Cuestionario 2
Aprendizaje Automático

Alumno: Christian Vigil Zamora
DNI:
Curso: 3º

27 de agosto de 2019

1. Identificar de forma precisa dos condiciones imprescindibles para que un problema de predicción pueda ser aproximado por inducción desde una muestra de datos. Justificar la respuesta usando los resultados teóricos estudiados.

Solución:
Para que un problema de predicción pueda ser aproximado por inducción desde una muestra de datos, se deben cumplir las siguientes dos condiciones imprescindibles:

- La primera condición sería que los datos ($\mathcal{D}$) deben estar formados por muestras que deben haber sido elegidas de forma independiente e idénticamente distribuidas en una distribución de probabilidad ($\mathcal{P}$), **i.i.d**.

- La segunda condición sería que el valor de la dimensión de Vapnik&Chervonenkis, es decir d_{vc} debe ser finito, para así obtener una buena generalización, ya que si d_{vc} fuese infinito, no sería posible usar ERM durante el aprendizaje.

2. El jefe de investigación de una empresa con mucha experiencia en problemas de predicción de datos tras analizar los resultados de los muchos algoritmos de aprendizaje usados sobre todos los problemas en los que la empresa ha trabajado a lo largo de su muy dilatada existencia, decide que para facilitar el mantenimiento del código de la empresa van a seleccionar un único algoritmo y una única clase de funciones con la que aproximar todas las soluciones a sus problemas presentes y futuros. ¿Considera que dicha decisión es correcta y beneficiará a la empresa? Argumentar la respuesta usando los resultados teóricos estudiados.

Solución:
Considero que dicha decisión es incorrecta y no beneficiará a la empresa, puesto que si tenemos en cuenta el **Teorema No-Free-Lunch (NFL)**, para todo algoritmo, siempre existe una muestra en la que falla, y no existe un modelo universal que se adapte de la mejor forma a todos los problemas. A la hora de afrontar un problema, la empresa debería plantear diferentes modelos, con diversas clases de funciones y decidir cuál se ajusta mejor a ese problema. Posteriormente, probar dicho modelo con diferentes algoritmos y evaluar cuál ofrece un mejor rendimiento. Si la empresa lleva a cabo su decisión, a la mínima que se enfrente a diferentes problemas, se evidenciará que no conlleva ningún beneficio, pues cada problema requiere unas restricciones diferentes o incluso a la hora de diseñar el modelo, se hacen suposiciones, de forma que si tenemos exclusivamente una única clase de funciones y un único algoritmo, lo más probable es que tarde o temprano de con problemas cuyas restricciones o suposiciones no pueda satisfacer.

3. ¿Que se entiende por una solución PAC a un problema de aprendizaje? Identificar el porqué de la incertidumbre e imprecisión.

Solución:
Una solución PAC, cuyas siglas significas "Probably Aproximately Correct", aplicada a un problema de aprendizaje vendría a ser el proceso por el que se trata de buscar una hipótesis que dada una muestra, la probabilidad de que el Error fuera de la muestra menos el Error dentro de la muestra sea menor que un valor Epsilon (ϵ), debe ser mayor o igual que 1 - Delta (δ). Para entender todo lo comentado, voy a mostrar un resultado PAC, el cual viene dado por la desigualdad de Hoeffding:
.
$$P\left(\mathcal{D} : |\mathrm{E}_{\mathrm{out}}(h) - \mathrm{E}_{in}(h)| < \epsilon\right) \geq 1 - \delta$$
considerando $\delta = 2e^{-2\epsilon^2 N}$. Esa expresión quiere decir que con una probabilidad de 1 - δ como mínimo, se va a cumplir que el error fuera de la muestra $\mathrm{E}_{\mathrm{out}}$ menos el error dentro de la propia muestra E_{in}, es decir, la diferencia entre ambos, es menor que ϵ.

- La incertidumbre en las soluciones PAC, identificada por δ, se debe a que como bien indica el significado de PAC, la solución es probablemente correcta, es decir, nunca tenemos la total certeza de que la solución es correcta, de ahí la incertidumbre.

