Aprendizaje Automático: Cuestionario 2

Anabel Gómez Ríos

6 de junio de 2016

Pregunta 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.

Pregunta 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 de 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?

Pregunta 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 que lo impida.

Pregunta 4. Considere la siguiente aproximación al aprendizaje. Mirando los datos, parece que los datos son linealmente separables, por tanto decidimos usar un simple perceptron 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.

Pregunta 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?

Pregunta 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}.

Pregunta 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?

Pregunta 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 $tanh(\alpha\mathbf{x}), \alpha \in R$).

Pregunta 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).

Pregunta 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?