Aprendizaje automático
Cuestionario

Alejandro García Montoro
agarciamontoro@correo.ugr.es

27 de abril 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 pájaros, 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.*

Solución. *Categorizar un grupo de animales vertebrados en pájaros, mamíferos, reptiles, aves y anfibios.*

La categorización automática de un grupo de animales vertebrados en sus diferentes clases puede abordarse como un problema de aprendizaje supervisado. Si tenemos una manera de describir un animal, ya sea con una imagen o, por ejemplo, una descripción cuantitativa de sus cualidades anatómicas, necesitaremos una muestra previamente clasificada para poder entrenar nuestro algoritmo de aprendizaje. Abordar este problema con aprendizaje no supervisado no tendría sentido, ya que tenemos una forma clara —aunque no definida analíticamente, si así fuera sería más sensato resolver este problema mediante diseño— de diferenciar entre clases; no usar esa información es perder oportunidades. Por último, el aprendizaje por refuerzo no tiene sentido en este caso; no hay definida una estructura de actión y recompensa.

Imaginemos ahora que la manera de describir a los animales es a partir de sus cualidades anatómicas. Si tenemos, por ejemplo, una serie de 30 características lógicas —como puede ser presencia o no de plumas, de pelo,

si son o no vivíparos...— que los describen, los datos de aprendizaje son los siguientes:

- $\mathcal{X} = \{(x_1, x_2, \ldots, x_{64}) \in \mathbb{Z}_2^{30}\}$
- $\mathcal{Y} = \{Pájaro, Mamífero, Reptil, Ave, Anfibio\}$

Clasificación automática de cartas por distrito postal.

Para clasificar automáticamente cartas por distrito postal basta leer el código postal escrito en el sobre; esto, evidentemente, se reduce a comprender cada uno de los dígitos escritos. Estos números suelen estar manuscritos, así que la diversidad de escrituras, tintas y papeles nos impiden intentar abordar este problema por diseño —no conocemos una regla exacta para asignar trazos manuscritos (que pueden tener formas diversas) a su correspondiente dígito—. Además, podemos tomar una gran muestra de dígitos manuscritos, clasificarlos a mano y usar esta información para intentar aprender bajo qué regla se asigna un trazo a un dígito concreto. Este caso es, por tanto, un claro ejemplo de aprendizaje supervisado con clasificación multi-etiqueta.

Imaginemos que tenemos imágenes escaneadas de dígitos manuscritos, de manera que después de tratarlas y procesarlas, se reducen s matrices 8x8 —o vectores de longitud 64— en las que cada entrada almacena la intensidad del trazo en ese punto, numerada de 0 a 255. Los datos de aprendizaje que debemos usar son entonces los siguientes:

- $\mathcal{X} = \{(x_1, x_2, \ldots, x_64) \in \mathbb{Z}_{256}^{64}\}$
- $\mathcal{Y} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

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

Este problema puede ser abordado con aprendizaje por refuerzo. La situación del mercado de valores puede no valer de un año a otro debido a la inestabilidad de las situaciones políticas, económicas y sociales, así que no es fácil generar una muestra de aprendizaje supervisado en la que conozcamos las etiquetas correctas para cada predicción.

Sin embargo, sí podemos recolectar predicciones anteriores, su nivel de éxito —cómo de cerca estuvieron de la posterior realidad— y generar una muestra de la forma (situación, predicción, nivel de éxito de la predicción), de manera que el algoritmo aprenda cómo de buena o mala —no exactamente si es la mejor— es una predicción. Además, esta estructura nos permitirá realimentar el algoritmo con el éxito de las predicciones que haga una vez se encuentre en producción, de forma paralela a su funcionamiento.

Si la situación del índice se puede describir con un número real, la predicción que se hace es otro número real y alimentamos el algoritmo con lo que pasó en los últimos 60 días, los datos de aprendizaje serán los siguientes:

- $\mathcal{X} = \{(x_1, x_2, \ldots, x_30, z) \in \mathbb{R}^{30} \times \mathbb{R}\}$

- $\mathcal{Y} = [0, 1]$, donde este número indica el valor de éxito de la predicción.

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.*

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

Este caso es el más adecuado para una aproximación por diseño. Podemos medir el nivel de tráfico medio en el cruce a lo largo del día, simular qué ocurre con cada uno de los ciclos disponibles y elegir el mejor según el criterio que se establezca.

No tendría sentido abordar este problema con una aproximación por aprendizaje, ya que tenemos una forma analítica de medir el impacto de cada ajuste de los parámetros del problema que queremos obtener.

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

Este caso, sin embargo, es más adecuado para una aproximación por aprendizaje. No existe una expresión analítica que nos diga, dados los datos de educación, edad, experiencia y estatus social, los ingresos medios de una persona. Sin embargo, es evidente que entre estas variables existe una relación.

