Universidad Nacional de Colombia
Departamento de Matemáticas
Matemáticas para el aprendizaje de máquina
Ejercicios libro (I-2023)

Daniel Santiago Pardo Gómez .....

1. Suponga que usamos un perceptron para detectar mensajes de spam. Podemos decir que cada correo es está representado por la frecuencia de ocurrencia de palabras clave, y la salida de el mensaje es +1 si el mensaje es considerado spam.

   a) Puedes pensar algunas palabras que tendrán con un peso grande en el perceptron?
   Algunas palabras con un peso grande podrían ser: Gratis, enlace, ganar, deuda, descarga, registrarse.

   b) ¿Que palabras tendrán un peso negativo?
   Palabras con un peso negativo pueden ser palabras que aparecen en correos comunes entre personas: Buenos días, gracias, cordialmente

   c) ¿Que parámetro en el perceptron afectará directamente afecta cuantos mensajes frontera terminaran siendo clasificados como spam?
   El parámetro b o bias determinara si el correo es spam o no, pues representa un desplazamiento del hiperplano de separación.

2. La regla de actualización de peso en (1.3) tiene una buena interpretación de lo que se mueve correctamente en dirección de clasificar $x(t)$:

   a) Muestre que $y(t)w^T(t)x(t) < 0$.
   Note que como t es alguna iteración el valor de $x(t)$ estará mal clasificado por $w(t)$, esto quiere decir que la predicción $w^T(t)x(t)$ tendrá signo diferente a $y(t)$, por lo tanto, $y(t)w^T(t)x(t) < 0$.

   b) Muestre que $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$.
   Recordemos que la regla de actualización de la regla $w(t+1) = w(t) + y(t)x(t)$, entonces:

$$\begin{aligned} y(t)w^T(t+1)x(t) &= y(t)(w(t) + y(t)x(t))^T x(t) \\ &= y(t)w^T(t)x(t) + y(t)y^T(t)x^T(t)x(t) \\ &> y(t)w^T(t)x(t) \end{aligned}$$

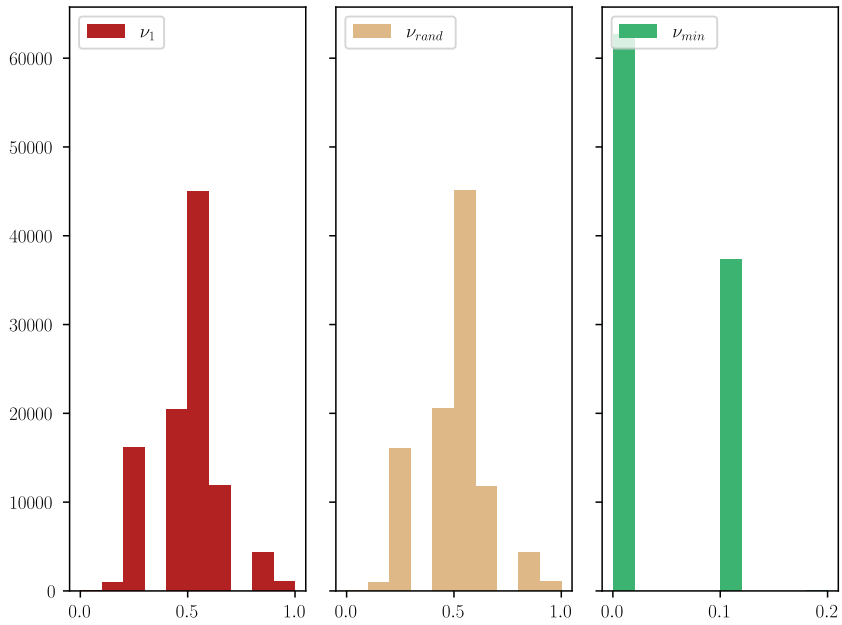
   La ultima desigualdad está dada por $y(t)y^T(t)x^T(t)x(t) = \|x\|\|y\| \leq 0$.

