CONTESTACIONES CUESTIONARIO-1.
Orden Aleatorio

- Suponga que disponemos de un conjunto de 1000 muestras etiquetadas de un problema de clasificación binaria....

  Respuesta: Dividir la muestra en conjuntos de entrenamiento y validación en rango (90 %-10 % a 80 %-20 %): $(N-K, K)$. Entrenar los 5 modelos y elegir el de menor error de validación. La cota

$$P(|E_{\text{out}} - E_{\text{val}}| < \epsilon) > 2 \times 5 \times e^{-2\epsilon^2 K}$$

$$|E_{\text{out}} - E_{\text{val}}| \leq \sqrt{\frac{ln10 - ln\delta}{2K}}$$

- Suponga un problema de predicción bien definido que hace uso de una clase de funciones $\mathcal{H}$. Suponga que dicha clase $\mathcal{H}$ es suficiente como para hacer que $E_{in} \to 0$ para cualquier muestra de un tamaño dado N....

  Respuesta: Dado que $E_{in} \to 0$ para toda muestra de tamaño N, en general, la clase $\mathcal{H}$ estará sobreajustando la muestra debido a la presencia natural de ruido estocástico. Al ir reduciendo la complejidad de la clase el sesgo de la estimación aumentará y la varianza disminuirá pero en total el error por sobreajuste disminuirá hasta alcanzar un mínimo. Si continuamos reduciendo la complejidad de $\mathcal{H}$ el error del sesgo crecerá más que la disminución de la varianza, dando lugar a un aumentando del sobreajuste. La complejidad de la clase en la que se alcanza el mínimo del error por sobreajuste depende del tamaño N de la muestra, siendo en general creciente con N.

- Suponga que debe etiquetar muestras en un problema de clasificación binario donde desconoce la función de etiquetado f, pero conoce la distribución muestral $\mathcal{P}$. ....

  Respuesta: Aplicar la regla de Bayes: asignar a la clase de mayor probabilidad. Esta regla garantiza el mínimo error posible.

- Considere la función de error $E_n(\mathbf{w}) = (\max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2$. Deducir la regla de adaptación de gradiente descendente para el parámetro $\mathbf{w}$....

  Respuesta:

$$\nabla E_n = \left\{ \begin{array}{cc} 0 & si \quad y\mathbf{w}^T\mathbf{x} \geq 1) \\ -2\mathbf{x}(y - \mathbf{w}^T\mathbf{x}) & si \quad y\mathbf{w}^T\mathbf{x} < 1 \end{array} \right.$$

  Equivalentemente: $\nabla E_n = -2y\mathbf{x} \cdot \max(0, 1 - y\mathbf{w}^T\mathbf{x})$

  Regla adaptación para $\mathbf{w}$:

$$\mathbf{w}_{t+1} = \left\{ \begin{array}{cc} \mathbf{w}_t & si \quad y\mathbf{w}^T\mathbf{x} \geq 1 \\ \mathbf{w}_t + 2\lambda\mathbf{x}(y - \mathbf{w}^T\mathbf{x}) & si \quad y\mathbf{w}^T\mathbf{x} < 1) \end{array} \right.$$

Equivalentemente: $\mathbf{w}_{t+1} = \mathbf{w}_t + 2\lambda y\mathbf{x} \cdot \max(0, 1 - y\mathbf{w}^T\mathbf{x})$

Regla adaptación para $E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} E_n(\mathbf{w})$:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{2\lambda}{N} \sum_{i:y_i\mathbf{w}^T\mathbf{x}_i<1} \mathbf{x}_i(\mathbf{w}^T\mathbf{x}_i - y_i)$$

Equivalentemente: $\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{2\lambda}{N}\sum_i y_i\mathbf{x}_i \cdot \max(1 - y_i\mathbf{w}^T\mathbf{x}_i)$

- Cuando resolvemos un problema de aprendizaje desde datos cuales de las siguientes opciones son correctas:...

  Respuesta: Opción (c). Suponemos una clase de funciones finita sin perdida de generalidad. Si con la clase de funciones usada E_{in} es grande hemos fallado. En caso contrario habremos obtenido una solución PAC.

  Equivalentemente: La desigualdad de Hoeffding nos permite establecer $E_{\text{in}} \approx E_{\text{out}}$. Entonces si se logra $E_{\text{in}} \approx 0$ tendremos una solución PAC, en caso contrario hemos fallado.

- Suponga el siguiente escenario. Un matemático que no cree en la teoría del aprendizaje desde datos le reta a que averigue una función de etiquetado $f : \mathcal{R} \to \{-1, +1\}$ ...

  Respuesta: Ya que la muestra no viene de distribución de probabilidad sobre $\mathcal{X} \times \mathcal{Y}$ el problema define una inducción determinista y por tanto no hay solución al mismo.

