Cuestionario 3

Luis Suárez Lloréns

Cuestión 1. Considere los conjuntos de hipótesis $\mathcal{H}_1$ y $\mathcal{H}_{100}$ que contienen funciones *booleanas* sobre 10 variables *booleanas*, es decir $\mathcal{X} = \{-1, +1\}^{10}$. $\mathcal{H}_1$ contiene todas las funciones *booleanas* que toman valor $+1$ en un único punto de $\mathcal{X}$ y -1 en el resto. $\mathcal{H}_{100}$ contiene todas las funciones *booleanas* que toman valor $+1$ en exactamente 100 puntos de $\mathcal{X}$ y -1 en el resto.

a) ¿Cuántas hipótesis contienen $\mathcal{H}_1$ y $\mathcal{H}_{100}$?

b) ¿Cuántos bits son necesarios para especificar una hipótesis en $\mathcal{H}_1$?

c) ¿Cuántos bits son necesarios para especificar una hipótesis en $\mathcal{H}_{100}$?

Argumente sobre la relación entre la complejidad de una clase de funciones y la complejidad de sus componentes.

Respuesta. a) La clase $\mathcal{H}^1$ tiene un sólo punto que vale 1, y el resto valen 0. Por tanto, es la elección de dicho punto es el parámetro a seleccionar. Por tanto, el número de funciones en $\mathcal{H}^1$ es igual al número de puntos. El número de puntos es $2^{10} = 1024$.

La clase $\mathcal{H}^{100}$ tiene por parámetros 100 puntos del espacio. Tenemos que elegir 100 puntos, sin repetición, y sin importar el orden. Por tanto, el número de datos es $\binom{1024}{100}$. Por tanto, tenemos 7.75×10^{140} funciones distintas.

b) Simplemente necesitamos almacenar el punto en el que vale 1. Por tanto son 10 bits.

c) De igual manera, necesitamos almacenar los 100 puntos. Cada punto necesita para ser almacenado 10 bits, luego necesitamos 1000 bits.

Podemos ver que un espacio de funciones más complejo, no sólo el número de funciones disponibles es mayor, si no que también el de sus componentes. Esto es lógico, pues el número de parámetros crece, y esto hace crecer tanto el número de posibilidades de la clase de funciones como la complejidad de representación de una posible solución, pues necesita dar valor a una variable más.

Cuestión 2. Suponga que durante 5 semanas seguidas, recibe un correo postal que predice el resultado del partido de fútbol del domingo, donde hay apuestas sustanciosas. Cada lunes revisa la predicción y observa que la predicción es correcta en todas las ocasiones. El día de después del quinto partido recibe una carta diciéndole que si desea conocer la predicción de la semana que viene debe pagar 50.000 euros. ¿Pagaría?

a) ¿Cuántas son las posibles predicciones gana-pierde para los cinco partidos?

b) Si el remitente desea estar seguro de que al menos una persona recibe de él la predicción correcta sobre los 5 partidos, ¿cuál es el mínimo número de cartas que deberá enviar?

c) Después de la primera carta prediciendo el resultado del primer partido, ¿a cuántos de los seleccionados inicialmente deberá de enviarle la segunda carta?

d) ¿Cuántas cartas en total se habrán enviado depués de las primeras cinco semanas?

e) Si el coste de imprimir y enviar las cartas es de 0.5 euros por carta, ¿cuánto ingresa el remitente si el receptor de las 5 predicciones acertadas decide pagar los 50.000 euros ?

f) ¿Puede relacionar esta situación con la función de crecimiento y la credibilidad del ajuste a los datos?

