Aprendizaje automático. Cuestionario de teoría 1

Jacinto Carrasco Castillo

2 de abril de 2016

Cuestión 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.

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 período de tiempo fijado.

Solución 1.

a El tipo de aprendizaje adecuado es el aprendizaje supervisado, pues, en primer lugar, queremos asignar una etiqueta concreta a cada animal y no simplemente agruparlos por sus características. En segundo lugar, sabemos a priori características de estos grupos, por lo que tampoco es necesario pasar por una fase previa de aprendizaje no supervisado para dotar de estructura a los datos.
Sin embargo, para este problema, por estar las categorías tan bien diferenciadas y, seleccionando bien las características a medir o tomar como entrada no habría lugar a dudas, sería más apropiada una aproximación por diseño.
En caso de seguir queriendo aplicar una aproximación por aprendizaje (supervisado), el tipo de dato de entrenamiento será un vector de características de dicho animal (tipo de reproducción, piel, sistema respiratorio, ...) y la clase de cada animal.

b En este caso sin dudas el tipo de aprendizaje debe ser supervisado, pues no nos podemos conformar con que se formen grupos de cartas, sino que

debemos ser capaces de asignar un distrito postal según el código postal escrito en la carta. Aquí no cabe pensar que podría resolverse mediante una aproximación por diseño, ya que no podemos determinar todas las formas posibles de escribir un código postal ni características propias fijas y exactas que debe tener la grafía de cada dígito. Para ello entrenaremos a nuestro clasificador con las imágenes de la dirección escrita en la carta (o el código postal) y el dato del distrito postal que le corresponda.

c Aquí el tipo de aprendizaje adecuado es por refuerzo, ya que, para empezar, la salida es gradual, es decir, un índice del mercado no solo sube, o baja, sino que lo hace además en una cierta magnitud y en un cierto tiempo. Los datos de aprendizaje a usar sería la situación sobre el valor a predecir su tendencia y las noticias relacionadas con la empresa. La salida aquí sería el porcentaje de variación en un determinado tiempo.
También podría aplicarse aprendizaje supervisado, pues lo que nos interesa es la salida, es decir, si un valor subirá o no, y no realizar grupos sin saber cuál será su resultado (esto podría ser interesante si quisiésemos separar en un principio en distintos grupos y posteriormente ver qué ocurrió con estos valores). En este caso los datos no tendrán como etiqueta el porcentaje de variación sino si sube o no (aunque podamos generar etiquetas según el porcentaje de subida).

Cuestión 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.

Solución 2.

a Para este problema es más adecuada una aproximación por aprendizaje, ya que para encontrar el ciclo óptimo usaremos datos como el número de vehículos que pasan por cada cruce, las retenciones, el tiempo a esperar, etc. Sin embargo, no podemos realizar una aproximación por diseño, ya que no tenemos unas circunstancias explícitas y concretas para las cuales aplicar un ciclo concreto.

b Para este problema es más adecuada una aproximación por aprendizaje, en concreto, realizando una regresión sobre los datos que nos permita determinar los ingresos medios. Para esto no es factible usar una aproximación por diseño, ya que no se sabe (no existe) una función que nos permita saber, ni siquiera asumiendo un cierto error (un error aceptable, se entiende), con total seguridad, los ingresos medios.

c Para este problema es más adecuada una aproximación por diseño, ya que existen unos protocolos realizados por expertos que determinan si es útil una campaña de vacunación y, aunque puedan fallar en algún caso, se entiende que se han llevado a cabo con las suficientes evidencias científicas y consideración de las consecuencias como para que sean efectivos.

Cuestión 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.

Solución 3. Los elementos a identificar son:

a Entrada

b Salida

c Función objetivo

d Datos de entrada

Como pone en las transparencias, no tenemos conocimiento sobre las propiedades que hacen sabrosos a los mangos, así que lo único que podemos hacer es hacer mediciones sobre las frutas: color, tamaño, peso, textura, etc. Esto lo anotaríamos sobre el mayor y más variado conjunto posible de mangos (teniendo en cuenta que puede darse que muchos mangos no estén en su estado óptimo de recolección y que tendremos que pagar a quien realice la cata de mangos para determinar si están buenos o no).

a La entrada será un elemento x del conjunto donde se mueven las variables que hemos medido (suponiendo que sigue el orden en el que se escribieron antes, $\mathcal{X} = \mathit{Color} \times \mathbb{R}^3 \times \mathbb{R}^+ \times \mathit{Textura}$).

b La salida, puesto que queremos saber cuál será el precio de los mangos y sólo pagarán por aquellos que sean aptos para la venta en los días siguientes, será un punto y en el conjunto de los números positivos y el 0: $\mathcal{Y} = \mathbb{R}_0^+$

c La función objetivo, desconocida, será $f : \mathcal{X} \longrightarrow \mathcal{Y}$, donde $f(x) = 0$, donde el vector $x \in \mathcal{X}$ se corresponde con las características de un mango que no está bueno.

d Los datos de entrad serán $(x_1, y_1), \ldots, (x_N, y_N)$, donde $x_i \in \mathcal{X}, y_i \in \mathcal{Y} \forall i = 1, \ldots, N$.

