Doble Grado en Ingeniería Informática y Matemáticas

APRENDIZAJE AUTOMÁTICO

Cuestionario 1

UNIVERSIDAD
DE GRANADA

Alberto Estepa Fernández - *albertoestep@correo.ugr.es*

Mayo de 2020

Índice

1. Ejercicio 1

Suponga que disponemos de un conjunto de 1000 muestras etiquetadas de un problema de clasificación binaria a partir de las cuales queremos entrenar 5 modelos distintos y elegir el mejor de ellos. Cada modelo es entrenado usando un algoritmo específico y una clase de funciones finita de tamaño M. Describa qué decisiones tomaría para ello y calcule la mejor cota para el error E_{out} del modelo elegido.

Queremos elegir el mejor de los 5 modelos distintos que tenemos. Necesitamos entrenar los modelos con un conjunto de las muestras del problema etiquetadas, que será el conjunto de entrenamiento (reservaremos un conjunto, el conjunto de validación, de las muestras etiquetadas para testear el modelo). El resultado de entrenar dichos modelos será un conjunto de hipótesis o funciones.

Para elegir que modelo es mejor para nuestros datos y nuestro problema de clasificación mediremos el error de validación (tasa de muestras del conjunto de validación mal clasificadas por nuestra función o hipótesis de cada modelo) para cada una de las hipótesis obtenida y obviamente el mejor será aquel que menor error tenga.

Este procedimiento es el conocido como 'model selection' (información obtenida del libro LEARNING FROM DATA de Yaser S. Abu-Mostafa, Malik Magdon-Ismail y Hsuan-Tien Lin. Capítulo 4, sección 3.2).

Para calcular la mejor cota del error E_{out}, entrenamos el mejor modelo obtenido con el conjunto de entrenamiento, obteniendo una función g_m y usando la desigualdad de Hoeffding:

$$E_{out}(g_m) \leq E_{val}(\bar{g}_m) + O(\sqrt{\frac{\ln M}{K}}) \tag{1}$$

donde K es el tamaño del conjunto validación y $\bar{g}_m$ la hipótesis del modelo elegida.

2. Ejercicio 2

Suponga que le encargan escribir un programa para etiquetar muestras sin etiqueta en un problema de clasificación binario. No se conoce la verdadera función de etiquetado f, pero si se conoce la distribución de las muestras P. ¿Cómo lo haría para garantizar un error de clasificación lo más bajo posible?

Queremos minimizar el error de clasificación para cualquier función de clasificación f que desarrollemos. Si las etiquetas son $\{1, -1\}$ éste error de clasificación se puede escribir como:

$$Error(f) = P[f(x) = 1, y = -1] + P[f(x) = -1, y = 1] \tag{2}$$

Es decir, los puntos mal clasificados. Así si escogemos:

$$f(x) = \begin{cases} 1 & si \quad P[y = 1] \geq P[y = -1] \\ \\ -1 & si \quad P[y = 1] < P[y = -1] \end{cases} \tag{3}$$

Probabilísticamente el error de clasificación es el más bajo posible.

3. Ejercicio 3

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

Como si $yw^T x \geq 1 \rightarrow (max(0, 1 - yw^T x))^2 = 0$, podemos expresar el error de la siguiente manera:

$$E(w) = \begin{cases} (1 - yw^T x)^2 & si \quad yw^T x < 1 \\ \\ 0 & si \quad yw^T x \geq 1 \end{cases} \tag{4}$$

Es trivial comprobar que es continua en 1 y de forma general en todo el dominio. Por otro lado sabemos que la regla de adaptación del gradiente descendente para w es:

$$w_{n+1} = w_n - \eta \nabla E(w_n) \tag{5}$$

Donde $\nabla E(w_n)$ es el gradiente del error evaluado en w_n. Es trivial que $E(w)$ es derivable. Vamos a calcular su valor para un w cualquiera:

$$\nabla E(w) = \begin{cases} -2yx(1 - yw^T x) & si \quad yw^T x < 1 \\ \\ 0 & si \quad yw^T x > 1 \end{cases} \tag{6}$$

Así sustituyendo dicho valor en la ecuación anterior, obtenemos que:

$$w_{n+1} = \begin{cases} w_n - \eta(-2yx(1 - yw^T x)) & si \quad yw^T x < 1 \\ \\ w_n & si \quad yw^T x \geq 1 \end{cases} \tag{7}$$