Respuesta. a) Hay 5 partidos, y para cada partido hay dos resultados posibles, ganar o perder. Luego tenemos 2^5 opciones, es decir 32 predicciones distintas.

b) Evidentemente, si hay 32 opciones, puede enviar a 32 personas distintas una predicción distinta a cada una. Así, al terminar el proceso, una y sólo una de las 32 personas habrá recibido las 5 predicciones correctas.

c) Tendrá que enviar la carta a la mitad de las personas, pues habrá enviado 16 "gana el equipo 1" y 16 "gana el equipo 2".

d) Habrá enviado 32 cartas la primera semana, 16 la segunda, 8 la tercera, 4 la cuarta y 2 la última semana. En total 62 cartas distintas.

e) Simplemente tenemos que realizar la siguiente operación:

$$50000 - 0.5 \times 62 = 49969$$

Se obtendrían 49969 €.

f) Esta situación nos muestra que, con pocas muestras, es muy fácil crear una predicción correcta. Conforme aumenta el número de partidos, el número de escenarios posible crecerá. Crear un ajuste que obtenga buenos resultados sobre muchas muestras será por tanto más creíble que con pocas, pues como hemos visto en este ejemplo, para un número pequeño de muestras, podríamos hasta generar a mano todas las diferentes posibilidades sin mucho problema.

Cuestión 3. En un experimento para determinar la distribución del tamaño de los peces en un lago, se decide echar una red para capturar una muestra representativa. Así se hace y se obtiene una muestra suficientemente grande de la que se pueden obtener conclusiones estadísticas sobre los peces del lago. Se obtiene la distribución de peces por tamaño y se entregan las conclusiones. Discuta si las conclusiones obtenidas servirán para el objetivo que se persigue e

identifique si hay algo que lo impida.

Respuesta. El enunciado dice que se obtiene una muestra suficientemente grande, pero no se asegura anda sobre la distribución de la misma. Para poder obtener unos buenos resultados sobre la distribución de los peces, necesitamos que las muestras estudiadas se obtengan de la misma distribución.

Esto no tiene porque ser así con el experimento propuesto. Al echar una red, puede que alcance a un banco de peces de un tipo concreto, y que haya tipos de peces que no se encuentran en la muestra por no encontrarse en esa zona concreta, obteniendo un resultado sesgado.

Ahora supongamos que también se cumpla que los datos representan bien todas las clases de peces. En este caso, no podremos asegurar que esa sea la distribución. Lo más que podemos asegurar es que sea muy probable que sea esa la distribución del tamaño de los peces del lago.

Determinar exactamente la distribución es una tarea prácticamente imposible. Si relajamos un poco el objetivo y que lo que queremos es una estimación de la distribución, podríamos decir que tenemos un alto grado de creencia de que el estudio cumple el objetivo.

Cuestión 4. Considere la siguiente aproximación al aprendizaje. Mirando los datos, parece que los datos son linealmente separables, por tanto decidimos usar un simple perceptrón y obtenemos un error de entrenamiento cero con los pesos óptimos encontrados. Ahora deseamos obtener algunas conclusiones sobre generalización, por tanto miramos el valor d_{VC} de nuestro modelo y vemos que es $d + 1$. Usamos dicho valor de d_{VC} para obtener una cota del error de test. Argumente a favor o en contra de esta forma de proceder identificando los posible fallos si los hubiera y en su caso cuál hubiera sido la forma correcta de actuación.

Respuesta. Este procedimiento tiene, de base, un gran problema. Observar los datos contamina de manera total el procedimiento, pues nos lleva directamente a elegir un procedimiento y un conjunto de funciones, en este caso lineales. Esto es incorrecto, pues ya estamos introduciendo conocimiento del problema.

Dado esto, hace que no tenga sentido la parte de obtener una cota del error, pues el error en el procedimiento cometido al seleccionar la función o el conjunto de funciones también rompe el concepto de la cota.

Cuestión 5. Suponga que separamos 100 ejemplos de un conjunto $\mathcal{D}$ que no serán usados para entrenamiento sino que serán usados para seleccionar una de las tres hipótesis finales g_1, g_2 y g_3 producidas por tres algoritmos de aprendizaje distintos entrenados sobre el resto de datos.
Cada algoritmo trabaja con un conjunto $\mathcal{H}$ de tamaño 500. Nuestro deseo es caracterizar la precisión de la estimación $E_{out}(g)$ sobre la hipótesis final seleccionada cuando usamos los mismos 100 ejemplos para hacer la estimación.

a ¿Qué expresión usaría para calcular la precisión? Justifique la decisión

b ¿Cuál es el nivel de contaminación de estos 100 ejemplos comparándolo con el caso donde estas muestras fueran usadas en el entrenamiento en lugar de en la selección final?

