Aprendizaje automático. Cuestionario de teoría 1

Jacinto Carrasco Castillo

29 de marzo 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.

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 $x(t)$.

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?

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

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

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

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?

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.

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.

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

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?

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**