APRENDIZAJE AUTOMÁTICO

****

TRABAJO 1.
PREGUNTAS TEORÍA.

****

Autora: Cristina Zuheros Montes.

- Correo: zuhe18@gmail.com
- Github: https://github.com/cristinazuhe

Fecha: 03 Abril 2016

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 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.

En este caso lo mejor será aprendizaje supervisado ya que necesitamos categorizar la salida en el tipo de animal del que se trate. Como datos de aprendizaje podremos usar: número de extremidades, tiene o no tiene pico, tiene o no tiene plumas, tiene o no tiene alas, tienen o no tiene cola, tacto, es o no migrador, tipo de alimentación, tamaño medio, velocidad movimiento….

b) Clasificación automática de cartas por distrito postal

Si nos ceñimos a la tarea, vemos no nos están dando ningún objetivo de salida categorizada, lo que nos lleva a pensar que lo mejor será un aprendizaje no supervisado. Los datos de aprendizaje serían los dígitos del distrito postal. Con ellos podríamos ver una nube de puntos con las muestras de las cartas. En esta nube podríamos distinguir zonas, que nos indicarían que dichas cartas están destinadas a una zona próxima, pero no sabremos la zona exacta. No nos lo están pidiendo.

Si nos lo pidieran, es decir, si en la clasificación automática nos piden que le asignemos una ciudad, una provincia….tendríamos que utilizar aprendizaje supervisado. Al igual que antes los datos de aprendizaje serían los dígitos del distrito postal, pero además conoceríamos su etiqueta que sería su ciudad, provincia…o la salida específica que nos pidieran.

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

Para este caso usaríamos aprendizaje por refuerzo ya que nos interesa ir simulando acciones y ver qué es lo que más nos interesa en el futuro. Los datos de aprendizaje serían los índices del mercado de valores. La idea es que si, por ejemplo, un índice de mercado baja en un periodo de tiempo, posteriormente vuelve a bajar, y así repetidas veces, pues vemos que éste no nos interesa.

2. ¿Cuales 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.

Usaremos aproximación por aprendizaje ya que necesitamos conocer bien las características de las muestras de entrada y sus correspondientes etiquetas.

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

Proponemos usar aproximación por diseño, podemos obtener toda la información de la persona y sacar el nivel medio de ingresos. No tenemos porqué conocer los datos de cada persona para aproximar f.

c) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.

En este caso podremos usar aproximación por diseño ya que no tenemos porqué conocer bien cada uno de los datos de la muestra. Basta con identificar zonas de riesgo que nos digan si interesa vacunar o no contra esa enfermedad. Podemos calcular f analíticamente.

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.

Supongamos que la fruta en cuestión son los mangos (como el ejemplo que viene en las transparencias de clase). Los elementos formales que vamos a pasar a describir a continuación, nos servirán para cualquier otro tipo de fruta, luego no estamos limitando el problema que se nos plantea.

Las etiquetas a usar, nuestra salida, será binaria Y={-1,1} donde -1 indicará que la fruta (mango) no será sabroso, mientras que 1 indica que el mango sí es sabroso. Estamos sólamente evaluando cómo de sabroso es un mango. Si queremos evaluar otro aspectos como podrían ser si el mango es o no dulce, si está o no maduro...tendríamos que trabajar con una clasificación multi-etiqueta. Ahora mismo nos quedaremos con la clasificación binaria que nos dice si los mangos son o no sabrosos, ya que los modelos que conocemos son para clasificaciones binarias.

Para asociarle a un mango una etiqueta -1 o 1 (no sabroso o sabroso), vamos a tener en cuenta un espacio de características X que podría estar formado por distintas características físicas del mango como son: peso, color, forma, textura, zona de crecimiento del mango…

La idea es seleccionar un conjunto de mangos como datos de entrenamiento y medirles las características de nuestro conjunto X, así como asociarle la etiqueta Y correspondiente según el mango esté o no sabroso. Si, por ejemplo, el mango es pequeño, verde y se cultivó en la zona B, el mango está sabro. Pero esta target function que nos asocia los valores de X con los valores de Y será desconocida.