- La imprecisión, identificada por ϵ aparece en las soluciones PAC puesto que según el problema sobre el que se aplique, tendremos una muestra de mayor o menor calidad, lo que hace que los errores dentro y fuera de la muestra puedan diferir demasiado.

4. Suponga un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraidos de una funcion desconocida $f : \mathcal{X} \to \mathcal{Y}$, donde $\mathcal{X} = \mathcal{R}$ e $\mathcal{Y} = \{-1, +1\}$. Para aprender f usamos un conjunto simple de hipótesis $\mathcal{H} = \{h_1, h_2\}$ donde h_1 es la función constante igual a $+1$ y h_2 la función constante igual a -1. Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). S elige la hipótesis que mejor ajusta los datos y C elige deliberadamente la otra hipótesis.
a) ¿Puede S producir una hipótesis que garantice mejor comportamiento que la aleatoria sobre cualquier punto fuera de la muestra? Justificar la respuesta

Solución:
a) No es posible garantizar que S va a producir una hipótesis con mejor rendimiento que C fuera de la muestra. Es obvio que con el conjunto de entrenamiento, el algoritmo S ofrece un mejor rendimiento que C, pues selecciona la hipótesis que mejor ajusta los datos, en éste caso, la hipótesis cuyo valor se corresponde con la etiqueta de la mayoría del conjunto de entrenamiento, pero eso no quiere decir que fuera de la muestra ocurra igual. Perfectamente, la muestra de test podría

contener una mayoría de puntos etiquetados con un valor contrario al fijado por la hipótesis elegida por S, y en ese caso C dar un mejor rendimiento. Aún considerando que el conjunto de datos datos fuera de la muestra han sido elegidos de forma independiente e idénticamente distribuidos, seguiríamos sin poder garantizar que S produce una hipótesis de mejor comportamiento, como mucho que la probabilidad de que S produzca un mejor comportamiento que C es mayor.

5. Con el mismo enunciado de la pregunta 4:
a) Asumir desde ahora que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$. Es posible que la hipótesis que produce C sea mejor que la hipótesis que produce S? Justificar la respuesta

Solución:
a) Sí, si es posible. El hecho de que todos los datos en la muestra de entrenamiento tengan la etiqueta +1, conlleva que el algoritmo elegirá evidentemente la hipótesis h_1. Por lo tanto, en la muestra de entrenamiento, el error E_{in} que obtendrá S será 0, mientras que el error E_{in} que obtendrá C siempre será del 100 %. Hasta ahí bien, pero lo que me ha llevado a responder que sí es posible es que la muestra de entrenamiento $\mathcal{D}$ probablemente no represente adecuadamente a la población, es decir, es posible que los únicos datos en toda la población que tienen etiqueta +1 hayan sido seleccionados en $\mathcal{D}$, de esa forma, cuando se evalúan x puntos fuera de la muestra de entrenamiento, la hipótesis que produce C es totalmente mejor que la que produce S. Y ese es sólo un ejemplo de los tantos que existen. En cuánto al aprendizaje, podríamos decir que prácticamente no sería posible puesto que los elementos de $\mathcal{D}$ nos hace ver que no han sido elegidos mediante la misma distribución de probabilidad que la muestra de test, condición fundamental también en problemas de éste tipo.

6. Considere la cota para la probabilidad de la hipotesis solución g de un problema de aprendizaje, a partir de la desigualdad generalizada de Hoeffding para una clase finita de hipótesis, $\mathbb{P}\left[|E_{out}(g) - E_{in}(g)| > \epsilon\right) < \delta$
a) ¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?
b) Si elegimos g de forma aleatoria ¿seguiría verificando la desigualdad?
c) ¿Depende g del algoritmo usado?
d) Es una cota ajustada o una cota laxa?
Justificar las respuestas.