Estamos ante un problema que requiere claramente de una aproximación por aprendizaje, en la que tendremos que tomar una muestra lo suficientemente amplia de personas, analizar sus datos e intentar aprender la relación que existe entre ellos.

Determinar si se debe aplicar una campaña de vacunación contra una enfermedad. Las enfermedades son diversas y tienen comportamientos muy inestables. Aunque quizás existan parámetros analíticos para intentar una aproximación por diseño, en este caso parece más sensato también intentar abordar el problema mediante aprendizaje.

Podríamos estudiar todos los casos documentados de enfermedades, analizar sus atributos, determinar si se aplicó o no una campaña de vacunación y cómo de exitosa fue la decisión. De esta manera, podríamos intentar aprender qué hacer ante la nueva enfermedad intentando aprender una relación entre sus atributos y los de enfermedades ya tratadas.

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.*

Solución. En este ejercicio tenemos que identificar $\mathcal{X}$, el conjunto de características medidas; $\mathcal{Y}$, el conjunto de etiquetas que podemos asignar a cada muestra; D, el conjunto de muestras de entrenamiento y f, la función objetivo.

En el problema queremos seleccionar qué fruta está en su punto óptimo de madurez y cuál no —ya que queremos escoger el producto que mejor salida tenga en el mercado después—. Si codificamos *maduro* como +1 y *no maduro* como -1, el conjunto de etiquetas será el siguiente:

$$\mathcal{Y} = \{+1, -1\}$$

Imaginemos ahora que el experto nos ha informado de que las siguientes características físicas son las que determinan la madurez de una pieza de fruta en concreto:

- Tamaño.
- Color.

Estas características las mediremos de la siguiente forma:

- El tamaño lo mediremos de forma continua, luego se mueve en $\mathbb{R}^+$
- El color lo mediremos con imágenes de las piezas de fruta, analizando después la media de los píxeles pertenecientes a la fruta en una escala de 256 valores en cada uno de los tres canales de color usuales: rojo, verde y azul. Por tanto, esta medida se mueve en $\mathbb{Z}_{256}^3$.

Así, concluimos que el conjunto de características es el siguiente:

$$\mathcal{X} = \mathbb{R}^+ \times \mathbb{Z}_{256}^3$$

Si en nuestra visita a la huerta tomamos 100 muestras de mangos, medimos su tamaño y color y analizamos, con ayuda del experto, si están o no maduros, nuestro conjunto de entrenamiento será de la siguiente forma:

$$\mathcal{D} = \{(x_i, y_i)\ i = 1, 2, \ldots, 100\ / x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$$

La finalidad del aprendizaje será encontrar la función objetivo, fjia pero desconocida, siguiente:

$$f \colon \mathcal{X} \to \mathcal{Y}$$

Tenemos así todos los elementos formales del aprendizaje. Para desarrollar el modelo completo, tendremos que seleccionar $\mathcal{H}$, el conjunto de hipótesis —esto deberíamos hacerlo antes de conocer las muestras— y $\mathcal{A}$, el algoritmo de aprendizaje.

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. [1]

Sea $(x(t), y(t))$, con $y(t) \in \{-1, +1\}$ la muestra mal clasificada respecto del modelo PLA; esto es, $sign(w^T(t)x(t)) \neq y(t)$ o, dicho de otra manera

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

Probar que la regla de adaptación de pesos del PLA es un movimiento en la dirección correcta para clasificar bien $x(t)$ consiste en determinar que la siguiente iteración del algoritmo estará más cerca de clasificar bien la muestra; esto es, tenemos que probar lo siguiente:

$$y(t)sign(w^T(t+1)x(t)) > y(t)sign(w^T(t)x(t)) \tag{1}$$

donde $w^T(t+1)$ es el vector de pesos que nos da la siguiente iteración del algoritmo, y que viene determinado como

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

Para probar la desigualdad 1 basta trabajar con la parte izquierda y la definición de $w^T(t+1)$ de la siguiente manera:

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

donde hemos usado que $y(t)^2 = 1$, pues $y(t) \in \{-1, +1\}$ y que $x^2(t) > 0$.

Hemos probado así que la regla del PLA permite acercarse al correcto etiquetado de la muestra.

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?*

Solución. *Si $p = 0{,}9$, ¿cuál es la probabilidad de que S produzca una hipótesis mejor que C?*

[1] Ejercicio resuelto con la ayuda de los pasos explicados en el Ejercicio 1.3 de [1].

En el ejercicio citado se intenta aprender la función

$$f\colon \mathcal{X} = \mathbb{R} \to \mathcal{Y} = \{-1, +1\}$$

con un conjunto de hipótesis $\mathcal{H} = \{h_1 \equiv +1, h_2 \equiv -1\}$. El algoritmo S elige la hipótesis que mejor ajusta los datos y el C justo la contraria. Además, se asume que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$.

