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 $f : \mathcal{R} \rightarrow \{-1, +1\}$ que él ha diseñado. Para ello le proporciona 100 valores de dicha función y una clase finita de funciones $\mathcal{H}$ y le pide que diga que función $h \in \mathcal{H}$ aproxima mejor a f.

a) ¿Cúal sería su contestación?

b) Considere ahora que el matemático le permitiera añadir una hipótesis extra al problema. ¿Añadiría algo ? ¿Qué y porqué?

Suponga que tiene la tarea de escribir un programa para etiquetar muestras en un problema de clasificación binaria. Pero solo conoce la distribución sde probabilidad sobre las muestras de población $\mathcal{P}$. ¿Cómo lo haría para garantizar una clasificación errónea lo más baja posible? Justificar la contestación

Considdere los cuatro elementos básicos de un problema de aprendizaje por inducción $\mathcal{H}, \mathcal{A}, \mathcal{P}, \mathcal{D}$, donde ($\mathcal{H}$-clase funciones, $\mathcal{A}$-Algoritmo de aprendizaje, $\mathcal{P}$-distribución de probabilidad and $\mathcal{D}$-muestras). Diga cúal de ellos condiciona con mayor importancia la solución de un problema de predicción. Argumente su contestación identificando las causas.

Considere un problema binario de clasificación en un espacio 3D. Le dan un conjunto de 1000 datos y alguien le dice que cree que son separables, Para corroborarlo ajusta un modelo perceptron y obtiene un error $E_{in} = 0$ y decide usar la dimensión de VC del perceptrón para calcular la cota de error de E_{out}. Analice la situación y diga si la información recibida afecta (si/no) al cálculo de la cota del error y cómo afecta.

Considere la función de error definida para una muestra $(\mathbf{x}, y)$ por $E(\mathbf{w})) = (\mathbf{max}(\mathbf{0}, \mathbf{1} - \mathbf{y}\mathbf{w}^{\mathbf{T}}\mathbf{x}))^{\mathbf{2}}$. Deducir la regla de adaptación de gradiente descendente para el parámetro $\mathbf{w}$. Usar el resultado para deducir la regla de adaptación para la función de error $\frac{1}{N}\sum_{n=1}^{N} E_n(\mathbf{w})$ asociada al promedio de n muestras.

¿Es posible aprender por inducción sin imponer condiciones a la clase de funciones?

a) En caso negativo, identifique alguna condición que sea especialmente relevante por su importancia y generalidad.

b) En caso negativo, indique como sería la técnica.

Utilice la teoría para justificar sus argumentos de forma clara y precisa.

Es bien sabido que la investigación en "Machine Learning" ha ido produciendo gradualmente diferentes algoritmos de considerable éxito para aproximar problemas de clasificación: k-NN, RL, Trees, SVM, RRNN, AdaBoost, etc. Sin embargo, en los últimos años los modelos de CNN se han mostrado increíbles superioridad sobre técnicas anteriores en multitud de problemas. Si le piden, como experto, que elija una técnica como la mejor entre las mencionadas, ¿cuál elegiría y por qué?

k-NN: k vecinos más cercanos, RL: regresión logística, RRNN: redes neuronales, SVM: máquinas de vectores de soporte, CNN: redes neuronales convolucionales

En un problema de clasificación binaria la función de perdida óptima que se intenta minimizar es $\sum_{(x,y)}[[sign(h(x)) \neq y]]$, es decir el número de fallos de predicción del clasificador h, donde y representa etiqueta y $[[\cdot]]$ es un predicado con valores en $\{0, 1\}$. Verificar que dicha minmización se puede realizar a través de otras funciones de error. En particular, probar que

a) $[[sign(h(x)) \neq y]] \leq (y - h(x))^2$

b) $[[sign(h(x)) \neq y]] \leq \frac{1}{\ln 2}\ln(1 + \exp(-h(x)y))$

Considere un modelo de "bin" para una hipótesis h que comete un error μ al aproximar una función determinística f. Ambas funciones se suponen binarias. Si usamos la misma h para aproximar una versión ruidosa de f dada por

$$P(y|\mathbf{x}) = \begin{cases} \lambda & y = f(x) \\ 1 - \lambda & y \neq f(x) \end{cases}$$

¿Qué error comete h al aproximar y, y con qué valor de λ el error de h es independiente de μ?

¿Cúal sería la dimensión de VC de un modelo de red de base radial definido a partir de K núcleos , donde $K \leq N$ (tamaño de la muestra). Analice las posibles opciones.

Suponga que le piden ajustar un modelo concreto de clasificador a un muestra de tamaño N. Una vez que lo ha logrado le anuncian que dispone de nuevas muestras de entrenamiento para mejorar la estimación inicial. Analice la situación y diga:

¿Cómo afectará el uso de más muestras al sesgo y varianza de la solución final respecto de la inicial? Justifique con argumentos sólidos de la teoría.

Suponga que dispone de un conjunto de 1000 muestras etiquetadas de un problema de clasificación binaria. Le piden ajustar el mejor modelo posible dentro de 5 tipos distintos de modelos paramétricos. Haga un análisis de la situación y describa que decisiones tomaría para ello y como calcularía el error E_{out} del modelo elegido. Valore los pros y contras de cada decisión tomada.

Considere que trabaja en una empresa dedicada a la construcción de máquinas para la clasificación automática de piezas de fruta por tipos y variedades. En particular todo tipo de piezas duras no arracimadas como naranjas, manzanas, peras, melocotón, mangos, etc, hasta un total de 50 tipos distintos de frutas. Analice la situación y diga:

A) ¿Considera que este problema es un problema que puede ser resuelto por diseño? Justifique su decisión. En caso positivo diga como.

B) ¿Considera que este problema es un problema que puede ser resuelto por aprendizaje? Justifique su decisión. En caso positivo diga:

a) ¿Que tipo de aproximación recomendaría: (1) global para todos los tipos, (2) por tipo de fruta, (3) otra (diga cúal). Justifique su elección.

b) Establezca quien es la función f de su problemna.

c) Un experto le indica que las siguientes variables: RGB del color predominante, grado de esfericidad un valor en [0,1], volumetría en cm^3 , y el tipo de piel, un valor de entre tres posibles, son las características más relevantes. Describa la codificación de cada pareja (x, y) , para que pueda ser usada por el ordenador.

1.- Analice la construcción del clasificador de Random Forest y discuta las características por las que es esperable tenga una alta eficacia.

2.- Compárelo con la SVM-soft y diga si es esperable que Random Forest sea una mejor solución en algunos casos.

3.- ¿Es Random Forest óptimo en algún sentido?

Justifique con solidez y precisión las contestaciones.