Preguntas de Teoría 1

Luis Suárez Lloréns

Preguntas

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

a) Categorizar un grupo de animales vertebrados en pajaros, 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.

Respuesta. a) Supervisado. Tenemos unos datos de entrada, y unas salidas esperadas bien definidas — mamífero, reptil... —. Por tanto, nos encontramos en la situación de aprendizaje supervisado.

b) No supervisado. No disponemos de una salida para los datos. Nuestro objetivo es encontrar relaciones entre los datos, lo cual es propio del aprendizaje no supervisado.

c) Refuerzo. Los datos no disponen de una salida directamente, pero en los mismos datos se puede saber si hay una subida o una bajada. Por tanto, al tener los resultados en los mismos datos, tenemos una situación de aprendizaje por refuerzo.

Cuestión 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) 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.

Respuesta. a) Aprendizaje. Para poder modelizar bien esta situación, necesitamos muchos datos sobre el tráfico del cruce. Tenemos que tomar todos estos datos, para conseguir optimizar el semáforo. Es decir, tenemos que aprender de los datos.

b) Aprendizaje. Otra vez, necesitamos una gran cantidad de datos, e intentar aprender de los mismos para conseguir una función que nos pueda indicar los ingresos medios de una persona. Al tener tantos datos y aprender de ellos, tenemos que aprender de los datos.

c) Diseño. Es un problema en el que puede que incluso, no tengamos datos de situaciones pasadas —por ejemplo, una nueva enfermedad— y por tanto, no podemos aprender de datos. Por tanto, la solución se encuentra por diseño.

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.

Respuesta. $\mathcal{X}$: Los datos que podemos usar son: color, tamaño, calibre, textura, peso.

$\mathcal{Y}$: 1 y -1, según sean agrios o no.

$\mathcal{D}$: Los datos indicados en $\mathcal{X}$ de la muestra que se ha podido recoger en la granja, así como la etiqueta que le asignamos a cada una de las frutas.

f: La función buscada, que nos indica si la fruta es agria o no

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).

Respuesta. Veamos que pasa si tenemos un dato que deberíamos clasificar como 1, y clasificamos como -1. Tenemos que $w^t(t)x_i < 0$ y $y_i = 1$. Los nuevos coeficientes serían $w(t+1) = w(t) + y_i x_i$. Entonces:

$$w^t(t+1)x_i = (w(t) + y_i x_i)^t x_i = w^t(t)x_i + (y_i x_i)^t x_i$$

Por tanto, los nuevos coeficientes clasificarán mejor — es decir $w^t(t+1)x_i > w^t(t)x_i$— si $(y_i x_i)^t x_i > 0$:

$$(y_i x_i)^t x_i = y_i \sum_j x_{i_j}^2 = \sum_j x_{i_j}^2 > 0$$

Por tanto, clasifica mejor después de dar el paso.

Para el caso de clasificar mal un dato con clasificación -1, pasa algo similar. En ese caso, clasificará mejor si $(y_i x_i)^t x_i < 0$:

$$(y_i x_i)^t x_i = y_i \sum_j x_{i_j}^2 = -\sum_j x_{i_j}^2 < 0$$

Entonces, al realizar un paso del PLA, mejoramos la clasificación.

Cuestión 5. Considere el enunciado del ejercicio 2 de la sección FACTIBILIDAD DEL APRENDIZAJE de la relación apoyo a) Si p = 0;9 ¿Cual 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?

