Aprendizaje Automático (2015-2016)
Grado en Ingeniería Informática
Universidad de Granada

Cuestionario 3

Laura Tirado López

12 de junio de 2016

1. Consideré los conjuntos de hipótesis $\mathcal{H}_1$ y $\mathcal{H}_{100}$ que contienen funciones Booleanas sobre 10 variables Booleanas, es decir $X = \{-1, +1\}^{10}$. $\mathcal{H}_1$ 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 $\mathcal{H}_{100}$? El total de elementos de $\mathcal{X}$ es 2^10, es decir, 1024 elementos. El número de hipótesis del conjunto $\mathcal{H}_1$ sería el número combinatorio $\binom{1024}{1}$, dado que estamos escogiendo sólo un elemento con el valor $+1$ sobre el total de elementos de $\mathcal{X}$. Este número es 1024. Siguiendo este mismo razonamiento el número de hipótesis del conjunto $\mathcal{H}_{100}$ es $\binom{1024}{100}$, que es igual a $7,746\dot{1}0^{140}$ aproximadamente.

   b) ¿Cuantos bits son necesarios para especificar una de las hipótesis en $\mathcal{H}_1$? Para especificar una de las hipótesis en $\mathcal{H}_1$ serían necesarios $\log_2(1024) = 10$ bits.

   c) ¿Cuantos bits son necesarios para especificar una de las hipótesis en $\mathcal{H}_{100}$? Para especificar una de las hipótesis en $\mathcal{H}_{100}$ serían necesarios $\log_2(7,746 \cdot 10^{140}) = 469$ bits.

   Nota: El resultado de $\log_2(7,746 \cdot 10^{140})$ es $468,023$, por lo que se redondea a la unidad superior para calcular el número de bits.

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

   La complejidad de una clase de funciones se refiere al tamaño de dicha clase, es decir, al número de hipótesis que contiene la clase. La complejidad de sus componentes podemos interpretarla como el número de bits necesarios para representar las hipótesis de una clase. Si partimos de estos conceptos, podemos ver que para este caso, cuanto mayor es la clase o más compleja, mayor es la complejidad de sus componentes. Sin embargo, al tratarse de números combinatorios, esta relacion no es estrictamente lineal, dado que a patir de la clase $\mathcal{H}_{\bigtriangledown\infty\in}$, la complejidad de la clase irá decrementándose. Por tanto, para clases que contienen este tipo de funciones Booleanas la relación entre la complejidad de una clase de funciones y la complejidad de sus componentes será que cuanto más compleja sea la clase más complejos serán sus componentes hasta la clase $\mathcal{H}_{\bigtriangledown\infty\in}$ a partir de la cual, la relación será a la inversa.

2. Suponga que durante 5 semanas seguidas, recibe un correo postal que predice el resultado del partido de fútbol del domingo, donde hay apuestas substanciosas. 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. ¿Pagaría?

   a) ¿Cuántas son las posibles predicciones gana-pierde para los cinco partidos? Las posibles predicciones gana-pierde para los cinco partidos son $2^5 = 32$, dado que para cada partido hay dos posibilidades.

b) Si el remitente desea estar seguro de que al menos una persona recibe de él la predicción correcta sobre los 5 partidos, ¿Cual es el mínimo número de cartas que deberá de enviar? Para estar seguro debería enviar una carta con cada posible predicción para que al menos una persona reciba la predicción correcta por lo que debería enviar 32 cartas.

c) Después de la primera carta prediciendo el resultado del primer partido, ¿a cuántos de los seleccionados inicialmente deberá de enviarle la segunda carta? Deberá enviar la segunda carta a $2^4 = 16$ personas.

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

$$\sum_{i=1}^{5} 2^i] = 62$$

cartas.

e) Si el coste de imprimir y enviar las cartas es de 0,5 por carta, ¿Cuánto ingresa el remitente si el receptor de las 5 predicciones acertadas decide pagar los 50.000? El coste total de imprimir y enviar las cartas será de 31. Si el receptor para los 50.000, el remitente habrá ingresado 49.939.

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

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.

   Las conclusiones obtenidas no servirán si aunque la muestra sea grande esta no es representativa de la población. En este caso, la muestra no sería representativa dado que la posibilidad de coger peces más grandes es mayor que la de coger peces pequeños con la red, a no ser que la red sea muy fina. También podría darse que las zonas en las que se eche la red haya un mayor número de peces pequeños que de peces grandes o viceversa. Teniendo en cuenta estos posibles escenarios, la muestra no seguiría una distribución uniforme, por lo que no sería representativa sobre la población y las conclusiones obtenidas no servirían para determinar la distribución del tamaño de los peces. El método de selección de la muestra siempre debe asegurar que la muestra se elija de forma aleatoria y que sea uniformemente distribuida.

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 cual hubiera sido la forma correcta de actuación.

Obtener cero como error de entrenamiento no garantiza una buena generalización, dado que podría haberse sobreajustado sobre los datos de entrenamiento. También podría ser que el conjunto de datos sea pequeño, por lo que es fácil ajustarlo, pero si se añadiese una mayor cantidad de datos, podría cambiar completamente la clasificación. En mi opinión, podrían haberse implementado varios modelos y evaluarlos con validación cruzada para determinar cuál es el que mejor clasifica los datos garantizando una buena generalización.

5. Suponga que separamos 100 ejemplos de un conjunyo $\mathcal{D}$ 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) ¿Que expresión usaría para calcular la precisión? Justifique la decisión

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

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,...,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, basadonos solo en E_{in}."*

7. *a*) Considere un núcleo Gaussiano en un modelo de base radial. ¿Que 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?

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

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

10. Para el perceptrom 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 $\mathbf{w} \to \infty$ ¿que le sucede al gradiente? ¿Cómo se realiciona esto con la dificultad de optimizar el perceptron multicapa?