Solución:

a) Para elegir g, se usa el algoritmo ERM (Empirical Risk Minimization), cuyo criterio es elegir la g que más minimiza el Error Empírico. El Error Empírico viene dado por la expresión: $R_{\text{emp}}(h) = \frac{1}{n}\sum_{i=1}^{n} D\left(h\left(x_i\right), y_i\right)$

b) Sí, seguiría verificando la hipótesis puesto que sabemos que g es elegida del conjunto finito de Hipótesis $\mathcal{H}$, por lo tanto siempre es cierto que
$|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \rightarrow |E_{\text{in}}\ (h_1) - E_{\text{out}}\ (h_1)\,| > \epsilon$
. or $|E_{\text{in}}\ (h_2) - E_{\text{out}}\ (h_2)\,| > \epsilon$
. $\cdots$ or $|E_{\text{in}}\ (h_M) - E_{\text{out}}\ (h_M)\,| > \epsilon$
donde B1 $\rightarrow$ B2 significa que el evento B1 implica B2, lo que en términos de probabilidad significa que $\mathbb{P}[B1] \leq \mathbb{P}[B2]$.
Si juntamos las reglas comentadas anteriormente tenemos que:
$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \mathbb{P}[|E_{\text{in}}\ (h_1) - E_{\text{out}}\ (h_1)\,| > \epsilon]$
. or $\mathbb{P}[|E_{\text{in}}\ (h_2) - E_{\text{out}}\ (h_2)\,| > \epsilon]$
. $\cdots$ or $\mathbb{P}[|E_{\text{in}}\ (h_M) - E_{\text{out}}\ (h_M)\,| > \epsilon]$,
en resumen: $\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}\ (h_m) - E_{\text{out}}\ (h_m)\,| > \epsilon\right]$.
Dicho ésto, si aplicamos la Desigualdad de Hoeffding sobre los M términos de una vez, obtenemos la expresión: $\mathbb{P}\left[|E_{\text{in}}\,(g) - E_{\text{out}}\,(g)| > \epsilon\right] \leq 2Me^{-2\epsilon^{-2}\text{N}}$ la cuál es una versión uniforme de la expresión de la Desigualdad de Hoeffding en función del Error que nos permite saber que no importa la h que seleccionemos como g.

c) Sí, g depende del algoritmo, lo evidencia el libro de teoría en el que se basa apoya la asignatura *"With multiple hypotheses in $\mathcal{H}$, the learning algorithm picks the final hypothesis g bases on $\mathcal{D}$, i.e. after generating the data set"*

d) Es una cota laxa. Tal y como indica el enunciado, nos encontramos ante un conjunto finito de hipótesis, por tanto la desigualdad de Hoeffding la podemos expresar de la siguiente forma: $\mathbb{P}\left[|E_{\text{in}}\,(g) - E_{\text{out}}\,(g)| > \epsilon\right] \leq 2Me^{-2\epsilon^2 N}$, dónde δ es la parte izquierda derecha de la ecuación, y como vemos está en función de M, siendo M el número de hipótesis de $\mathcal{H}$. El motivo por el que es una cota laxa es debido a que el límite de probabilidad $\mathbb{P}\left[|E_{\text{in}}\,(g) - E_{\text{out}}\,(g)| > \epsilon\right] \leq 2Me^{-2\epsilon^2 N}$ es un factor de M más suelto que si sólo tuviésemos una hipótesis.

7. ¿Por qué la desigualdad de Hoeffding definida para clases $\mathcal{H}$ de una única función no es aplicable de forma directa cuando el número de hipótesis de $\mathcal{H}$ es mayor de 1? Justificar la respuesta.