Respuesta.

Cuestión 6. Considere la tarea de seleccionar una regla del vecino más cercano. ¿Qué hay de erróneo en la siguiente lógica que se aplica a la selección de k? (Los límites son cuando $N \to \infty$). *"Considere la posibilidad de establecer la clase de hipótesis H_{NN} con N reglas, las $k - NN$ hipótesis, usando $k = 1, \ldots, N$. Use el error dentro de la muestra para elegir un valor de k que minimiza E_{in}. Utilizando el error de generalización para N hipótesis, obtenemos la conclusión de que $E_{in} \to E_{out}$ porque $\log N/N \to 0$. Por lo tanto concluimos que asintóticamente, estaremos eligiendo el mejor valor de k, basándonos sólo en E_{in}."*

Respuesta. Este procedimiento tiene varios errores.

Por un lado, seleccionar el k para el que el error en la muestra E_{in} es mínima, va a seleccionar siempre $k = 1$. Esto es lógico, pues con $k = 1$ seleccionamos siempre como clase el más cercano, lo que dentro de la muestra es siempre es el mismo punto que estás evaluando, luego el error E_{in} va a ser 0.

Podría pasar que de verdad $k = 1$ sea la mejor elección, pero sabemos que no tiene porque serlo. Es más, $k = 1$ realiza un gran sobreaprendizaje de la muestra lo que, como sabemos, reduce enormemente la capacidad de generalización.

Por otro lado, tenemos que no es adecuado de todas formas seleccionar la cota para N funciones. Esto se debe a que cada una de las hipótesis —cada uno de los distintos k— hay infinitas funciones en esa clase, y no una sola como indica el ejercicio. Es más, para $k = 1$, podemos clasificar perfectamente cualquier conjunto de puntos, por lo que hemos visto antes. Esto nos indica que la cota de Vapnik-Chervonenskis es infinita. Es claro que una sola función no puede hacer eso. Luego cada k nos da un conjunto de funciones, y no se puede realizar la cota que se nos indica.

Cuestión 7. a) Considere un núcleo Gaussiano en un modelo de base radial. ¿Qué representa $g(x)$ (ecuación 6.2 del libro LfD) cuando $||x|| \to \infty$ para el modelo RBF no-paramétrico versus el modelo RBF paramétrico, asumiendo los $\mathbf{w}_n$ fijos.

b) Sea Z una matriz cuadrada de características definida por $Z_{nj} = \phi_j(\mathbf{x}_n)$ donde $\phi_j(\mathbf{x})$ representa una transformación no lineal. Suponer que Z es invertible. Mostrar que un modelo paramétrico de base radial, con $g(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x})$ y $\mathbf{w} = Z^{-1}\mathbf{y}$, interpola los puntos de forma exacta. Es decir, que $g(\mathbf{x}_n) = \mathbf{y}_n$, con $E_{in}(g) = 0$.

c) ¿Se verifica siempre que $E_{in}(g) = 0$ en el modelo no-paramétrico?

Respuesta.

Cuestión 8. Verificar que la función sign puede ser aproximada por la función tanh. Dados $\mathbf{w}_1$ y $\epsilon > 0$, encontrar $\mathbf{w}_2$ tal que $|sign(\mathbf{x}_n^T\mathbf{w}_1) - tanh(\mathbf{x}_n^T\mathbf{w}_2)| \leq \epsilon$ para $x_n \in \mathcal{D}$ (Ayuda: analizar la función $tanh(\alpha\mathbf{x}), \alpha \in R$).

Respuesta. Sabemos que la tangente hiperbólica lleva los valores muy grandes en valores que son casi 1 o -1. Por tanto, la idea que perseguimos es convertir el número que obtengamos en la operación $x_n^T w_1$ en un número lo suficientemente grande como para quedarse a un ϵ del valor correspondiente, 1 o -1.

