CONTESTACIONES CUESTIONARIO-2.

- **Pregunta.1:** No existe "data Snooping". La información de que la muestra es separable NO aporta ningún tipo de información sobre de la clase de funciones que se ha usado en la afirmación, ya que las clases conocidas van desde dimension 1 a infinito. Por tanto la cota es calculable y para una confianza del 95 %sería

$$E_{\text{out}} \leq \sqrt{\frac{8}{1000} \ln\left(\frac{4(2000^4+1)}{0{,}05}\right)} = 0{,}5275$$

- **Pregunta.2:** De la función de crecimiento se deduce $m_{\mathcal{H}}(2) = 2^2$ y $m_{\mathcal{H}}(3) < 2^3$ por tanto la dimensión VC es 2 y la cota es

$$E_{\text{out}} \leq E_{\text{in}} + \sqrt{\frac{8}{1000} \ln\left(\frac{4(2000^2+1)}{0{,}05}\right)} = E_{\text{in}} + 0{,}3958$$

  Alternativamente se puede usar la cota que usa $m_{\mathcal{H}}(2N)$

- **Pregunta.3:** Sea $d_{vc}(\mathcal{H}_i) = d_i$. De la desigualdad dada en el enunciado se obtiene que $m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) < 2^N$ cuando $d_1 + d_2 + 2 \leq N$. Por tanto $d_1 + d_2 + 2$ es un punto de ruptura para $\mathcal{H}_1 \cup \mathcal{H}_2$ y $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_1 + d_2 + 1$. Entonces aplicando recursion(o inducción) tenemos $d_{vc}(\cup_{k=1}^{K} \mathcal{H}_k) = d_{vc}(\cup_{k=1}^{K-1} \mathcal{H}_k \cup \mathcal{H}_K) \leq d_{vc}(\cup_{k=1}^{K-1} \mathcal{H}_k) + d_K + 1 \leq d_{vc}(\cup_{k=1}^{K-2} \mathcal{H}_k) + d_K + d_{K-1} + 2 \leq ... \leq \sum_{k=1}^{K} d_k + K - 1$. Por construcción $d_1 = \cdots = d_k = d + 1$ ya que las valores de los elementos de la muestra no afectan a la dimensión de VC. Luego la cota es $d_{vc}(\cup_{k=1}^{K} \mathcal{H}_k) \leq K(d+2) - 1$

- **Pregunta.4:** Sea $m_{\mathcal{H}}(N) = K$ entonces K es el máximo de dicotomías para N puntos. Pero $m_{\mathcal{H}}(N+N) \leq m_{\mathcal{H}}(N) m_{\mathcal{H}}(N)$ ya que las dicotomias de $2N$ puntos se pueden construir por extensión de las de N añadiendo y etiquetando con las mismas funciones otro conjunto igual. Como máximo se podrán generar el producto de dicotomías, pero no hay garantias de que cada una de ellas se pueda generar con una única función.

  Alternativamente: si $m_{\mathcal{H}}(2N) = 2^{2N}$ entonces $m_{\mathcal{H}}(2N) = m_{\mathcal{H}}(N)^2$. Si $m_{\mathcal{H}}(N) = 2^N$ y $m_{\mathcal{H}}(2N) < 2^{2N}$ es obvio. Sea $k < N$ el mayor entero tal que $m_{\mathcal{H}}(k) = 2^k$, entonces $2^{2k} < m_{\mathcal{H}}(N)^2 < 2^{2N}$ pero $m_{\mathcal{H}}(2N) \leq 2^k \times 2^k = 2^{2k}$

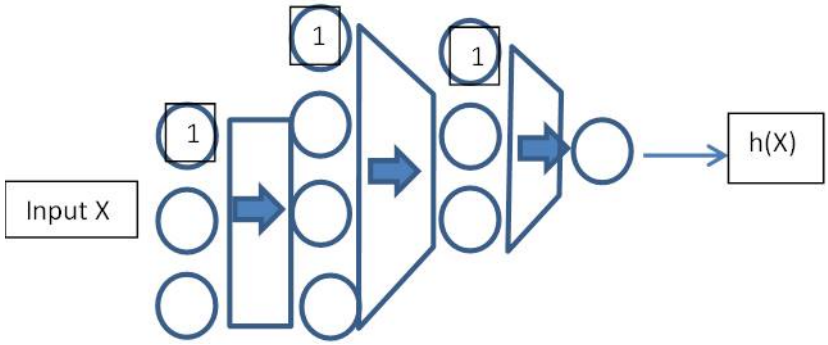
- **Pregunta.5:**

$s^{(1)}$	$x^{(1)}$	$s^{(2)}$	$x^{(2)}$	$s^{(3)}$	$x^{(3)}$	$10^5 \times \delta^{(3)}$	$10^4 \times \delta^{(2)}$	$10^4 \times \delta^{(1)}$
1,5	0,9051	−0,0186	−0,0186	3,4602	0,9980	−1,5554	0,311	−0,0717
1,3	0,8617	1,1961	0,8325				−0,1433	−0,1009
0,7	0,6044							0,7740

$$\frac{\partial \mathbf{e}}{\partial \mathbf{W}_1} = 10^{-3} \times \begin{pmatrix} -0{,}0072 & -0{,}0101 & 0{,}0774 \\ -0{,}0143 & -0{,}0202 & 0{,}1548 \\ -0{,}0143 & -0{,}0202 & 0{,}1548 \end{pmatrix}, \frac{\partial \mathbf{e}}{\partial \mathbf{W}_2} = 10^{-4} \times \begin{pmatrix} -0{,}3110 & -0{,}1433 \\ -0{,}2815 & -0{,}1297 \\ -0{,}2680 & -0{,}1234 \\ -0{,}1879 & -0{,}0866 \end{pmatrix}$$

$$\frac{\partial \mathbf{e}}{\partial \mathbf{W}_3} = 10^{-4} \times \begin{pmatrix} -0{,}1555 \\ 0{,}0029 \\ -0{,}1295 \end{pmatrix}$$