c) En lo que respecta a la clasificación de $x(t)$, argumenta que el movimiento desde $w(t)$ a $w(t+1)$ es un movimiento "en la dirección correcta".
Como probamos en *a*) mientras en todas las iteraciones $x(t)$ esté mal clasificado tendremos que $y(t)w^T(t)x(t) < 0$, de manera análoga, cuando tengamos que $x(t)$ está bien clasificado los signos de $y(t)$ y $w^T(t)x(t)$ serán iguales y $y(t)w^T(t)x(t) \geq 0$, como vimos en el punto anterior en cada iteración el valor de $y(t)w^T(t)x(t)$ aumenta, esto quiere decir que w^T va en la dirección correcta.

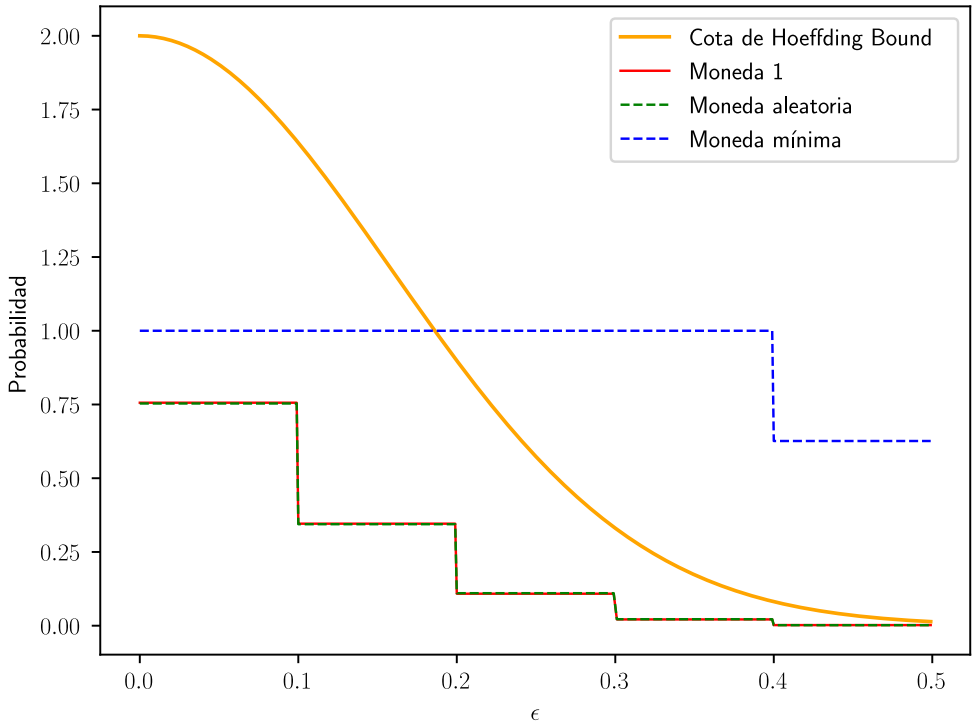
3. Este es un experimento que ilustra la diferencia entre un solo contenedor y múltiples contenedores. Realiza una simulación por computadora lanzando 1,000 monedas. Lanza cada moneda de forma independiente 10 veces. Enfoquémonos en 3 monedas de la siguiente manera: c_1 es la primera moneda lanzada; c_{rand} es una moneda que eliges al azar; $c_{\text{mín}}$ es la moneda que tuvo la frecuencia mínima de caras (elige la más temprana en caso de empate). Sean ν_1, ν_{rand} y $\nu_{\text{mín}}$ las fracciones de caras obtenidas para las respectivas tres monedas.

   a) ¿Cuál es el valor esperado μ para las tres monedas seleccionadas?
   Como son monedas convencionales su distribución de probabilidad es uniforme, por lo tanto las tres monedas tendrán el mismo valor esperado de caras de $\frac{1}{2}$.