- Considere el mismo escenario de la pregunta anterior. Ahora el matemático le permite que Ud. añadir una información extra...

  Respuesta: Una distribución de probabilidad sobre el espacio $\mathcal{X} \times \mathcal{Y}$ y una muestra i.i.d de la misma. En esas condiciones una clase finita de funciones es siempre PAC-aprendible desde una muestra independiente

- En regresión hemos introducido la función de error $\mathbf{e}(\mathbf{x}) = (h(\mathbf{x}) - y)^2$ entre predicción y etiquetas. Sin embargo posibles medidas aberrantes de las y producen valores de error enormes que condicionan la búsqueda de buenas soluciones....

  Respuesta: Claramente *Error* alcanzará su mínimo cuando $h(\mathbf{z})$ tome un valor entre z_1 y z_N. Analizando el valor *Error* cuando $h(\mathbf{z})$ avanza desde z_1 a z_N se observa que el valor de *Error* decrece linealmente con el número de muestras que deja a su izquierda y crece con el número de muestras que deja a su derecha. Por tanto alcanzara el mínimo cuando tenga tantas muestras a su derecha como a su izquierda. Esto es por defínción la mediana($\mathbf{z}$). Dado que la mediana solo tiene en cuenta el número de datos a su izquierda y derecha las observaciones aberrantes no tienen mayor influencia en el valor de estimador que cualquier otro punto. En conclusión minimizar la métrica L_1 de los errores produce soluciones robustas

Equivalentemente: Consideremos $h(\mathbf{z}) \in [z_1, z_N], h(\mathbf{z}) \neq z_k, k = 1, .., N$ y supongamos j el número de valores z menores de $h(\mathbf{z})$. Entonces $Error = \sum_{k=1}^{j}(h(\mathbf{z}) - z_k) + \sum_{k=j+1}^{N}(z_k - h(\mathbf{z})) = (j - N + j) \times h(\mathbf{z}) - \sum_{k=1}^{j} z_k + \sum_{k=j+1} z_j^N$. $Error$ alcanza un mínimo respecto de $h(\mathbf{z})$ cuando $2j = N$. Es decir, $h(\mathbf{z})$ =mediana($\mathbf{z}$). Alternativamente: Se acepta el uso de la derivada del valor absoluto (no derivable en 0), si esta bien usada.

- Considere un modelo de "bin"para modelar una hipótesis h que comete un error μ al aproximar una función determinística f. ....

  Respuesta: $P[h(x) \neq y] = P[h(x) \neq f(x)]P[f(x) = y] + P[h(x) = f(x)]P[f(x) \neq y] = (1 - \mu)(1 - \lambda) + \mu\lambda = 2\mu\lambda - \mu - \lambda + 1$ Cuando $2\mu\lambda - \mu = 0 \rightarrow \lambda = \frac{1}{2}$

- Consideremos un problema de clasificación binaria probabilístico con etiquetas $\{+1, -1\}$ donde la solución probabilística sin costes esta dada por la función

  $$g(x) = \mathbb{P}[y = +1|\mathbf{x}]$$

  Suponga que la matriz de coste...

  Respuesta: De acuerdo a la regla de Bayes la regla sería asignar a la clase más probable. Los pesos se incorporan a esta regla haciendo que el coste promedio (error) de ambas clases sea igual, es decir $1000 \cdot g(\mathbf{x}) = 10 \cdot (1 - g(\mathbf{x}))$. Por tanto asignar $+1$ si $g(\mathbf{x}) \geq \frac{1}{101}$.

- Es bien conocido que la investigación en "Machine Learning"ha ido produciendo distintas técnicas exitosas para aproximar problemas de clasificación: k-nn, árboles,SVM, NN, AdaBoost, etc...

  Respuesta: Ninguno o cualquiera. No existe el mejor algoritmo para todas las distribuciones como se deduce de los resultados del teorema Non-Free Lunch.

- Considere $E_n(\mathbf{w}) = \mathbf{max(0, 1 - y_n w^T x_n)}$. Mostra que $E_n(w)$ es una cota superior...

  Respuesta: $[[sign(\mathbf{w}^T\mathbf{x_n}) \neq sign(y_n)]]$ toma valores 0 y 1. Cuando toma el valor 0 $max(0, 1 - y_n\mathbf{w}^T\mathbf{x_n})$ es también igual a 0 y para el valor 1 $max(0, 1 - y_n\mathbf{w}^T\mathbf{x_n})$ toma valores extrictamente mayores a 1.

  Por el resultado anterior $E_{\text{in}} = \frac{1}{N}\sum_n[[sign(\mathbf{w}^T\mathbf{x_n}) \neq sign(y_n)]] \leq \frac{1}{N}\sum_n E_n(w)$