Respuesta. Antes de responder a los dos apartados, veamos cuales son las hipótesis que tenemos como S y C. En el segundo apartado del ejercicio 2, se dice que se suponga que $\mathcal{D} = \{1, 1, ..., 1\}$. Por tanto, S tomará siempre la

hipótesis $h_1 = 1$ y C tomará la hipótesis $h_{-1} = -1$.

a) El ejercicio pide estudiar el error fuera de la muestra. Hay dos maneras de interpretar la pregunta, como el error total fuera de la muestra o como el error para una muestra concreta. Vamos a empezar con el error total. Este se define como:

$$E_{out}(h) = \mathcal{P}\left[f(x) \neq h(x)\right]$$

Por tanto, sabiendo que la hipótesis de S es h_1 y la de C es h_{-1}, tenemos:

$$E_{out}(h_1) = \mathcal{P}\left[f(x) \neq h_1(x)\right] = \mathcal{P}\left[f(x) \neq 1\right] = \mathcal{P}\left[f(x) = -1\right] = 0.1$$

$$E_{out}(h_{-1}) = \mathcal{P}\left[f(x) \neq h_{-1}(x)\right] = \mathcal{P}\left[f(x) \neq -1\right] = \mathcal{P}\left[f(x) = 1\right] = 0.9$$

Luego es mejor S.

En cuanto al error asociado a una muestra de tamaño N, S clasifica mejor la muestra cuando hay más 1 que -1. La probabilidad de que pase esto, se puede calcular como:

$$\sum_{i=\mathbb{Z}(\frac{N}{2}+1)}^{N} \binom{N}{i} 0.9^i 0.1^{N-i}$$

Para una muestra de tamaño 1, S clasifica mejor con probabilidad 0.9. Para una muestra de tamaño 3, la probabilidad es 0.975. Y este valor sigue aumentando conforme aumenta N. Por tanto, S nos va a clasificar mejor una muestra dada casi siempre.

b) Usando los mismos cálculos que en el apartado anterior, podemos ver que si p < 0.5, C va a clasificar mejor que S. La diferencia es más marcada conforme más se acerque a 0 — pero $p \neq 0$, pues $\mathcal{D}$ contiene 1—.

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

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

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 $M = 100$.

Respuesta. Vamos a igualar la parte derecha de la fórmula, y vamos a despejar N:

$$2Me^{-2N\epsilon^2} = 0.03; e^{-2N\epsilon^2} = \frac{0.03}{2M}$$

$$-2N\epsilon^2 = log\left(\frac{0.03}{2M}\right); N = \frac{log\left(\frac{0.03}{2M}\right)}{-2\epsilon^2}$$

Al sustituir, obtenemos la solución al problema anterior, y solo tenemos que tomar el número natural inmediatamente superior:

$$M = 1 \Rightarrow N^{'} = 839.94 \Rightarrow N = 840$$

$$M = 10 \Rightarrow N^{'} = 1300.45 \Rightarrow N = 1301$$
$$M = 100 \Rightarrow N^{'} = 1760.97 \Rightarrow N = 1761$$

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

Respuesta. Para clasificar disponemos de M intervalos. Primero, vamos a ver que con M intervalos no se puede separar —shatter— una muestra de 2M+1 puntos distintos. Después veremos que sí puede separar una muestra de menor tamaño.

En el caso de tener 2M+1 puntos distintos, al estar en la recta real, los ordenamos. Después, les damos como etiquetas $\{1, -1, 1, -1, ..., -1, 1\}$. Entonces, para cada uno tendríamos que tener un intervalo, ya que los -1 obligan a que no pueda crearse un intervalo que pueda clasificar dos 1. Por tanto, necesitaríamos M+1 intervalos para clasificar la muestra y solo tenemos M. Luego no podemos separar una muestra de 2M+1 puntos distintos. Para terminar, queda la opción de que haya puntos repetidos en el conjunto, pero si tomamos dos elementos que sean iguales y les asignamos etiquetas distintas, tampoco se podría separar.

Si tenemos 2M puntos, podríamos clasificar una situación como la anterior. Además, otras situaciones serían más fáciles de clasificar pues encontraríamos clasificaciones consecutivas iguales, que podrían ser clasificadas por el mismo intervalo — o por el mismo hueco, si fueran -1 —. Por tanto, podemos separar una muestra de 2M puntos separados.