Cuestión 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 4. Sea w la solución en el paso t del modelo PLA. Si $x(t)$, con etiqueta $y(t)$ está mal clasificado, significa que $\texttt{sign}(w^T x(t)) \neq y(t)$. Dado que sólo uno de los elementos $\{w^T x(t), y(t)\}$ es negativo y el otro es positivo, se tiene que el producto $y(t) w^T x(t) < 0$. Ahora bien, veamos qué pasa al actualizar w a w_{new}. Como queremos probar que la solución se mueve "en la buena dirección", veamos qué le ocurre a $y(t) w_{new}^T x(t)$:

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

Como $x(t)^T x(t) = \|x(t)\| > 0$, ya que todos los puntos del espacio tienen la coordenada homogénea a 1 y por tanto su norma no puede ser 0, estamos añadiendo a $y(t) w^T x(t)$ una cantidad positiva. En resumen:

$$y(t) w_{new}^T x(t) > y(t) w^T x(t)$$

Esto significa que estamos más cerca de clasificar el valor (que sería $y(t) w_{new}^T \cdot x(t) > 0$). Como sabemos, el algoritmo PLA converge, aunque no podemos afirmar que en un único paso clasifiquemos correctamente $x(t)$ pero sí que clasificará correctamente el punto en un número finito de pasos.

Cuestión 5. Considere el enunciado del ejercicio 2 de la sección `FACTIBILIDAD DEL APRENDIZAJE` de la relación 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?

Solución 5. a Teniendo en cuenta que la hipótesis que sería mejor es tomar como función h_1 puesto que $\mathbb{P}[f(x) = 1] = 0{,}9$, nos interesa saber con qué probabilidad en $\mathcal{D}$ habrá más muestras con la etiqueta 1 que con la etiqueta -1. La probabilidad de que haya un cierto número de 1s en $\mathcal{D}$ sigue una distribución binomial ($\mathbb{B}(25, 0{,}9)$), por tanto, tenemos que sumar las probabilidades de los casos en los que hay más 1s que -1s, es decir:

$$\sum_{k=13}^{25} \left(\begin{array}{c} 25 \\ k \end{array} \right) 0{,}9^k 0{,}1^{25-k} = 0{,}9999998$$

b La hipótesis considerada mejor será h_1 si $p > 0{,}5$ y h_2 si $p < 0{,}5$. Entonces se pregunta si existe p tal que C produzca una hipótesis mejor, es decir:

$$\text{Si } p > 0{,}5, \sum_{k=0}^{12} \left(\begin{array}{c} 25 \\ k \end{array} \right) p^k (1-p)^{25-k} > 0{,}5,$$

$$\text{Si } p < 0{,}5, \sum_{k=13}^{25} \left(\begin{array}{c} 25 \\ k \end{array} \right) p^k (1-p)^{25-k} > 0{,}5,$$

lo cual no puede darse, pues significa que la probabilidad de que se den más apariciones del elemento que aparece con menor probabilidad sea más alta que la probabilidad de que se den más apariciones del elemento que aparece con menor probabilidad. Para $p = 0{,}5$, ambos algoritmos de aprendizaje tienen la misma probabilidad de realizar una hipótesis mejor, pues son las dos igual de buenas.

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

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

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

Solución 6. Como es simplemente despejar, lo hacemos en función de M, ε y la cota del error, que llamaremos α y posteriormente sustituimos.

$$2M \exp^{-2\varepsilon^2 N} \leq \alpha; \quad -2\varepsilon^2 N \leq \log(\frac{\alpha}{2M})$$

$$N \geq \frac{-\log(\frac{\alpha}{2M})}{2\varepsilon^2}$$

- $M = 1, \quad N \geq \lceil 839{,}941 \rceil = 840$
- $M = 10, \quad N \geq \lceil 1300{,}458 \rceil = 1301$
- $M = 100, \quad N \geq \lceil 1760{,}975 \rceil = 1761$

Cuestión 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?

Solución 7. El más pequeño punto de ruptura es $2m + 1$, puesto que hasta $2m$ podemos separar cualquier partición de los datos sin más que situar un intervalo en cada conjunto de puntos contiguos de la clase $+1$. El problema llega cuando hay $m + 1$ conjuntos de puntos contiguos de la clase $+1$, y lógicamente el menor número de puntos para que esta circunstancia pueda darse es cuando estos conjuntos son de únicamente un punto, y los conjuntos de puntos que los separan son también de un punto.

