Aprendizaje Automático: Cuestionario 1

Alejandro Alcalde, *Universidad de Granada*

3 de abril de 2016

Índice

1. Ejercicio 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.

- ***Categorizar un grupo de animales vertebrados en pajaros, mamíferos, reptiles, aves y anfibios.***

  Este caso sería supervisado. Como variables se podrían usar por ejemplo distintas propiedades que todos los vertebrados tienen. Por ejemplo si tiene plumas o no, densidad ósea, si tiene pelo o no, si pone huevos, si tiene cola o no etc. Al igual que el problema de la moneda, es posible hacer este problema no supervisado. Bastaría con no dar las etiquetas y al visualizar los datos nos daríamos cuenta de que los datos están separados en una serie de clusters. En ese caso pues, no podríamos decir qué tipo de animal es, solo que pertenece a la clase 1, 2, 3, 4 o 5.

- ***Clasificación automática de cartas por distrito postal***

  Supervisado. En este caso estamos ante un caso de clasificación multiclase. La respresentación del problema puede aproximarse de dos modos, una más sencilla que otra.

  Como hemos visto en clase de prácticas, debemos ser capaces de reconocer dígitos. Podemos intentar reconocer características propias de cáda dígito, como simetría vertical e intensidad media. Esto nos da dos variables para trabajar, nuestra $\mathcal{X}$ estaría formada por la dos variables simetría vertical e intensidad media, $\mathcal{X} = (Sim, int)$.

Otra aproximación podría ser usar todo el conjunto de características del dígito, es decir, si el dígito está formado por una imagen de 16×16 píxeles, usaríamos los valores de cada píxel para aprender a clasificar el dígito. Esta aproximación es sin duda más compleja, ya que en lugar de 2 variables tenemos 256. En la primera aproximación $\mathcal{X} \in \mathcal{R}^2$, mientras que en la segunda $\mathcal{X} \in \mathcal{R}^{256}$.

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

  Si usamos valores históricos índices anteriores, sabremos si el mercado de valores subió o bajó para esos datos, por tanto estamos ante un ejemplo de aprendizaje supervisado.

2. Ejercicio 2

¿Cuales de los siguientes problemas son más adecuados para una aproximación por aprendizaje y cuales 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.***

  Por aprendizaje, ya que la diferencia principial entre la aproximación por aprendizaje y por diseño reside en que para la última, es posible calcular la f analíticamente sin necesidad de ver ningún dato, debido a que el problema está bien especificado. En este caso no lo está, y necesitaríamos datos para saber a qué horas hay más tráfico en el cruce, cómo es de fluido el tráfico etc. Por tanto no podemos calcular la f analíticamente. Una vez tuvieramos los datos y los visualizáramos, tendríamos una mejor idea de cómo influye el ciclo de las luces en la fluidez del tráfico.

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

  Aquí los datos del problema están bien definidos, por tanto se trata de una aproximación por diseño. Bastaría con acceder a los datos mencionados en el enunciado y realizar una media de los ingresos que obtienen las personas en base a las distintas variables especificadas (Nivel de educación, edad etc.)

- ***Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.***

Aqui tenemos un ejemplo claro que confirma la aproximación por aprendizaje, *Google Flu*, ya que los datos del problema en este caso no están bien definidos, necesitamos aprender mediante datos. En el caso de Google, se usaron más de 30.000 millones de búsquedas diárias y encontraron una combinación de 45 términos de búsqueda que al usarse con un modelo matemático presentaba una correlación fuerte entre su predicción y las cifrar oficiales de la enfermedad.

3. 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.

Los elementos formales del problema son:

- Las variables de entrada $\mathbf{x}$. Basandonos en un estudio previo de la fruta, podemos intentar extrar una serie de variables que influyan en el sabor, como el tamaño, el color, la dureza de la piel, si está maduro o no, el peso, textura etc. Una vez elegidas, estas serán el vector de entrada $\mathbf{x}$ que usemos. Si consideramos que vamos a usar dos características para determinar el precio, por ejemplo el color y la textura, el vector $\mathbf{x}$ estaría formado por (x_0, x_1, x_2), con $x_0 = 1$ para poder añadir el umbral y simplificar la fórmula.

