Aprendizaje automático: Cuestionario 3.

Cristina Zuheros Montes.

13/06/2016.

Todas las preguntas tienen el mismo valor

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

   (a) ¿Cuantas hipótesis contienen $\mathcal{H}_\infty$ y $\mathcal{H}_\infty{''}$
   Solución
   En el primer caso, tendremos un sólo punto con valor distinto a los del resto. De modo que tendremos $|\mathcal{H}_\infty| = 2^{10}$ hipótesis.
   Para el segundo caso, tendremos 100 puntos con valor distinto a los del resto. De modo que tendremos $|\mathcal{H}_\infty| = 2^{10*100} = 2^{1000}$ hipótesis.

   (b) ¿Cuantos bits son necesarios para especificar una de las hipótesis en $\mathcal{H}_\infty$?
   Solución
   Por el apartado anterior, vemos que necesitamos 10 bits.

   (c) ¿Cuantos bits son necesarios para especificar una de las hipótesis en $\mathcal{H}_\infty{''}$?
   Solución
   Por el apartado a), vemos que necesitamos 1000 bits.

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

2. Suponga que durante 5 semanas seguidas, recibe un correo postal que predice el resultado del partido de futbol 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 diciendole que si desea conocer la predicción de la semana que viene debe pagar 50.000?. ¿Pagaría?

   (a) ¿Cuantas son las posibles predicciones gana-pierde para los cinco partidos?
   Solución
   Tendremos 5 partidos posibles, uno para cada día. Y cada partido tiene 2 predicciones posibles. El número de predicciones posibles para los cinco partidos, será: $2^5 = 32$

   (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?
   Solución
   Se tendrán que enviar como mínimo 32 cartas, las 32 cartas con predicciones diferentes que hemos visto en el apartado anterior. Con sólamente estas 32 cartas, es seguro que al menos una persona va a recibir la predicción correcta para todos los días, aunque también habrá al menos una persona que recibirá todas las predicciones incorrectas.

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

Solución
De las 32 cartas que mandamos, la mitad de ellas habrán obtenido una predicción correcta. De modo que será a $2^4 = 16$ personas a las que tendremos que enviarle la segunda carta.

(d) ¿Cuantas cartas en total se habrán enviado después de las primeras cinco semanas?
Solución
Por el mismo razonamiento que seguimos en el apartado anterior, tenemos que mandar: $2^5 + 2^4 + 2^3 + 2^1 + 2^0 = 63$ cartas.

(e) Si el coste de imprimir y enviar las cartas es de 0.5? por carta, ¿Cuanto ingresa el remitente si el receptor de las 5 predicciones acertadas decide pagar los 50.000??
Solución
Hemos dicho que tendremos que enviar unas 63 cartas, a 0.5 cada una saldrían por un coste total de $63 * 0.5 = 31.5$ Si el receptor para 50000, el remitente ganará $50000 - 31.5 = 49968.5$

(f) ¿Puede relacionar esta situación con la función de crecimiento y la credibilidad del ajuste de los datos?
Solución
Las predicciones que mandamos a una persona durante las 5 semanas representarán una dicotomía. A medida que aumentamos el número de cartas que se mandan, iremos teniendo mayor credibilidad y mayor porcentaje de acierto.

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.

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.

5. Suponga que separamos 100 ejemplos de un conjunyo 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 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 w_n fijos.

   (b) Sea Z una matriz cuadrada de características definida por $Z_{nj} = \Phi_j(x_n)$ donde $\Phi_j(x)$ representa una transformación no lineal. Suponer que Z es invertible. Mostrar que un modelo paramétrico de base radial, con $g(x) = w^T\Phi(x)$ y $w = Z^{-1}y$, interpola los puntos de forma exacta. Es decir, que $g(x_n) = y_n$, 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 w_1 y $\epsilon > 0$ encontrar w_2 tal que $|\text{sign}(x_n^T w_1) - \tanh(x_n^T w_2)| \leq \epsilon$ para $x_n \in D$ (Ayuda: analizar la función $\tanh(\alpha x)$, $\alpha \in \mathrm{R}$)
Solución
Esta desigualdad se puede verificar probando que ($w_2 = k * w_1$):

$$lim_{k\to\infty} \tanh(k * x_n^T w_1) = \text{sign}(x_n^T w_1)$$

Es decir, tenemos que ver que la función tangente hiperbólica es una aproximación suave de la función signo. Veamos los límites, verificando que obtenemos la función signo.

$$lim_{x_n^T w_1\to\infty} \tanh(k * x_n^T w_1) = lim_{x_n^T w_1\to\infty} \frac{e^{k*x_n^T w_1} - e^{-k*x_n^T w_1}}{e^{k*x_n^T w_1} + e^{-k*x_n^T w_1}} \frac{e^{-k*x_n^T w_1}}{e^{-k*x_n^T w_1}} =$$

$$lim_{x_n^T w_1\to\infty} \frac{1 - e^{-2k*x_n^T w_1}}{1 + e^{-2k*x_n^T w_1}} = \frac{1-0}{1+0} = 1$$

$$lim_{x_n^T w_1\to-\infty} \tanh(k * x_n^T w_1) = lim_{x_n^T w_1\to-\infty} \frac{e^{k*x_n^T w_1} - e^{-k*x_n^T w_1}}{e^{k*x_n^T w_1} + e^{-k*x_n^T w_1}} \frac{e^{k*x_n^T w_1}}{e^{k*x_n^T w_1}} =$$

$$lim_{x_n^T w_1\to-\infty} \frac{e^{-2k*x_n^T w_1} - 1}{e^{-2k*x_n^T w_1} + 1} = \frac{0-1}{0+1} = -1$$

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 ¿cuantos 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 $h(x) = \tanh(x^T w)$, sea el error de ajuste $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(\tanh(x_n^T w) - y_n)^2$. Mostrar que

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

si $w \to \infty$ ¿que le sucede al gradiente? ¿Cómo se relaciona esto con la dificultad de optimizar el perceptron multicapa?
Solución
En primer lugar sabemos que:

$$\frac{\partial}{\partial_w} tanh(w) = sech^2(w) = 1 - tanh(w)^2$$

Luego tendremos:

$$\nabla E_{in}(w) = \frac{2}{N}\sum_{n=1}^{N}(\tanh(x_n^T w) - y_n) * \frac{\partial}{\partial_w} tanh(x_n^T w)$$

$$= \frac{2}{N}\sum_{n=1}^{N}(\tanh(x_n^T w) - y_n) * (1 - tanh(x_n^T w)^2) * x_n$$

Quedando probada la igualdad. Si $w \to \infty$, entonces el gradiente tiene a 0, pues:

$$lim_{w\to\infty} tanh(x_n^T w)^2 = 1$$