APRENDIZAJE AUTOMÁTICO

Trabajo.2: Cuestiones de Teoría

Fecha de entrega: 28 Abril de 2018. Valor máximo: 12.5 puntos + BONUS

NORMAS DE DESARROLLO Y ENTREGA DE TRABAJOS

Para este trabajo como para los demás es obligatorio presentar un informe escrito (hacerlo en pdf, MS Word). **Sin este informe se considera que el trabajo NO ha sido presentado.**

Normas para el desarrollo de los Trabajos: EL INCUMPLIMIENTO DE LAS NORMAS (*) SIGNIFICA PERDIDA DE 2 PUNTOS POR CADA INCUMPLIMIENTO.

- En su informe de contestación debe incluir todas las preguntas en el orden y tal y como se les formula en este documento. (*)
- Las contestaciones irán a continuación de cada pregunta, dejando en blanco las que no conteste.(*)
- Todas las contestaciones deben ser justificadas con argumentos. Sin argumentos la pregunta se considera no contestada.
- Todas las justificaciones matemáticas deben contener todos y cada uno de los pasos de la misma. En caso de duda la contestación no se considerará válida.
- Cualquier desarrollo matemático hecho a mano que no presente la claridad y calidad de un editor de ecuaciones (MS WORD, latex) no se considerará válida. Se recomienda vivamente usar el editor latex para estos casos.
- **Forma de entrega**: Subir el pdf a la web de DECSAI.

Todas las preguntas tienen el mismo valor

1. Identificar de forma precisa dos condiciones imprescindibles para que un problema de predicción puede ser aproximado por inducción desde una muestra de datos. Justificar la respuesta usando los resultados teóricos estudiados.

2. El jefe de investigación de una empresa con mucha experiencia en problemas de predicción de datos tras analizar los resultados de los muchos algoritmos de aprendizaje usados sobre todos los problemas en los que la empresa ha trabajado a lo largo de su muy dilatada existencia, decide que para facilitar el mantenimiento del código de la empresa van a seleccionar un único algoritmo y una única clase de funciones con la que aproximar todas las soluciones a sus problemas presentes y futuros. ¿Considera que dicha decisión es correcta y beneficiará a la empresa? Argumentar la respuesta usando los resultados teóricos estudiados.

3. Supongamos un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraidos de una funcion desconocida $f : \mathcal{X} \to \mathcal{Y}$, donde $\mathcal{X} = \mathbb{R}$ e $\mathcal{Y} = \{-1, +1\}$. Para aprender f usamos un conjunto simple de hipótesis $\mathcal{H} = \{h_1, h_2\}$ donde h_1 es la función constante igual a $+1$ y h_2 la función constante igual a -1. Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). S elige la hipótesis que mejor ajusta los datos y C elige deliberadamente la otra hipótesis.
   a) ¿Puede S producir una hipótesis que garantice mejor comportamiento que la aleatoria sobre cualquier punto fuera de la muestra? Justificar la respuesta

4. Con el mismo enunciado de la pregunta.3:
   a) Asumir desde ahora que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$. ¿Es posible que la hipotesis que produce C sea mejor que la hipótesis que produce S ?. Justificar la respuesta

5. Considere la cota para la probabilidad del conjunto de muestras de error $\mathcal{D}$ de la hipótesis solución g de un problema de aprendizaje, a partir de la desigualdad de Hoeffding, sobre una clase finita de hipótesis,

$$\mathbb{P}[|E_{out}(g) - E_{in}(g)| > \epsilon) < \delta(\epsilon, N, |\mathcal{H}|)$$

   a) Dar una expresión explícita para $\delta(\epsilon, N, |\mathcal{H}|)$
   b) Si fijamos $\epsilon = 0{,}05$ y queremos que el valor de δ sea como máximo 0,03 ¿cual será el valor más pequeño de N que verifique estas condiciones cuando $\mathcal{H} = 1$?.
   c) Repetir para $\mathcal{H} = 10$ y para $\mathcal{H} = 100$

   ¿Que conclusiones obtiene?

6. Considere la cota para la probabilidad del conjunto de muestras de error $\mathcal{D}$ de la hipótesis solución g de un problema de aprendizaje, a partir de la desigualdad de Hoeffding, sobre una clase finita de hipótesis,

$$\mathbb{P}[|E_{out}(g) - E_{in}(g)| > \epsilon) < \delta$$

   a) ¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?
   b) Si elegimos g de forma aleatoria ¿seguiría verificando la desigualdad?
   c) ¿Depende g del algoritmo usado?.
   d) Es una cota ajustada o una cota laxa?

Justificar las respuestas

7. ¿Por qué la desigualdad de Hoeffding no es aplicable de forma directa cuando el número de hipótesis de $\mathcal{H}$ es mayor de 1? Justificar la respuesta.

8. Si queremos mostrar que k^* es un punto de ruptura para una clase de funciones $\mathcal{H}$ cuales de las siguientes afirmaciones nos servirían para ello:
   a) Mostrar que existe un conjunto de k^* puntos $x_1, \ldots, x_{k^*}$ que $\mathcal{H}$ puede separar ("shatter").
   b) Mostrar que $\mathcal{H}$ puede separar cualquier conjunto de k^* puntos.
   c) Mostrar un conjunto de k^* puntos $x_1, \ldots, x_{k^*}$ que $\mathcal{H}$ no puede separar
   d) Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de k^* puntos
   e) Mostrar que $m_{\mathcal{H}}(k) = 2^{k^*}$

9. Para un conjunto $\mathcal{H}$ con $d_{VC} = 10$, ¿que tamaño muestral se necesita (según la cota de generalización) para tener un 95 % de confianza (δ) de que el error de generalización (ϵ) sea como mucho 0.05?

10. Considere que le dan una nuestra de tamaño N de datos etiquetados $\{-1, +1\}$ y le piden que encuentre la función que mejor ajuste dichos datos. Dado que desconoce la verdadera función f, discuta los pros y contras de aplicar los principios de inducción ERM y SRM para lograr el objetivo. Valore las consecuencias de aplicar cada uno de ellos.

BONUS

Los BONUS solo serán tenidos en cuenta si en el cuestionario obligatorio se ha consiguido al menos un 75 % de los puntos totales.

1. (1 puntos) Supongamos que tenemos un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraidos de una funcion desconocida $f : \mathcal{X} \to \mathcal{Y}$, donde $\mathcal{X} = \mathbb{R}$ e $\mathcal{Y} = \{-1, +1\}$. Para aprender f usamos un conjunto simple de hipótesis $\mathcal{H} = \{h_1, h_2\}$ donde h_1 es la función constante igual a $+1$ y h_2 la función constante igual a -1.

   Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). S elige la hipótesis que mejor ajusta los datos y C elige deliberadamente la otra hipótesis. Suponga que hay una distribución de probabilidad sobre $\mathcal{X}$, y sea $P[f(x) = +1] = p$
   a) Si $p = 0{,}9$ ¿Cual es la probabilidad de que S produzca una hipótesis mejor que C?
   b) ¿Existe un valor de p para el cual es más probable que C produzca una hipótesis mejor que S?

2. (1 puntos) Consideremos el modelo de aprendizaje "M-intervalos"donde la clase de funciones H está formada por $h : \mathbb{R} \to \{-1, +1\}$, con $h(x) = +1$ si el punto está dentro de uno de m intervalos arbitrariamente elegidos y -1 en otro caso. Calcular la dimensión de Vapnik-Chervonenkis para esta clase de funciones.