- Luego, basándonos en esas variables, necesitamos representar la salida, en este caso queremos saber el precio de la pieza, luego la salida será un número real ($y \in \mathcal{R}$)

- La función objetivo, desconocida, como siempre, sería de la forma $f : \mathcal{X} \rightarrow y$ y transformaría un vector de entrada al precio de la pieza de fruta.

- Para los datos, necesitaríamos un histórico de frutas con buen sabor etiquetadas con su precio en el mercado. Para así poder entrenar al modelo. Los datos sería de la forma $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$. Donde cada $\mathbf{x}_n$ está compuesto por la variables mencionadas anteriormente.

- Llegados a este punto, solo queda especificar cual es el modelo de aprendizaje a usar, para ello es neceario definir el conjunto de hipótesis $\mathcal{H}$ y el algoritmo de aprendizaje. Como la salida es un valor real, podemos usar regresión lineal para nuestro modelo, de este modo, el conjunto de hipótesis $\mathcal{H}$ será:

$$h(\mathbf{x}) = \sum_{i=0}^{d} w_i x_i = \mathbf{w}^T \mathbf{x}$$

- Para el algoritmo de aprendizaje, podemos obtener la pseudo-inversa de $\mathbf{X}$, siendo esta una matriz con todos los datos de entrada y el vector $\mathbf{y}$ conteniendo las etiquetas, y calcular su pseudoinversa como:

  $$\mathbf{X}^\dagger = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

  y el vector de pesos $\mathbf{w} = \mathbf{X}^\dagger \mathbf{y}$

4. Ejercicio 4

Suponga un modelo PLA y un dato $\mathbf{x}(t)$ 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 $\mathbf{x}(t)$.

$\mathbf{wx}$ es el producto escalar entre los dos vectores, cuando el ángulo formado por ambos es menor de 90 grados, el signo del producto es positivo (Ya que la interpretación geométrica de este producto es $\|\mathbf{w}\|\,\|\mathbf{x}\|\cos(\theta)$). Cuando el ángulo sea mayor de 90 grados el signo será negativo.

Cuando hay un punto mal clasificado, su etiqueta no coincide con el resultado del producto escalar, es aquí cuando la regla de adaptación entra en juego:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + y\mathbf{x}$$

Esta regla suma o resta a $\mathbf{w}$ el vector $y\mathbf{x}$, en función de la etiqueta correcta. Pueden darse dos casos:

- Si $y = 1$ pero el punto está mal clasificado y tenemos un -1, $\mathbf{w}(t) + y\mathbf{x}$ sumará al vector y por tanto ahora estará bien clasificado al formar $\mathbf{x}$ e $\mathbf{w}$ un ángulo inferior a 90 grados.

- Analogamente para $y = -1$, salvo que en este caso se le resta al vector, formando ahora un grado mayor a 90 grados y etiquetándose correctamente como un -1.

5. Ejercicio 5

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

- Si $p = 0{,}9$ ¿Cual es la probabilidad de que S produzca una hipótesis mejor que C?

  Siendo $\mathbb{P}[f(x) = 1] = 0{,}9$ sabemos entonces que $\mathbb{P}[S = f(x)] = 0{,}9$ y por tanto

  $$\mathbb{P}[C = f(x)] = 1 - \mathbb{P}[S = f(x)] = 1 - 0{,}9 = 0{,}1$$

  Con lo cual podemos deducir que $\mathbb{P}[\mathbb{P}[S = f(x)] > \mathbb{P}[C = f(x)]] = 1$

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

  Sí, para cualquier valor de p por debajo de $0{,}5$ C puede producir una hipótesis mejor que S.