Supuesto en este ejercicio que $p = P[f(x) = +1] = 0{,}9$, podemos estudiar cuáles es la probabilidad de que S produzca una hipótesis mejor que C estudiando la siguiente probabilidad:

$$P[E^S_{out} < E^C_{out}]$$

donde E^S_{out} indica el error fuera de la muestra de la hipótesis producida por S y E^C_{out} el error fuera de la muestra de la hipótesis producida por C.

Estos errores fuera de la muestra se corresponden con la probabilidad de que la hipótesis producida *no* sea la función objetivo f; es decir, tenemos la siguiente expresión:

$$P[E^S_{out} < E^C_{out}] = P[P[f \neq h_S] < P[f \neq h_C]]$$

donde h_S es la hipótesis producida por S y h_C la producida por C.

Como por hipótesis todas las muestras en $\mathcal{D}$ están etiquetadas como $+1$, S elegirá la función h_1 y C la h_2. Por tanto, la expresión anterior se puede escribir como:

$$P[E^S_{out} < E^C_{out}] = P[P[f \neq +1] < P[f \neq -1]]$$

Pero por hipótesis sabemos que $P[f(x) = +1] = p = 0{,}9$. Por tanto, tenemos que $P[f(x) \neq +1] = 1 - p = 0{,}1$ y $P[f(x) \neq -1] = P[f(x) = +1] = p = 0{,}9$. Hemos reducido nuestra expresión a lo siguiente, que ya podemos calcular:

$$P[E^S_{out} < E^C_{out}] = P[1 - p < p] = P[0{,}1 < 0{,}9] = 1 \tag{2}$$

ya que 0,1 es siempre menor estricto que 0,9.

Hemos probado entonces que si $p = 0{,}9$, es seguro que S producirá una hipótesis mejor que C.

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

Siguiendo el mismo razonamiento, llegamos a la expresión 2. Sin embargo, como queremos ver cuándo es más probable que C produzca una hipótesis mejor que S, tenemos que darle la vuelta a la desigualdad. Nuestro problema se reduce entonces a determinar el p de manera que

$$P[E^S_{out} > E^C_{out}] = P[1 - p > p]$$

Operando con la desigualdad $1 - p > p$ tenemos

$$\begin{aligned} 1 - p &> p \\ 1 &> 2p \\ p &< \frac{1}{2} \end{aligned}$$

Podemos concluir entonces que es más pobable que C produza una hipótesis mejor que S si y sólo si $p < 0{,}5$.

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

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

y estudiar lo que se nos pide. Tenemos entonces:

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

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

$$N_M = \left\lceil \frac{1}{-2\varepsilon^2} ln\left(\frac{k}{2M}\right) \right\rceil \tag{3}$$

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 3, lo que nos da los siguientes valores:

$$\begin{aligned} N_1 &= 840 \\ N_{10} &= 1301 \\ N_{100} &= 1761 \end{aligned}$$

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 probar que no hay ninguna muestra de ese tamaño de manera que $\mathcal{H}$ sea capaz de separar por completo.

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 implementar con $\mathcal{H}$ 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.

Podríamos pensar que esto prueba que existe una única muestra que $\mathcal{H}$ no puede separar. Sin embargo, cualquier muestra de $2M + 1$ puntos en la recta real es de la forma descrita, y en todas y cada una de ellas existe al menos esta dicotomía que $\mathcal{H}$ no puede implementar.

Tenemos así que cualquier muestra de $2M + 1$ puntos tiene al menos 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. En este caso no conocemos nada acerca del punto de ruptura. Por definición, k es un punto de ruptura si $m_{\mathcal{H}}(k) < 2^k$. Esto no quiere decir que *exista* una muestra de k puntos para la que H no sea capaz de implementar todas las dicotomías; sino que $\mathcal{H}$ es incapaz de implementar todas las dicotomías *para todas* las muestras de k puntos.

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.*

Solución. En este caso ya sí podemos afirmar, usando el mismo argumento que en el ejercicio anterior, que k^* es un punto de ruptura.

Por otro lado, es evidente que si k^* es un punto de ruptura, $k \geq k^*$ también lo es. Esto es claro: si $\mathcal{H}$ es incapaz de separar una muestra de k^* puntos, añadir puntos a la muestra no hará sino hacerla más compleja, luego $\mathcal{H}$ seguirá siendo incapaz de separarla.

Por último, poco podemos decir sobre los $k < k^*$. Pueden ser puntos de ruptura, si se da el caso de que k^* no sea el más pequeño punto de ruptura —basta usar el mismo argumento que en el párrafo anterior—, o pueden no serlo, como en el caso sencillo del perceptron en el plano, donde sabemos que $k^* = 4$ es un punto de ruptura y $k = 3 < k^*$ no lo es.