Solución:
La desigualdad de Hoeffding sólo es aplicable de forma directa cuando $\mathcal{H} = \{\text{h}\}$ es decir, cuando sólo existe una única hipótesis. Además, dicha hipótesis h es fijada antes de conocer la muestra de datos. De ésta forma, la desigualdad de Hoeffding se expresa mediante la ecuación: $P\left(\mathcal{D} : |\text{E}_{\text{out}}(h) - \text{E}_{\text{in}}(h)| > \epsilon\right) \leq 2e^{-2\epsilon^2 N}$.
Por lo tanto, si el número de hipótesis es mayor de 1, lo anterior no se cumple e imposibilita que se aplique la desigualdad de forma directa. En los casos en los que el número de hipótesis es mayor de 1, el algoritmo de aprendizaje selecciona la hipótesis final g, la cuál está basada en la muestra de datos, es decir, se selecciona después de conocer la muestra de datos, de ahí que no pueda aplicarse de forma directa la desigualdad. Cuando el número de hipótesis es mayor de 1, el estado que nos gustaría producir no es $\mathbb{P}\left[|E_{\text{in}}\left(h_m\right) - E_{\text{out}}\left(h_m\right)| > \epsilon\right]$ es pequeña"para una $h_m \in \mathcal{H}$ particular, sino "$\mathbb{P}\left[|E_{\text{in}}\left(g\right) - E_{\text{out}}\left(g\right)| > \epsilon\right]$ es pequeña"para la hipótesis final g.

8. Si queremos mostrar que k^* es un punto de ruptura para una clase de funciones $\mathcal{H}$ cuales de las siguientes afirmaciones nos servirían para ello:
a) Mostrar que existe un conjunto de k^* puntos $x_1, \ldots, x_k$. que $\mathcal{H}$ puede separar "shatter").
b) Mostrar que $\mathcal{H}$ puede separar cualquier conjunto de k^* puntos.
c) Mostrar un conjunto de k^* puntos $x_1, \ldots, x_k$. que $\mathcal{H}$ no puede separar
d) Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de k^* puntos
e) Mostrar que $m_{\mathcal{H}}(k) = 2^{k^*}$

Solución:
a) Incorrecto. Ésta afirmación no nos serviría pues presenta una incoherencia. Si queremos demostrar que k^* es un punto de ruptura, la teoría nos dice que "-*That is,* $\mathcal{H}$ *CANNOT shatter a sample of size* k", es decir, que si es un punto de ruptura, no puede separar un conjunto de k^* puntos.

b)Incorrecto. De hecho, es justo lo contrario, si k^* es un punto de ruptura, no puede separar ningún conjunto de k^* puntos, tal y como indica la teoría citada en el apartado anterior.

c) Incorrecto. El hecho de que $\mathcal{H}$ no pueda separar un conjunto de k^* puntos, no nos sirve para mostrar que k^* es un punto de ruptura, puesto que podría

existir otro que conjunto que $\mathcal{H}$ sí pueda separar, y para que k^* sea punto de ruptura, necesitamos que $\mathcal{H}$ no pueda separar **NINGÚN** conjunto de k^* puntos.

d) Correcto. Ésta afirmación sí nos serviría, pues indica justo lo que estamos buscando para mostrar que k^* es un punto de ruptura. El hecho de que $\mathcal{H}$ no pueda separar ningún conjunto de k^* puntos, nos indica que la condición $m_{\mathcal{H}}(k) < 2^k$ se cumple y que por tanto, k^* es un punto de ruptura.

e) Incorrecto. Para que k^* sea punto de ruptura, la condición $m_{\mathcal{H}}(k) < 2^k$ debe cumplirse, y en éste apartado no sucede eso. Es más, según la condición de éste apartado, $\mathcal{H}$ separaría el conjunto de k^* puntos, cosa que como ya sabemos, es incorrecta si consideramos k^* como punto de ruptura.

9. Para un conjunto $\mathcal{H}$ con $d_{VC} = 10$, ¿que tamaño muestral se necesita (según la cota de generalización) para tener un 95 % de confianza (δ) de que el error de generalización (ϵ) sea como mucho 0.05?