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

a $k^\star$ es un punto de ruptura.

b $k^\star$ no es un punto de ruptura.

c Todos los puntos de ruptura son estrictamente mayores que $k^\star$.

d Todos los puntos de ruptura son menores o iguales que $k^\star$.

e No conocemos nada acerca del punto de ruptura.

Solución 8.

a Falso. Puede existir un conjunto de $x_1, x_2, \ldots, x_{k^\star}$ puntos sobre los cuales la clase $\mathcal{H}$ implemente $2^{k^\star}$ dicotomías, con lo que $k^\star$ no es un punto de ruptura.

b Falso. No podemos afirmar que $k^\star$ no es un punto de ruptura, pues es posible que no exista un conjunto de $k^\star$ puntos para los que el número de dicotomías sea $2^{k^\star}$

c Falso. Por el apartado anterior, $k^\star$ podría ser un punto de ruptura, e incluso ser menores, puesto que si $k < k^\star$ es punto de ruptura, $k^\star$ lo es.

d Falso. Por lo comentado también en el apartado anterior, si $k^\star$ es punto de ruptura, $k > k^\star$ son puntos de ruptura.

e Verdadero. Que exista un conjunto de $k^\star$ puntos tal que que $\mathcal{H}$ implemente $2^{k^\star}$ dicotomías no aporta ninguna información sobre el punto de ruptura.

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

a $k^\star$ es un punto de ruptura.

b $k^\star$ no es un punto de ruptura.

c Todos los $k \geq k^\star$ son puntos de ruptura.

d Todos los $k < k^\star$ son puntos de ruptura.

e No conocemos nada acerca del punto de ruptura.

Solución 9.

a Verdadero. Si esto se da para todo conjunto de $k^\star$ puntos, también se da para aquel conjunto tal que $m_{\mathcal{H}}$ es máximo y esta es la definición de punto de ruptura.

b Falso. Contradice el apartado anterior.

c Verdadero. Por ser $k^\star$ punto de ruptura, todos los $k \geq k^\star$ son puntos de ruptura. Si existiese una configuración para la que se implementan 2^k dicotomías, para todo subconjunto de $k^\star$ puntos se estarían implementando $2^{k^\star}$ dicotomías, lo que contradice que $k^\star$ sea un punto de ruptura.

d Falso. Por ejemplo, 4 es punto de ruptura para el Perceptron 2D, pero no así 3.

e Falso. Sabemos que $k^\star$ es punto de ruptura.

Cuestión 10. Si queremos mostrar que $k^\star$ es un punto de ruptura cuales de las siguientes afirmaciones nos servirían para ello:

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

b Mostrar que $\mathcal{H}$ puede separar cualquier conjunto de $k^\star$ puntos.

c Mostrar que un conjunto de $k^\star$ puntos $x_1, \ldots, x_{k^\star}$ que $\mathcal{H}$ no se puede separar.

d Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de $k^\star$ puntos.

e Mostrar que $m_{\mathcal{H}}(k) = 2^{k^\star}$

Solución 10.

a Nos ayudaría a decir que $k^\star$ no es un punto de ruptura, pues $m_{\mathcal{H}}(k^\star) = 2^{k^\star}$ y se alcanza en el conjunto dado.

b Nos ayuda a decir que $k^\star$ no es un punto de ruptura, al igual que en el apartado anterior, sólo que vale cualquier conjunto.

c No nos sirve, pues puede existir otro conjunto de $k^\star$ puntos que sí sea separable, y por tanto $k^\star$ no sea un punto de ruptura. También podría darse que ningún conjunto de $k^\star$ puntos fuese separable, esto es, $m_{\mathcal{H}}(k^\star) < 2^{k^\star}$ y por tanto $k^\star$ sería un punto de ruptura.

d Esto nos dice que $k^\star$ es un punto de ruptura, pues para todo conjunto de $k^\star$ puntos el número de dicotomías es $< 2^{k^\star}$.

e Esta afirmación nos puede aportar formación dependiendo de la relación entre k y $k^\star$. Si $k = k^\star$, claramente $k^\star$ no es un punto de ruptura. Si $k > k^\star$ no nos da información, pues puede darse $m_{\mathcal{H}}(k) = 2^{k^\star} < 2^k$ (k sería un punto de ruptura) y darse o bien $m_{\mathcal{H}}(k^\star) = 2^{k^\star}$ o $m_{\mathcal{H}}(k^\star) < 2^{k^\star}$, con lo que no deducimos nada. El caso $k < k^\star$ no puede darse.

Cuestión 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?