Por otra parte si ahora definimos como $Error(w) = \frac{1}{N}\sum_{n=1}^{N} E_n(w)$ en vez de $E(w)$, aplicando el mismo procedimiento anterior, donde E_n representa la función de error E en la n-ésima iteración, calculamos $\nabla Error(w)$:

$$\nabla Error(w) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N}(-2y_n x(1 - y_n w^T x_n)) & si \quad y_n w^T x_n < 1 \\ \\ 0 & si \quad y_n w^T x_n > 1 \end{cases} \tag{8}$$

Pero ésto es igual a:

$$\nabla Error(w) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N}\nabla E_n(w) & si \quad yw^T x < 1 \\ \\ 0 & si \quad yw^T x \geq 1 \end{cases} \tag{9}$$

Así, por último sustituyendo en la ecuación de la adaptación del gradiente descendente:

$$w_{n+1} = w_n - \eta\frac{1}{N}\sum_{n=1}^{N}\nabla E_n(w_n) \tag{10}$$

4. Ejercicio 4

Cuando resolvemos un problema de aprendizaje desde datos, diga cúales de las siguientes opciones son correctas: (Justificar la elección)

1. **Después de aprender obtenemos una función g, con garantía de que aproxima bien a la función f, fuera de la muestra.**

   Falso, no tenemos garantía de que aproxima bien a la función f fuera de la muestra.

2. **Después de aprender obtenemos una función g, que con alta probabilidad dicha aproximará a la función f, bien fuera de la muestra.**

   Falso, tampoco tenemos una alta probabilidad de que aproxime a la función f de forma correcta fuera de la muestra. Puede ser que la función g que hemos aproximado no sea de la clase correcta.

3. **Logramos una de estas dos cosas:**

   a) **Obtendremos una función g que con alta probabilidad aproximará a la función f bien fuera de la muestra.**

   b) **Diremos que hemos fallado.**

   Verdadero: la desigualdad de Hoeffding se puede aplicar al problema de aprendizaje, lo que nos permite hacer una predicción fuera de los datos proporcionados para el aprendizaje. Podemos obtener algo de información sobre f, aunque no podemos obtener f de forma general. Podemos obtener la tasa de error que provoca nuestra función h al aproximarse a f. Si dicho error está cerca a cero,

podemos predecir que h se aproximará bien a toda la entrada espacio. Si no, no hemos tenido suerte y diremos que hemos fallado.

(Extraído del libro LEARNING FROM DATA de Yaser S. Abu-Mostafa, Malik Magdon-Ismail y Hsuan-Tien Lin. Capítulo 1, sección 3.2.)

5. Ejercicio 5

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

Es imposible 'averiguar' exactamente la función f con una total garantía de que sea la correcta con dicha información. Los 100 datos son una muestra escasa y no posee propiedades de ser aleatoria, pueden ser escogidos a consciencia para que ningun algoritmo se aproxime la función: por ejemplo coger los 100 elementos de una misma clase y 0 de la otra.

6. Ejercicio 6

Considere de nuevo el escenario descrito para averiguar la función f del matemático. Si el matemático le permitiera añadir una hipótesis extra al problema. ¿Añadiría algo? ¿Qué y por qué?

Deberíamos añadir que la muestra sea aleatoria iid. (independiente e identicamente distribuida), lo que permite usar la desigualdad de Hoeffding para acotar el error de la aproximación.

(Información obtenida del libro de: LEARNING FROM DATA de Yaser S. Abu-Mostafa, Malik Magdon-Ismail y Hsuan-Tien Lin. Apéndice)

7. Ejercicio 7

En regresión hemos introducido la función de error cuadrática $e(x) = (h(x) - y)^2$ entre predicción y etiqueta. Sin embargo medidas aberrantes en los valores de las etiquetas producen valores de error enormes que condicionan la búsqueda de buenas soluciones. Con objeto de rebajar la influencia de estos errores aberrantes podemos decidir usar el valor absoluto de la diferencia en lugar del cuadrado como medida de error, es decir $e(x) = |h(x) - y|$. Suponga que dispone de una muestra de N valores unidimensionales $z = \{z_1 \leq z_2 \leq \ldots \leq z_N\}$. Calcule que función de la muestra $h(z)$ alcanza el mínimo del error $Error = \sum_i |h(z) - z_i|$ y argumente a

la vista del estimador que le salga si el cambio de norma para eliminar observaciones aberrantes es una buena o mala decisión.