Solución:
Primero, obtengo el valor de δ. Puesto que el enunciado nos indica que el nivel de confianza es del 95 % sabemos que $\delta = 1$ - NC, por lo tanto $\delta = 1 - 0{,}95 = 0{,}05$. A continuación, para obtener el tamaño muestral necesario, voy a utilizar la cota de generalización, dejando la ecuación en función de N: $N \geq \frac{8}{\varepsilon^2} \ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)$, de la que conocemos $\delta = 0{,}05$ y puesto que el error de generalización debe ser como mucho 0.05 también sabemos $\epsilon = 0{,}05$. N va a ser el tamaño muestral que necesitamos y $m_{\mathcal{H}}(2N)$ es la función de crecimiento (Growth Function) la cuál dado un tamaño de muestra N y una clase H, devuelve el número máximo de formas distintas en que N puntos pueden ser clasificados usando H. Por la dimensión-VC de Vapnik&Chervonenkis sabemos que podemos expresar la función de crecimiento mediante 2 límites. En éste caso vamos a utilizar $N^{d_{VC}} + 1$. Por tanto, si sustituimos los datos que conocemos en la ecuación comentada anteriormente, obtenemos que $N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4\left((2N)^{10}+1\right)}{0{,}05}\right)$. Como vemos, nos queda una ecuación ímplicita en N, es decir vamos a resolver la ecuación de forma iterativa, partiendo de un valor N inicial, el cuál se irá actualizando en cada iteración. Tras múltiples iteraciones, llegamos a la conclusión de que se necesita un tamaño muestral de N = 452.957 para tener un 95 % de confianza de que el error de generalización sea como mucho 0.05. He de decir que para obtener el resultado de éste ejercicio, programé un script en Python para el que primeramente obtuve que el tamaño de N debía ser aproximadamente 460.000, pero posteriormente afiné aún más para obtener el valor de N exacto.

10. Considere que le dan una nuestra de tamaño N de datos etiquetados $\{-1, +1\}$ y le piden que encuentre la función que mejor ajuste dichos datos. Dado que desconoce la verdadera función f, discuta los pros y contras de aplicar los principios de inducción ERM y SRM para lograr el objetivo. Valore las consecuencias de aplicar cada uno de ellos.

Solución:
Primeramente, voy a empezar comentando las ventajas y desventajas del principio de inducción **ERM**:
Ventajas:

- Nos permite elegir seleccionar una hipótesis cercana a f, la cuál mejor minimiza el error empírico.
- Si la dimensión VC es finita, obtenemos una buena generalización.
- Si el valor de N es suficientemente grande, podremos conseguir que el error fuera de la muestra $((E_{\text{in}}))$ sea muy similar al error dentro de la muestra (E_{in}), y como hemos dicho en la primera ventaja, se elige la hipótesis que mejor minimiza, por tanto los errores obtenidos serán próximos a 0.

Desventajas:

- Si la dimensión VC es infinita, de ninguna forma podremos llevar a cabo el aprendizaje con ERM.
- Si el valor de N no es suficientemente grande, no podemos garantizar la generalización y ésta desventaja nos abre paso a la siguiente.
- Partiendo de la desventaja anterior, es posible que la hipótesis elegida ajuste los datos demasiado bien, pero sólo en la muestra, por tanto E_{in} será cercano a 0, pero cuando evaluamos fuera de la muestra nos encontramos con que E_{out} es muy grande, lo que se conoce como el problema de 'Overfitting'.

A continuación, voy a hablar sobre SRM:
Ventajas:

- Nos aporta una solución a la última desventaja comentada en ERM, la técnica llamada 'Regularización'.
- Afronta de mejor manera problemas que contiene ruido.
- Utiliza la **Caída de peso** (Weight Decay), asociada a cada hipótesis, por lo que no sólo se centra en minimizar el E_{in} sino también $\Omega(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$, para así evitar el 'Overfitting' y reducir también la dimensión de VC.

Desventajas:

- Su principal inconveniente es que para la mayoría de conjuntos de hipótesis, encontrar la solución es una tarea que requiere de un gran cómputo.

Para finalizar, a la hora aplicar un principio u otro, lo primero que tendría en cuenta es el tamaño de la muestra, pues si tenemos una muestra de tamaño N considerable, es bastante probable que con el ERM obtengamos unos resultados satisfactorios, sin tener que realizar un cómputo muy complejo, mientras que si la la muestra de datos es pequeña o contiene ruido, sería mejor opción aplicar SRM, pues nos garantiza evitar Overfitting y obtener una mejor solución pese a que el cómputo que emplearía en ello sería mucho mayor que ERM.