conjunto de k^* puntos. Si H puede separar un conjunto de k^* puntos, entonces k^* no es punto de ruptura.

b) Esto nos vale, por el mismo razonamiento del apartado a), para probar que k^* no es punto de ruptura.

c) Esto no nos vale porque aunque haya un conjunto de k^* puntos que no puede separar puede que haya otro que sí, y entonces ya no sería punto de ruptura.

d) Esto sí nos vale para probar que es un punto de ruptura puesto que es justo lo que necesitamos por definición.

e) Depende del valor de k. Si $k = k^*$, nos vale para mostrar que k^* no es un punto de ruptura (aunque no para mostrar que sí), ya que si $m_H(k^*) = 2^{k^*}$ es porque hay un conjunto que de k^* puntos que H puede separar. Si $k > k^*$, no tenemos ninguna información sobre k^* para poder decir ni si es un punto de ruptura o no. El caso $k < k^*$ no puede darse puesto que no puede haber más de 2^k dicotomías para un conjunto de k puntos.

Pregunta 11. Para un conjunto 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 $0,05$?

Lo hacemos como el ejemplo de las diapositivas de clase, es decir, de forma iterativa. Queremos que

$$\sqrt{\frac{8}{N} ln(\frac{4((2N)^{d_{VC}} + 1))}{\delta}} \leq \varepsilon \tag{2}$$

Es decir, que

$$N \geq \frac{8}{\varepsilon^2} ln(\frac{4((2N)^{d_{VC}} + 1)}{\delta}) \tag{3}$$

donde $d_{VC} = 10$, $\varepsilon = 0{,}05$, $1 - \delta = 95\,\%$, con lo que $\delta = 0{,}05$. Sustituyendo nos queda

$$N \geq \frac{8}{0{,}05^2} ln(\frac{4(2N)^{10} + 4}{0{,}05}) \tag{4}$$

Empezamos probando con $N = 1000$ en el segundo miembro a ver qué valor de N nos da y con el valor que nos dé lo volvemos a meter como N en el segundo miembro y repetimos esto hasta que nos salga un valor que cumpla con el mayor o igual.
Para $N = 1000$ obtenemos $N \geq 257251{,}4$, con el que obtenemos $N \geq 434853{,}1$, con el que obtenemos $N \geq 451651{,}6$, lo que sigue sin cumplir la desigualdad. Continuamos y nos da $N \geq 452864{,}8$ con el que obtenemos $N \geq 452950{,}3$, con el que obtenemos $N \geq 452956{,}4$, con el que obtenemos $N \geq 152956{,}8$, con el que obtenemos $N \geq 452956{,}9$. Al meter este valor en el segundo miembro de la desigualdad volvemos a obtener $N \geq 452956{,}9$, con lo que ya sí se cumple la igualdad y tenemos que el tamaño muestral que se necesita para tener tal porcentaje de confianza es como mínimo 452957, ya que N tiene que ser un número entero.

Pregunta 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 $D = (x_1, x_1^2), (x_2, x_2^2)$. El algoritmo de aprendizaje devuelve la línea que ajusta estos dos puntos como g (i.e. H consiste en funciones de la forma $h(x) = ax + b$).

a) Dar una expresión analítica para la función promedio $\overline{g}(x)$.

b) Calcular analíticamente los valores de E_{out}, **bias** y **var**.

2. Bonus

Pregunta 13. Considere el enunciado del ejercicio 2 de la sección ERROR Y RUIDO de la relación de apoyo.

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

$$E_{in}(h) = \sum_{n=1}^{N} |h - y_n| \tag{5}$$

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).

b) Suponga que y_N es modificado como $y_N + \epsilon$ donde $\epsilon \to \infty$. 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}?

Pregunta 14. Considere el ejercicio 12.

a) Describir un experimento que podamos ejecutar para determinar (numéricamente) $\overline{g}(x), E_{out}$, **bias** y **var**.

b) Ejecutar el experimento y dar los resultados. Comparar E_{out} con **bias+var**. Dibujar en unos mismos ejes $\overline{g}(x), E_{out}$ y $f(x)$.

3. Bibliografía

[1]: Learning from data: A short course, Abu-Mostafa, Magdon-Ismail y Lin

[2]: Transparencias de clase disponibles en Decsai