Derivando el error (que está en función de una función de la muestra) e igualando a cero obtendremos el mínimo del error. Así la derivada del error es:

$$\sum_i \frac{h(z) - z_i}{|h(z) - z_i|} \; si \; h(z) \neq z_i \forall i \tag{11}$$

En dichas condiciones, calculamos el mínimo, pero nos damos cuenta que, la derivada toma el valor 0 si el número de elementos de la muestra se reparten de forma equitativa entre los que son menores que $h(z)$ y los que son mayores que $h(z)$ (ya que los valores de los sumandos son 1 o -1 al estar divididos por el modulo), es decir, por ejemplo, error alcanza el mínimo en la mediana de z, (aunque hay otras funciones que pueden cumplir esta propiedad).

Repitiendo el experimento del enunciado para ésta función de la muestra (la mediana) obtenemos que si perturbamos de forma aberrante un punto (por ejemplo haciendolo muy grande), la mediana no sufre modificaciones grandes (puede ocurrir que el punto éste en el lado izquierdo de h(z) en la recta real y al hacerlo grande, pase al lado derecho o viceversa, lo que podría modificar de forma leve el resultado, pero no son cambios aberrantes.).

Así como el objetivo era reducir la influencia de los errores aberrantes, ésta nueva norma resulta bastante eficaz.

8. Ejercicio 8

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

$$P(y|x) = \begin{cases} \lambda & si \quad y = f(x) \\ \\ 1 - \lambda & si \quad y \neq f(x) \end{cases} \tag{12}$$

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

El error que comete h al aproximar y es $(1 - \lambda)(1 - \mu) + \lambda\mu$.

Ésto es debido a que sabemos que $P[h \neq f] = \mu$, entonces $P[h \neq f] = 1 - \mu$

Notando por g a la versión ruidosa de f, tenemos por el Teorema de la Probabilidad Total que:

$$P[h=g] = P[h=f]P[h=g|h=f] + P[h\neq f]P[h=g|h\neq f] = (1-\lambda)(1-\mu)+\mu\lambda \tag{13}$$

Vemos claramente que con un valor de $\lambda = 0.5$, el error de h es independiente de μ, ya que:

$$P[h=g] = \mu\lambda + (1-\lambda)(1-\mu) = \mu\lambda + 1 - \mu - \lambda + \mu\lambda = 2\mu\lambda + 1 - \mu - \lambda \tag{14}$$

Así sacando factor común μ obtenemos

$$P[h=g] = \mu(2\lambda - 1) + 1 - \lambda \tag{15}$$

Para eliminar μ de la ecuación, necesitamos que $(2\lambda - 1) = 0$. Así $\lambda = 0.5$.

9. Ejercicio 9

Consideremos un problema de clasificación binaria probabilístico con etiquetas $\{1, -1\}$ donde la solución probabilística sin costes está dada por la función $g(x) = P[y = 1|x]$. Suponga que su problema tiene asociada una matriz de coste que está dada por:

Costes

	Veradero +1	Verdadero -1
desición: +1	0	10
desición: -1	1000	0

Identifique la regla de clasificación asociada a $g(x)$. A continuación incorpore los costes a dicha función y deduzca una nueva regla de clasificación que incopore la información de los costes. Expresar con claridad los pasos que vaya dando y las hipótesis que haga para calcular la solución. ¿Cómo influyen los pesos en la regla final?