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 $m_{\mathcal{H}}(k) = 2^{k^*}$*

Solución. Por definición, k^* es un punto de ruptura si $m_{\mathcal{H}}(k^*) < 2^{k^*}$. Como la función de crecimiento está definida en términos del máximo número de dicotomías que H puede implementar, podemos concluir que para probar que k^* es un punto de ruptura hay que mostrar que $\mathcal{H}$ no puede separar *ningún* conjunto de k^* puntos.

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?*

Solución. La cota de generalización nos dice que, para tener una confianza del $1 - \delta$ de que el error de generalización sea como mucho ε, tenemos que tomar un número de muestras N tal que cumpla la siguiente desigualdad:

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

Una forma poco elegante, pero rápida de implementar en este caso, de conseguir tal número es iterando sobre N y comprobar si la desigualdad se satisface. El siguiente script en Python nos es suficiente para determinar el menor N necesario para tener un 95 % de confianza —como $1 - \delta = 0{,}95$, entonces $\delta = 0{,}05$— de que el error de generalización sea como mucho $\varepsilon = 0{,}05$:

```
from math import log

# Define las constantes del ejercicio
eps = 0.05
delta = 0.05
dVC = 10

# Iteramos desde 1 hasta 500.000
for N in range(1, 500000):
    rhs = (8 / eps**2) * log((4*(2*N)**dVC + 4) / delta)
```

```
11
12          if(N >= rhs):
13              # Si se cumple la desigualdad, imprime N y sal del bucle
14              print(N)
15              break
```

Así, obtenemos que el número de muestras tiene que ser mayor o igual que 452957.

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

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

La función promedio se puede calcular como sigue:

$$\bar{g}(x) = \mathbb{E}_{\mathcal{D}_n}[g^{\mathcal{D}_n}(x)] \approx \frac{1}{K}\sum_{n=1}^{K} g^{\mathcal{D}_n}(x)$$

donde $g^{\mathcal{D}_n}(x)$ es la función que devuelve el algoritmo respecto del n-ésimo conjunto de datos $\mathcal{D}_n$.

Llamando $\mathcal{D}_n = \{(a_n, a_n^2), (b_n, b_n^2)\}$, sabemos la expresión explícita de $g^{\mathcal{D}_n}(x)$, ya que no es más que la recta :

$$g^{\mathcal{D}_n}(x) = m_n x + c_n$$

donde m_n es la pendiente de la recta y c_n el punto de corte con el eje vertical; es decir:

$$m_n = \frac{b_n^2 - a_n^2}{b_n - a_n}$$
$$c_n = a_n^2 - \frac{a_n(b_n^2 - a_n^2)}{b_n - a_n}$$

Tenemos entonces la siguiente expresión de la función promedio:

$$\bar{g}(x) \approx \frac{1}{K}\sum_{n=1}^{K}(m_n x + c_n) = \frac{1}{K}\left(\sum_{n=1}^{K} m_n\right)x + \frac{1}{K}\sum_{n=1}^{K} c_n \tag{4}$$

Esto es, la función promedio es aproximadamente la recta cuya pendiente es la media de las pendientes entre cada par de puntos generados y cuyo corte con el eje vertical es la media de los cortes con el eje vertical de las rectas generadas para cada par de puntos.

El *aproximadamente* lo podemos eliminar, y por tanto tener una igualdad, si consideramos todos los conjuntos $\mathcal{D}_n$ posibles. Como los datos están uniformemente distribuidos en $[-1, 1]$, tanto la media de las pendientes como la media de los puntos de corte será 0 —ya que la media de una muestra uniforme en $[-1, 1]$ y las operaciones en m_n y c_n no alteran la media—.

Podemos concluir entonces que

$$\bar{g}(x) = 0x + 0 = 0 \quad \forall x$$

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 $f(x)$.*

Solución. Podemos desarrollar un experimento sencillo atendiendo a la expresión de la función promedio obtenida en 4. Así, basta tomar una muestra lo suficientemente grande de una distribución uniforme en el intervalo $[-1, +1]$, tomar sus cuadrados y generar K conjuntos $\mathcal{D}_n$.

Con cada uno de estos conjuntos podemos calcular $g^{\mathcal{D}_n}(x)$; esto es, la pendiente y punto de corte con el eje vertical de la recta obtenida con los dos puntos de $\mathcal{D}_n$. Por 4, basta luego tomar la media de las pendientes y de los puntos de corte y reconstruir así $\bar{g}(x)$.

Referencias

[1] Abu-Mostafa, Yaser S., Malik Magdon-Ismail y Hsuan Tien Lin: *Learning from data : a short course*. AMLBook.com, United States, 2012.