Solución 11. Planteamos la ecuación implícita correspondiente:

$$N \geq \frac{8}{\varepsilon^2} \log\left(\frac{4((2N)^{d_{VC}} + 1)}{\delta}\right)$$

En nuestro caso, $\delta = 1 - 95\,\% = 0{,}05,\ \varepsilon = 0{,}05$

$$N \geq \frac{8}{0{,}05^2} \log\left(\frac{4((2N)^1 0 + 1)}{0{,}05}\right)$$

Resolvemos la ecuación de manera iterativa:

```
cota=1
for(i in 1:10){
    cota=3200*log((4*(2*cota)^10 + 4)/0.05)
    print(cota)
}
```

y en 8 iteraciones ya llegamos al punto fijo de la ecuación, $N \geq \lceil 452956{,}9 \rceil$, luego $N \geq 452957$

Cuestión 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 es 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 (i.e. $\mathcal{H}$ consiste en funciones de la forma $h(x) = ax + b$).

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

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

Solución 12.

a Partimos de los datos $S = \{(x_1, x_1^2), (x_2, x_2^2)\}$. La función g^S será la recta que pase por estos dos puntos: $g^S(x) = \frac{x_2^2 - x_1^2}{x_2 - x_1}(x - x_1) + x_1^2 = (x_2 + x_1)(x - x_1) + x_1^2 = (x_1 + x_2)x - x_1x_2$. Se define $\bar{g}(x)$ como $\mathbb{E}_S[g^S(x)]$:

$$\bar{g}(x) = \mathbb{E}_S[g^S(x)] = \mathbb{E}_S[(x_1 + x_2)x - x_1x_2]$$

Entonces usamos la linealidad de la esperanza, que x es independiente de S y que la distribución de x_1 y x_2 son independientes con lo que la esperanza del producto es el producto de las esperanzas:

$$\bar{g}(x) = (\mathbb{E}_S[x_1] + \mathbb{E}_S[x_2]) \cdot x - \mathbb{E}_S[x_1]\mathbb{E}_S[x_2]$$

Por ser $\mathbb{E}_S[x_1] = \frac{1-1}{2} = 0 = \mathbb{E}_S[x_2]$

$$\bar{g}(x) = 0$$

b Para hallar E_{out} usaremos $E_{out} = \mathbb{E}_S[E_{out}(g^S)] = \mathbb{E}_x[bias(x) + var(x)] = bias + var$

(2)

$bias = \mathbb{E}_x[bias(x)] = \mathbb{E}_x[\mathbb{E}_S[(\bar{g}(x) - f(x))^2]] = \mathbb{E}_x[\mathbb{E}_S[x^4]] = \mathbb{E}_x[x^4] =$

$$= \mathbb{E}_x[x^4] = \int_{-1}^{1} \frac{x^5}{2 \cdot 5} dx = \frac{1}{5}$$

$$var = \mathbb{E}_x[var(x)] = \mathbb{E}_x[\mathbb{E}_S[(g^S(x) - \bar{g}(x))^2]] = \mathbb{E}_x[\mathbb{E}_S[g^S(x)^2]];$$

Calculamos aparte $\mathbb{E}_S[g^S(x)^2]$, usando que $\mathbb{E}_S[x_1] = 0$ y que x_1, x_2 son independientes.

$$\begin{aligned} \mathbb{E}_S\left[g^S(x)^2\right] &= \mathbb{E}_S\left[(x_1+x_2)^2x^2 + x_1^2x_2^2 - 2xx_1x_2(x_1+x_2)\right] = \\ &= x^2(\mathbb{E}_S\left[x_1^2\right] + \mathbb{E}_S\left[x_2^2\right]) + \mathbb{E}_S\left[x_1^2\right]\mathbb{E}_S\left[x_2^2\right] = \\ &= \frac{2x^2}{3} + \frac{1}{9} \end{aligned} \tag{3}$$

donde hemos usado $\mathbb{E}_S\left[x_1^2\right] = \int_{-1}^{1} \frac{x_1^2}{2} dx_1 = \frac{1}{3} = \mathbb{E}_S\left[x_2^2\right]$ (pues son i.d.). Por tanto:

$$var = \mathbb{E}_x[var(x)] = \mathbb{E}_x\left[\frac{2x^2}{3} + \frac{1}{9}\right] = \frac{1}{3}$$

Y nos queda:

$$E_{out} = bias + var = \frac{1}{5} + \frac{1}{3} = \frac{8}{15}$$

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

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

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

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 $y_N + \varepsilon$, donde $\varepsilon \leftarrow \varepsilon$. 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} ?

Cuestión Opcional 2. Considere el ejercicio 12.

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

2. Ejecutar el experimento y dar los resultados. Comparar E_{out} con **bias**+ **var**. Dibujar en unos mismos ejes