(Análogo al problema 3.16 del libro de LEARNING FROM DATA de Yaser S. Abu-Mostafa, Malik Magdon-Ismail y Hsuan-Tien Lin)

Consideramos la regla de clasificación asociada a $g(x)$ como:

$$f(x) = \begin{cases} 1 \quad si \quad P[y=1|x] \geq 0.5 \\ \\ 0 \quad si \quad P[y=-1|x] < 0.5 \end{cases} \tag{16}$$

Hemos tomado un umbral de $k = 0.5$ porque es lo habitual. Para incorporar la información de los costes en la función, lo habitual es calcular el umbral como:

$$k = \frac{10}{1000 + 10} \tag{17}$$

Así obtenemos una nueva regla de clasificación asociada a $g(x)$ como:

$$f(x) = \begin{cases} 1 & si \quad P[y=1|x] \geq \frac{1}{101} \\ \\ 0 & si \quad P[y=-1|x] < \frac{1}{101} \end{cases} \tag{18}$$

De esta forma, los pesos influyen notablemente en la regla final. Donde antes necesitabamas que $g(x) \geq 0.5$ para asignarle la etiqueta $+1$, ahora necesitamos simplemente que $g(x) \geq 0.0099$

10. Ejercicio 10

Es bien conocido que la investigación en 'Machine Learning' ha ido produciendo paulatinamente distintos algoritmos de considerable éxito para aproximar problemas de clasificación: k-NN, RL, Árboles, SVM, RRNN, AdaBoost, etc. Sin embargo en los últimos años la redes CNN han mostrado una increíble superioridad frente a las técnicas anteriores en multitud de problemas. Si le piden, como experto, que elija una técnica como la mejor entre las mencionadas, ¿cúal elegiría y por qué?

Nota: k-NN: k vecinos más cercanos, RL: Regresión logística., RRNN: redes neuronales, SVM: Maquinas de vectpres de Soporte, CNN: redes neuronales convolucionales

(Consultado en https://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf)

No podemos elegir una técnica como la mejor entre las anteriormente mencionadas. Ésto es lo que recoge el Teorema de No-Free-Lunch. Concretamente sabemos que existe al menos una distribución de probabilidad para la cuál nuestro algoritmo falla. Por tanto, si nos encontramos con tal distribución el algoritmo no dará buenos resultados.

11. Ejercicio 11

Suponga un problema de predicción bien definido que hace uso de una clase de funciones H. Suponga que dicha clase H es suficiente como para hacer que $E_{in} \to 0$ para cualquier muestra de un tamaño dado N. Analice las fuentes del error en el problema de sobreajuste de esta situación y su evolución cuando se va reduciendo la complejidad de la clase H.

Si E_{in} tiende a cero para cualquier muestra finita es un indicador claro de que puede que el modelo tenga un grado elevado de sobreajuste.

Para solucionar ésto podemos ir reduciendo la complejidad de la clase H, lo que disminuye el sobreajuste pero aumenta posiblemente el error del modelo escogido.

12. Ejercicio 12

Considere $E_n(w) = max(0, 1 - y_n w^T x_n)$. Mostrar que $E_n(w)$ es una cota superior para $[[sign(w^T x_n) \neq y_n]]$. Por tanto $\frac{1}{N}\sum_n E_n(w)$ es una cota superior para E_{in}.

Recordemos que la notacion $[[x]] \in \{0, 1\}$ y valdrá 1 si x se evalua a True o 0 si x se evalua a False.

Sea $[[y_n \neq signo(w^T x_n)]] = 1 \rightarrow y_n \neq signo(w^T x_n) \rightarrow y_n signo(w^T x_n) < 0$.

Así, deducimos que $y_n w^T x_n < 0$ y entonces $-y_n w^T x_n > 0$. Sumando así 1 a los dos miembros obtenemos que $E_n(w) = 1 - y_n w^T x_n > 1$. Así hemos comprobado que $[[sign(w^T x_n) \neq y_n]] < E_n(w)$ si $[[y_n \neq signo(w^T x_n)]] = 1$.

Por otro lado, sea $[[y_n \neq signo(w^T x_n)]] = 0$, que se da si $y_n = signo(w^T x_n)$, pero se deduce trivialmente que $[[sign(w^T x_n) \neq y_n]] \leq E_n(w)$ si $[[y_n \neq signo(w^T x_n)]] = 0$, ya que $E_n \geq 0$ siempre.

Por tanto hemos comprobado que $E_n(w)$ es una cota superior de $[[sign(w^T x_n) \neq y_n]]$.

Para concluir el ejercicio, tenemos que $E_{in} = \frac{1}{N}\sum_{n=1}^{N}[[sign(w^T x_n) \neq y_n]] \leq \frac{1}{N}\sum_{n=1}^{N} E_n$. Así comprobamos que $\frac{1}{N}\sum_{n=1}^{N} E_n$ es una cota superior de E_{in}.