Ya tenemos un conjunto de datos de entrenamiento etiquetados. Podremos el usar el algoritmo PLA para hacer la clasificación binaria. De modo que cuando nos den algún otro mango para saber si es o no sabroso, representamos sus características y vemos en qué parte del hiperplano que nos ha dado PLA queda.

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).

Supongamos, sin pérdida de generalidad, que x(t)=(1,x1,x2) es el punto que se encuentra mal clasificado con respecto al hiperplano que haya encontrado PLA definido por w=(w0,w1,w2).

- Supongamos que x(t) tiene como etiqueta el valor y=+1:

Esto quiere decir, que la clasificación por el hiperplano nos da como etiqueta -1:
$(w0\ w1\ w2)^t*(1\ x1\ x1) = w0+w1*x1+w2*x2$ tiene signo negativo → -1 (1)

Al hacer la siguiente iteración del algoritmo PLA , tenemos que el nuevo hiperplano queda definido como:
wnew = (w0n, w1n, w2n) = (w0, w1, w2) + 1(1, x1, x2) =(w0 + 1, w1+x1, w2+x2)
Veamos que ahora sí que tendríamos el punto bien clasificado:
$(w0 + 1, w1+x1, w2+x2)^t*(1\ x1\ x2)= w0 + 1 + w1*x1 + x1*x1 + w2*x2 + x2*x2$
Haciendo uso de (1), nos quedaría:
$(w0 + 1, w1+x1, w2+x2)^t*(1\ x1\ x2) = x1*x1 + x2*x2$ tiene signo positivo → +1 y el dato queda bien clasificado.

- Supongamos que x(t) tiene como etiqueta el valor y=-1:

Esto quiere decir, que la clasificación por el hiperplano nos da como etiqueta +1:
$(w0\ w1\ w2)^t*(1\ x1\ x1) = w0+w1*x1+w2*x2$ tiene signo positivo → +1 (2)

Al hacer la siguiente iteración del algoritmo PLA , tenemos que el nuevo hiperplano queda definido como:
wnew = (w0n, w1n, w2n) = (w0, w1, w2) - 1(1, x1, x2) =(w0 - 1, w1-x1, w2-x2)
Veamos que ahora sí que tendríamos el punto bien clasificado:
$(w0 - 1, w1-x1, w2-x2)^t*(1\ x1\ x2)= w0 - 1 + w1*x1 - x1*x1 + w2*x2 - x2*x2$
Haciendo uso de (), nos quedaría:
$(w0 - 1, w1-x1, w2-x2)^t*(1\ x1\ x2) = - x1*x1 - x2*x2$ tiene signo negativo→ -1 y el dato queda bien clasificado.

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?

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

$$P[|Eout(g) - Ein(g)| > \varepsilon] \leq 2Me^{-2\varepsilon^2 N} \text{ para cualquier } > 0.$$

Si fijamos = 0,05 y queremos que la cota probabilística $2Me^{-2\varepsilon^2 N}$ sea como máximo 0.03 ¿cual será el valor más pequeño de N que verifique estas condiciones si M = 1?. Repetir para M = 10 y para M = 100

- Para M=1:

$P[|Eout(g) - Ein(g)| > \varepsilon] \leq 2e^{-0.005N}$. Queremos que la cota probabilística sea como máximo 0.03, luego tendremos:

$0.03 \leq 2e^{-0.005N} \rightarrow 0.015 \leq e^{-0.005N} \rightarrow \ln(0.015) \leq -0.005N \rightarrow 4.1997 \geq 0.005N \rightarrow 839.9 \geq N$

En definitiva tenemos que el valor más pequeño de N es 840.

- Para M=10:

$P[|Eout(g) - Ein(g)| > \varepsilon] \leq 20e^{-0.005N}$. Queremos que la cota probabilística sea como máximo 0.03, luego tendremos:

$0.03 \leq 20e^{-0.005N} \rightarrow 0.0015 \leq e^{-0.005N} \rightarrow \ln(0.0015) \leq -0.005N \rightarrow 6.50229 \geq 0.005N \rightarrow 1300.45 \geq N$

En definitiva tenemos que el valor más pequeño de N es 1301.