6. Ejercicio 6

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

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

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

Fijamos $\epsilon = 0{,}05$, queremos $2Me^{-2N\epsilon^2} \leq 0{,}03$

- Probemos el primer caso, $M = 1$:

$$
\begin{aligned}
2Me^{-2N\epsilon^2} &\leq 0{,}03 \\
\ln(2e^{-2N\epsilon^2}) &\leq \ln 0{,}03 \\
\ln 2 + \ln(e^{-2N\epsilon^2}) &\leq \ln 0{,}03 \\
-2N\epsilon^2 &\leq \ln 0{,}03 - \ln 2 \\
N &\leq \frac{\ln 0{,}03 - \ln 2}{-2\epsilon^2} \\
N &\leq \frac{\ln 0{,}03 - \ln 2}{-2 \cdot 0{,}05^2} \\
N &\leq 839{,}94
\end{aligned}
$$

Con este resultado, concluimos que para una cota de $0{,}03$, el número de datos mínimo necesario es 840.

- Para $M = 10$:

$$
\begin{aligned}
20e^{-2N\epsilon^2} &\leq 0{,}03 \\
\ln(20e^{-2N\epsilon^2}) &\leq \ln 0{,}03 \\
\ln 20 + \ln(e^{-2N\epsilon^2}) &\leq \ln 0{,}03 \\
-2N\epsilon^2 &\leq \ln 0{,}03 - \ln 20 \\
N &\leq \frac{\ln 0{,}03 - \ln 20}{-2\epsilon^2} \\
N &\leq \frac{\ln 0{,}03 - \ln 20}{-2 \cdot 0{,}05^2} \\
N &\leq 1300{,}46
\end{aligned}
$$

Ahora, al haber incrementado M, necesitamos más datos para satisfacer la cota, siendo ahora el número mínimo de datos necesario 1301. A partir de este número de datos, la cota se satisface.

- Para $M = 100$

$$
\begin{aligned}
N &\leq \frac{\ln 0{,}03 - \ln 200}{-2 \cdot 0{,}05^2} \\
N &\leq 1056{,}15
\end{aligned}
$$

Sin embargo, al sustituir este dato en la desigualdad $2Me^{-2N\epsilon^2} \leq 0{,}03$ no obtenemos un dato coherente (1,013). Esto se debe a que para un

M tan grande no hay garantía de que se cumpla la cota, ya que M respresenta el número de hipótesis. Al ser el conjunto de hipótesis tan grande, estamos ampliando la cota para la probabilidad de que el evento no deseado ocurra (Este evento es $\mathbb{P}[|E_{out}(g) - E_{in}(g)| > \epsilon]$), en este caso, con $M = 100$ hemos fijado la cota a ≤ 1, cosa que no aporta información ya que la probabilidad del evento siempre estará por debajo de 1. Luego para $M = 100$ no conseguimos información útil.

7. Ejercicio 7

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

Una forma de resolverlo sería calcular función de crecimiento de los *M-Intervalos*, la cual sabemos es $\binom{N+1}{2} + \binom{N+1}{4} + \binom{N+1}{6} \dots \binom{N+1}{M} + 1$ y luego ir probando hasta que obtengamos un valor menor que 2^N, pero obtener la expresión en polinomios de esta combinatoria es complicado. Así que vamos a intentar generalizar a partir de lo que sabemos sobre el punto de ruptura para un intervalo:

Para un intervalo sabemos que el punto de ruptura está en 3, ya que no seríamos capaces de generar el siguiente etiquetado: **oxo**, donde **o** corresponde a un 1, y el **x** a un -1. Si probamos para dos intervalos, la configuración que no podemos generar es **oxoxox**, ya que con dos intervalos no somos capaces de hacer ese etiquetado, luego el punto de ruptura para $M = 2$ está en 5 puntos. Para tres invervalos, no podemos etiquetar **oxoxoxo**, luego el punto de ruptura para $M = 3$ está en 7 puntos, para $M = 4$ tenemos **oxoxoxoxo**, lo cual fija en 9 el punto de ruptura. Podríamos seguir, pero se puede intuir una regla, los puntos de ruptura para cada intervalo son $3, 5, 7, 9, \dots$ podemos deducir que el punto de ruptura para M intervalos viene dado por $2M + 1$.

8. Ejercicio 8

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

- k^* es un punto de ruptura.

Verdadero, ya que la definición de punto de ruptura establece que cualquier n para el que $m_{\mathcal{H}}(n) < 2^n$ es un punto de ruptura.

- k^* no es un punto de ruptura.

  Falso.

- Todos los puntos de ruptura son estrictamente mayores que k^*.