Por tanto, tomamos $w_2 = w_1\alpha$. Por simplificar la notación, llamamos β a $x_n^T w_1$. Entonces la ecuación queda de la siguiente forma:

$$|\text{sign}(\beta) - \tanh(\beta\alpha)| \leq \epsilon$$

Tanto si sign(β) es positivo como si es negativo, el problema es el mismo. Tenemos que encontrar un valor de α que haga el producto lo suficientemente grande. Pero β es una constante, luego siempre podemos encontrar un número α que haga el producto lo suficientemente grande para que se cumpla la inecuación.

Cuestión 9. Sean V y Q el número de nodos y pesos en una red neuronal,

$$V = \sum_{l=0}^{L} d^{(l)}, \quad Q = \sum_{l=1}^{L} d^{(l)}(d^{(l+1)} + 1)$$

En términos de V y Q, ¿cuántas operaciones se realizan en un pase hacia adelante (sumas, multiplicaciones y evaluaciones de θ)? (Ayuda: analizar la complejidad en términos de V y Q).

Respuesta. Número de evaluaciones de θ
En cada nodo de la red, se aplica una vez la función θ. Por tanto, es directo que se realizan V evaluaciones de la función θ.

Número de productos:
En el proceso de paso hacia delante, se realiza en cada capa el producto del vector de las entradas por la matriz de pesos. Por tanto, se realizan tantos productos como pesos haya en la matriz de pesos. Al sumar todas las capas, obtenemos tantas multiplicaciones como pesos haya en la red. Es decir, se realizan Q productos.

Número de sumas:
En cada capa, se realiza el producto del vector de las entradas por la matriz de pesos. Ya hemos comentado que se realizan tantos productos como pesos haya en la capa, y tras esto, hay que sumar los valores, para obtener las entradas para la función θ. En una capa, se realizan número de pesos menos el número de salidas sumas. En total, al sumar todas las capas, tenemos que se realizan $Q - V$ sumas.

Cuestión 10. Para el perceptron sigmoidal $h(x) = tanh(\mathbf{x}^T\mathbf{w})$, sea el error de ajuste $E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)^2$. Mostrar que

$$\nabla E_{in}(\mathbf{w}) = \frac{2}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)(1 - tanh(\mathbf{x}_n^T\mathbf{w})^2)\mathbf{x}_n$$

Si $\mathbf{w} \to \infty$ ¿qué le sucede al gradiente? ¿Cómo se relaciona esto con la dificultad de optimizar el perceptron multicapa?

Respuesta. Primero, vamos a realizar el gradiente de la función E_{in}.

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)^2 = \frac{2}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)' =$$

$$= \frac{2}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)' =$$

$$= \frac{2}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)(1 - tanh(x_n^T w)^2)(x_n^T w)' =$$

$$\frac{2}{N}\sum_{n=1}^{N}(tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)(1 - tanh(\mathbf{x}_n^T\mathbf{w})^2)\mathbf{x}_n$$

Ahora que tenemos calculado el gradiente, podemos seguir con el siguiente apartado. En él, se nos pregunta que pasa con la expresión del gradiente si $w \to \infty$.

Obtenemos que se calcula la tangente hiperbólica de un número que también tiende a infinito, y esa tangente hiperbólica tiende a 1 o a -1. Entonces el término $1 - \tanh(x_n^T w)^2$ tiende a 0. Entonces el producto tiende a 0 y la sumatoria también.

Obtenemos entonces lo siguiente:

$$\text{Si } w \to \infty, \nabla E_{in} \to 0$$

Esto es un gran problema. Las redes neuronales utilizan el descenso del gradiente o alguna variante del mismo para aprender los pesos. Pero si el gradiente tiende a 0, la variación de los pesos también tendería a 0, y el método se atascaría y no podría aprender.

Si a esto le añadimos que, dado al "back-propagation", el resto de las derivadas que dependen de esta, irán a 0 también. Luego no sólo una capa queda parada en el aprendizaje, si no todas. Por tanto, este problema puede bloquear totalmente el aprendizaje de una red neuronal.