Grafo:



- **Pregunta.6:** Un árbol es una función definida como una combinación lineal de funciones binarias {0,1}, en donde cada función binaria se asocia con una región de una partición disjunta y completa del espacio muestral a partir de hiperplanos paralelos a los ejes. El algoritno ID3 ajusta una función de partición por adición iterativa de funciones boleanas haciendo uso de una heurística (Mínima Entropía) que elige en cada caso la función boleana de partición que implique mayor disminución de varianza. Comienza con una función simple de sesgo alto y baja varianza. La partición de los nodos aumenta la complejidad de la función que define el árbol por lo que disminuimos en sesgo y aumentamos varianza. En el caso extremo de $E_{\text{in}} = 0$ tendriamos un árbol de sesgo nulo y muy alta varianza. Por tanto la poda de nodos usando un criterio de óptimalidad entre número de nodos y calidad de ajuste alcanza un equilibrio sesgo-varianza.

- **Pregunta.7:** Dos mayores ventajas del diseño del algoritmo SVM: (1) la solución es el resultado un problema de optimización convexa que garantiza el mejor estimador lineal para casos separables. (2) No depende de la dimensionalidad de las muestras. Mayor Desventaja del algortimo: Su baja eficiencia $\mathcal{O}(N^3)$. No es computacionalmente abordable una optimización cuadrática cuando el número de muestras supera un límite bajo (5000 aprox.). Esto hace muy complicado su aplicación en problemas de tamaño medio y grande. La mayor contribución de SVM-Soft es una nueva función de perdida $Hinge - Loss$ de máximo margen en casos con ruido, para ser usada con SGD.

- **Pregunta.8:** (a)El sesgo del clasificador RF es muy bajo por construcción, ya que se usan árboles sin podar. La varianza es alta, debido a la alta varianza de los árboles. Pero el promedio de un número grande $B >> 0$ de árboles i.d. permite decrecer la varianza hasta un límite debido a la ausencia de independencia entre los árboles. Para favorecer la independencia, se construye cada árbol a partir de un conjunto aleatorio de variables lo que reduce en gran parte el efecto de correlaciones entre variables. (b) En problemas con ruido aporta una técnica, mucho más simple, que parte de una hipótesis equivalente $E_{\text{in}} \approx 0$ y que por construcción trata de minimizar el error E_{out}, lo que le hace superior a SVM en estos casos; Selección natural del conjunto de test; natural multiclase. (c) Por construcción RF no es óptimo en ningún sentido.

- **Pregunta.9:** Los aparejos de pesca usados para la explotación comercial necesariamente tendrán un tamaño mínimo de captura para no dañar el crecimiento de las futuras generaciones. Por ello el muestreo estará siempre sesgado en la estimación de los tamaños más pequeños, y por tanto lo estará en la predicción de la producción futura.

- **Pregunta.10:** La convergencia del algoritmo Perceptron, en casos separables, garantiza la convergencia también con el nuevo orden establecido. Cuando se alcanza la solución del Perceptron (separable) el nuevo criterio de adaptación seguirá moviendo el hiperplano hasta alcanzar la posición de máxima distancia a las clases. Esta distancia estará definida por los puntos más cercanos de cada clase tras no poder mejorar. La solución es óptima ya que maximiza la distancia y coincide con la de SVM. Los puntos finales de adaptación definen a los puntos soporte.

- **Pregunta.11:** Por la condición de complementariedad de las activaciones de KKT, el problema dual identifica a partir de los valores de $\alpha^* > 0$ los vectores soporte que definen la solución del primal. Los vectores soporte están a distancia $= 1$ de la solución y verifican $(\alpha^* > 0)$, pero puntos con $\alpha^* = 0$ pueden estar tambíen a distancia $= 1$ de la solución. Por tanto para los puntos con $\alpha^* = 0$ su distancia a la solución es ≥ 1

- **Pregunta.12:** Necesitamos saber si el diseño del experimento es acorde a la teoría del aprendizaje. Es decir, si la clase de funciones del experimento es capaz o no de generar todos los posibles etiquetados. Para ello consideramos un árbol que se ramifica en cada nodo de acuerdo a los posibles Q-resultados de la semana. Cada nivel de árbol una semana. Entonces por construcción las hojas del nivel k-ésimo representan todos los posibles Q^k etiquetas de los k semanas. Si reetiquetamos con acierto/ fallo las etiquetas de los nodos la clase de funciones tiene función de crecimiento igual a 2^N ya que genera todos los etiquetados. Como consecuencia, si el número de correos iniciales enviados con cada posible resultado es mayor de Q^k, y se envía siempre a los ganadores, hay garantía de haber enviado k correos con k aciertos a la misma persona. No se pagaría porque el resultado es aleatorio.