  Verdadero, independientemente de que k^* sea el punto de ruptura mínimo (El primero en romper), a partir de él el resto también serán puntos de ruptura, y serán mayores que k^*.

- Todos los puntos de ruptura son menores o iguales a k^*.

  Falso, por lo dicho en c.

- No conocemos nada acerca del punto de ruptura.

  Falso, No sabemos si es el primer punto de ruptura, pero al saber $m_{\mathcal{H}}(n) < 2^n$, podemos asegurar que en efecto es punto de ruptura.

9. Ejercicio 9

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

- k^* es un punto de ruptura.

  Verdadero. Tomemos como ejemplo el Perceptrón, su punto de ruptura está en 4 puntos, pero podemos tomar un conjunto de 4 puntos en el que sí podamos generar todas las 2^4 posibles etiquetas, sin embargo, existe una disposición de 4 puntos que el perceptrón no puede separar, y por tanto 4 es su punto de ruptura, al decir que $\mathcal{H}$ implementa $< 2^{k^*}$ para todo conjunto de puntos k^*, se está diciendo que existe un punto de ruptura para ese conjunto de puntos.

- k^* no es un punto de ruptura.

  Falso

- Todos los $k \geq k^*$ son puntos de ruptura.

  Verdadero. Encontrando el primer punto de ruptura, todos los subsecuentes puntos lo son. Así, para los segmentos, encontrado el primper punto de ruptura, que es 2 todos los demás puntos lo son $(3, 4, 5 \dots)$

- Todos los $k < k^*$ son puntos de ruptura.

  Falso, por debajo de un punto de ruptura, puede haber o no más puntos de ruptura. Si k^* es el conjunto mínimo de puntos para el cual hay punto de ruptura, es decir, el punto de ruptura mínimo, no existen puntos de ruptura por debajo.

- No conocemos nada acerca del punto de ruptura.

  Falso. Conocemos que todo conjunto k^* implementa $\mathcal{H} < 2^{k^*}$, con lo cual sabemos que habrá puntos de ruptura.

10. Ejercicio 10

Si queremos mostrar que k^* es un punto de ruptura cuales de las siguientes afirmaciones nos servirían para ello:

La c, Mostrar un conjunto de k^* puntos $x_1, x_2, \ldots, x_{k^*}$ que $\mathcal{H}$ no puede separar. Como ejemplo ponemos el Perceptrón, basta con mostrar que con un conjunto de 4 puntos x_1, x_2, x_3, x_4 el algoritmo no es capaz de lograr todas las posibles 2^4 separaciones posibles.

11. 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 $0{,}05$?

Al querer un $95\,\%$ de confianza, $\delta = 0{,}05$

$$
\begin{aligned}
N &\geq \frac{8}{\epsilon^2} \ln\left(\frac{4((2N)^{d_{VC}} + 1)}{\delta}\right) \\
N &\geq \frac{8}{0{,}05^2} \ln\left(\frac{4((2 \cdot 100)^{10} + 1)}{0{,}05}\right) \\
N &\geq 183000
\end{aligned}
$$

Sustituimos ahora el valor de N por el resultado:

$$
\begin{aligned}
N &\geq \frac{8}{0{,}05^2} \ln\left(\frac{4((2 \cdot 183000)^{10} + 1)}{0{,}05}\right) \\
N &\geq 423954
\end{aligned}
$$

Volvemos a sustituir:

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4((2 \cdot 423954)^{10} + 1)}{0{,}05}\right)$$
$$N \geq 450839$$

Seguimos...

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4((2 \cdot 450839)^{10} + 1)}{0{,}05}\right)$$
$$N \geq 452806$$

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4((2 \cdot 452806)^{10} + 1)}{0{,}05}\right)$$
$$N \geq 452946$$

Como vemos, ya estamos convergiendo a 452000, luego para una dimensión de 10, el tamaño de los datos debe estar en torno a 452000. Según hemos estudiado, el tamaño de la dimensión guarda una proporción con la cantidad de datos en un factor de $10{,}000$, lo cual se cumple en este caso. Pero en la práctica esta costante se suele reducir a 10.

12. Ejercicio 12

13. Bonus 1

14. Bonus 2