Cuestión 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

Respuesta. a. Falso. Puede existir otro conjunto para el que $\mathcal{H}$ sí implemente 2^{k^*} dicotomías. Por tanto, no podemos decir que sea punto de ruptura.

b. Falso. La situación es similar al apartado anterior. Para que k^* sea punto de ruptura, tendríamos que encontrar otro conjunto para el que $\mathcal{H}$ sí implemente 2^{k^*} dicotomías. Como desconocemos si existe o no, no podemos decir que no sea punto de ruptura.

c. Falso. Hemos visto en el apartado "a" que k^* podría ser punto de ruptura, luego no podemos afirmar esto.

d. Falso. Aparte de la falta de información que tenemos para afirmar algo así, tenemos otra contradicción en que si k es punto de ruptura, cualquier punto mayor que k también lo es.

e. Cierto. Viendo las afirmaciones anteriores, en especial a y b, vemos que no conocemos nada relativo a si k^* es punto de ruptura o no.

Cuestión 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

Respuesta. a. k^* es punto de ruptura, pues el conjunto de dicotomías de $\mathcal{H}$ es menor que 2^{k^*}. Por tanto, $\mathcal{H}$ no puede separar —shatter— la muestra una muestra de tamaño k^*.

b. Al ser el apartado "a." cierto, este es falso.

c. Verdadero. Si no puede separar k^* puntos, no puede separar más puntos aún.

De una manera más formal, podríamos verlo de la siguiente forma. Supongamos que fuera falso, es decir, existe $k^{'} > k^*$ que no es punto de ruptura. Entonces, $\mathcal{H}$ puede separar $k^{'}$ puntos. Por tanto, también podría separar completamente una cantidad menor de puntos, pues podemos agregar puntos hasta llegar a $k^{'}$ y dividir todos estos puntos. Por tanto, k^* no sería punto de ruptura, y entonces tenemos una contradicción.

d. No tenemos información ninguna para afirmar esto. No obstante, no es necesariamente falso. Hay $\mathcal{H}$ que no pueden separar siquiera un elemento —por ejemplo $\mathcal{H} = \{-1\}$ no podría clasificar bien un punto que vale 1—. Por tanto, no podamos decir que la frase sea falsa en cualquier caso, pese que en la mayoría de los casos sí lo sería.

e. Falso, tenemos información de los puntos de ruptura de k^* en adelante. Pero sí que es verdad que desconocemos completamente con lo que pasa para $k < k^*$.

Cuestión 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 $m_H(k) = 2^{k^*}$

Respuesta. a. Si puede separar un conjunto de puntos de tamaño k^*, entonces $m_{\mathcal{H}}(k^*) = 2^{k^*}$. Por tanto no cumple la definición, luego no es un punto de ruptura.

b. Si divide a cualquier conjunto de k^*, entonces dividirá a uno. Entonces podemos reducir este apartado al anterior, luego no es un punto de ruptura.

c. No es suficiente para decir que sea punto de ruptura, pues puede existir un conjunto distinto de k^* que si pueda separar. Por tanto no nos sirve para saber si es o no punto de ruptura.

d. Si es suficiente para decir que es punto de ruptura. Al no poder separar ningún conjunto, sabemos que el número de dicotomías que puede generar $\mathcal{H}$ es menor que 2^{k^*}. Por tanto, $m_{\mathcal{H}}(k^*) < 2^{k^*}$, por ser $m_{\mathcal{H}}(k)$ el máximo número de dicotomías posible en una muestra de tamaño. Entonces, cumple la definición de punto de ruptura

e. Si $m_{\mathcal{H}}(k) = 2^{k^*}$, entonces es falso $m_{\mathcal{H}}(k^*) < 2^{k^*}$, y por tanto no cumple la definición. Luego no es un punto de ruptura.

Cuestión 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 $0,05$?