- Para M=100:

$P[|Eout(g) - Ein(g)| > \varepsilon] \leq 200e^{-0.005N}$. Queremos que la cota probabilística sea como máximo 0.03, luego tendremos:

$0.03 \leq 200e^{-0.005N} \rightarrow 0.00015 \leq e^{-0.005N} \rightarrow \ln(0.00015) \leq -0.005N \rightarrow 8.80487 \geq 0.005N \rightarrow 1760.97 \geq N$

En definitiva tenemos que el valor más pequeño de N es 1761.

7. Consideremos el modelo de aprendizaje "M-intervalos"donde h : R → {−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?

El más pequeño punto de ruptura k sería 2M+1:

Para el problema de 1-intervalo, ya hemos visto en clase que k=3. Basta considerar:

\+ - +

Para el problema de 2-intervalos, sería k=5. Basta considerar:

\+ - + - +

Vamos viendo que para el problema de M-intevalos, sería k=2M+1. Basta considerar:

\+ - + … + - + (M negativos intercalados)

8. Suponga un conjunto de k* puntos x1, x2, . . . , xk* sobre los cuales la clase H implementa < 2^{k*} dicotomías. ¿Cuales 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

En este caso la opción correcta es la e), no podemos saber nada sobre el punto de ruptura ya que sólo sabemos qué pasa con un conjunto de k* puntos, necesitaríamos que se cumpliese para todo conjunto de k* puntos.

9. Para todo conjunto de k * puntos, H implementa < 2^{k*} dicotomías. ¿Cuales 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 ≥ k* son puntos de ruptura

d) todos los k < k* son puntos de ruptura

e) no conocemos nada acerca del punto de ruptura

Las afirmaciones correctas son la a) y la c). Sabemos que a) es cierta por definición de punto de ruptura. Además c) es cierto, pues cualquier k superior a un punto de ruptura será un punto de ruptura.

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

a) Mostrar que existe un conjunto de k* puntos x1, . . . , xk* que H puede separar ("shatter").

b) Mostrar que H puede separar cualquier conjunto de k * puntos.

c) Mostrar un conjunto de k* puntos x1, . . . , xk* que H no puede separar

d) Mostrar que H no puede separar ningún conjunto de k* puntos

e) Mostrar que mH(k) = 2^{k*}

Concepto: if for some value k, $m_{\mathcal{H}}(k) < 2^k$, then k is a break point for $\mathcal{H}$

La afirmación que nos serviría sería la d). Tendríamos que mostrar que H no puede separar ningún conjunto de k* puntos.

Si hay un sólo conjunto de k* puntos que H puede separar (opción a) , entonces se tendría mH(k) = 2^{k*} (opción e), pero esto implica que k* no es un punto de ruptura. Luego estas dos opciones no nos serviría. Además la opción b tampoco nos sirve, pues para ver el punto de ruptura tendremos que ver que H no puede separar ciertos conjuntos, pero no que sí los pueda separar.

Finalmente, la opción c tampoco nos serviría pues no nos basta con tener un sólo conjunto de k* puntos que no separe, sino que hace falta que no separe ninguno (opción d, que sería la que correcta).

11. Para un conjunto H con $d_{VC} = 10$, ¿que 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?

Vamos a hacer uso de la fórmula vista en clase:

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

En nuestro caso $\varepsilon = 0.05$ $\delta = 0.05$**. Para obtenemos N hemos creado el siguiente código en R:**

```
N=1
total=2
epsilon= 0.05
delta= 0.05
dvc= 10

while(N < total){
 pri = 8/((epsilon)^(2))
 num = 4*((2*N)^dvc + 1)
 n_log = log(num/delta)
 total = pri*n_log
 N=N+1
}
print("N encontrado!")
```

Y el resultado obtenido es un tamaño muestral de N= 452957.

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 {x1, x2} y que la función objetivo es $f(x) = x^2$. Por tanto es conjunto de datos completo es D = {(x1, x2 1),(x2, x2 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 ḡ(x).

b) Calcular analiticamente los valores de Eout, bias, y var.