Respuesta. Si intentamos despejar N, no podemos conseguirlo, por culpa del logaritmo. Podemos llegar hasta:

$$N = \frac{8log(\frac{4((2N)^{d_{vc}}+1)}{\delta})}{0.05^2}$$

Para obtener el valor, trabajamos de manera iterativa. Damos un valor a N, evaluamos la parte derecha de la ecuación, y lo tomamos como el siguiente N. Este proceso ha nos ha llevado a que $N \approx 40790.2$, por tanto el N que buscamos es 40791.

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 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.

Respuesta. Dado los datos de los que aprendemos (x_1, x_1^2) y (x_2, x_2^2), la recta que conseguimos es la siguiente:

$$(x_1 + x_2)x - x_1x_2$$

a) Para encontrar $\bar{g}$, tenemos que hacer la esperanza en $\mathcal{D}$ de g_d. Esto se traduce en una integral doble en x_1 y x_2 entre -1 y 1 en ambas, multiplicando por la función de densidad, que es $\frac{1}{4}$ — la función de densidad de una sola de una sola de las variables es $\frac{1}{2}$.—. Por tanto.

$$\bar{g} = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1+x_2)x - x_1x_2 d_{x_1} d_{x_2} = \frac{x}{2}\int_{-1}^{1} x_2 d_{x_2} = 0$$

b) Para calcular el error fuera de la muestra, lo haremos como la suma de bias y varianza.

bias:

$$bias(x) = (\bar{g}(x) - f(x))^2 = x^4$$

$$bias = E_x\left[x^4\right] = \frac{1}{2}\int_{-1}^{1} x^4 dx = \frac{1}{5}$$

varianza:

$$var(x) = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\left((x_1+x_2)x - x_1x_2\right)^2 d_{x_1} d_{x_2} =$$

$$\frac{1}{4}\int_{-1}^{1} x_2^2(2x^2+\frac{2}{3}) + x_2(-\frac{4}{3}x) + \frac{2}{3}x^2 d_{x_2} =$$

$$\frac{2}{3}x^2 + \frac{1}{9}$$

$$var = E_x\left[var(x)\right] = \frac{1}{2}\int_{-1}^{1}\frac{2}{3}x^2 + \frac{1}{9} d_x = \frac{1}{3}$$

Por tanto, E_{out} es $\frac{8}{15}$.

Bonus

Bonus 1. 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,

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

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}?

Respuesta. a) Vamos primero a modificar la expresión que nos da el ejercicio. sea m el primer punto a la derecha de h.

$$E_{in}(h) = \sum_{n=1}^{N} |h-y_n| = \sum_{n=1}^{m-1} (h-y_n) + \sum_{n=m}^{N} (y_n-h) = \sum_{n=m}^{N} y_n - \sum_{n=1}^{m-1} y_n + (-N+2m)h$$

Ahora, supongamos que tenemos un número de datos par. Supongamos que tenemos h_{med} y otro h en el intervalo inmediatamente mayor. La comprobación para el caso de h menor se realiza del mismo modo.

Entonces, las sumatorias solo difieren en un elemento, x_m. Si restamos al error cometido h, el cometido por h_{med} tenemos:

$$-2x_m + 2h > 0, \textit{ pues } \; x_m < h$$

Por tanto, clasifica mejor h_{med} que un h del intervalo superior. Este argumento se puede repetir para los demás intervalos superiores.

Si tenemos un número de datos impar, podríamos eliminar uno de los extremos. Nos quedaría el caso par, por tanto h tiene que estar en el intervalo intermedio. Al añadir el punto que hemos quitado, para que h lo clasifique lo mejor posible, tiene que ir al extremo del intervalo más próximo al dato. Este punto es la mediana de todos los datos.

b) Al sumarle un número infinito al último valor, la mediana no se ve afectada, mientras que la media va a infinito — el nuevo parámetro añade a la media $\frac{\epsilon}{N}$—.

Por tanto, podemos ver que la mediana no le afecta una gran perturbación en un dato, y a la media sí.