   b) Repite este experimento completo un gran número de veces (por ejemplo, 100,000 ejecuciones del experimento completo) para obtener varias instancias de ν_1, ν_{rand} y $\nu_{\text{mín}}$, y traza los histogramas de las distribuciones de ν_1, ν_{rand} y $\nu_{\text{mín}}$. Observa que las monedas que resulten ser c_{rand} y $c_{\text{mín}}$ pueden variar de una ejecución a otra.
   Tras hacer la simulación (ver el código en el notebook) obtenemos los siguientes histogramas:



c) Utilizando *b*), traza las estimaciones para $\mathbb{P}[|\nu - \mu| > \epsilon]$ en función de ϵ, junto con la cota de Hoeffding $2e^{-2\epsilon^2 N}$ (en el mismo gráfico).



d) ¿Que monedas cumplen con la cota de Hoeffding y cuáles no? Explica por qué.

La primera moneda y la moneda aleatoria cumplen con la cota de Hoeffding, pero

la moneda con la mínima cantidad de caras no la cumple. Esto puede explicarse por la selección de la moneda después del experimento, pues estamos tomando los datos que tengan una característica particular, esto hace que el análisis probabilístico del que viene la desigualdad no se cumpla.

e) Relaciona la parte *d*) con los múltiples recipientes en la Figura 1.10.
La figura 1.10 muestra como elegir un contenedor para cada hipótesis, en el experimento de las monedas podemos elegir que moneda escoger antes de hacer el experimento, de esta manera nos aseguramos de que el experimento si pueda ser tratado como una variable aleatoria.

4. Se nos proporciona un conjunto de datos $\mathcal{D}$ con 25 ejemplos de entrenamiento de una función objetivo desconocida $f : \mathcal{X} \to \mathcal{Y}$, donde $\mathcal{X} = \mathbb{R}$ y $\mathcal{Y} = \{-1, +1\}$. Para aprender f, utilizamos un conjunto de hipótesis simple $\mathcal{H} = \{h_1, h_2\}$ donde h_1 es la función constante +1 y h_2 es la función constante -1.
Consideramos dos algoritmos de aprendizaje, S (inteligente) y C (loco). S elige la hipótesis que concuerda más con $\mathcal{D}$ y C elige deliberadamente la otra hipótesis. Veamos cómo se desempeñan estos algoritmos fuera de la muestra desde los puntos de vista determinista y probabilístico. Supongamos, en el punto de vista probabilístico, que hay una distribución de probabilidad en $\mathcal{X}$, y sea $\mathbb{P}[f(\mathbf{x}) = +1] = p$.

   a) ¿Puede S producir una hipótesis que garantice un mejor rendimiento que el azar en cualquier punto fuera de $\mathcal{D}$?
   No, S no puede garantizar una mejor hipótesis dado que el espacio de hipótesis es muy limitado y en muchos casos la muestra $\mathcal{D}$ con la que trabajamos puede dar una idea errada de la función f.

   b) Supongamos, para el resto del ejercicio, que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$. ¿Es posible que la hipótesis que produce C resulte ser mejor que la hipótesis que produce S?
   Si, puede pasar que $\mathcal{D}$ tenga ejemplos de la función con valor -1, pero para valores en $\mathcal{X} - \mathcal{D}$ la función tenga imágenes en +1, para este caso si el método probabilístico acierta tendrá una mejor hipótesis/

   c) Si $p = 0{,}9$, ¿cuál es la probabilidad de que S produzca una hipótesis mejor que C?
   Para que C produzca una mejor hipótesis debe ocurrir que C elija a h_1 lo que ocurre con una probabilidad de $\frac{1}{2}$ y que S tome h_2 esto sucede si menos de la mitad de ejemplos tienen como valor a +1 esto ocurre con una probabilidad de:

$$\sum_{i=0}^{12} (0{,}9)^i (0{,}1)^{12-i} \approx 1{,}62 \times 10^{-7}$$

Dado que ambos eventos son independientes tendremos que la probabilidad de obtener una mejor hipótesis es de aproximadamente $8{,}1 \times 10^{-8}$, entonces S obtendrá una hipótesis mejor o igual a la de C con una probabilidad de $1 - 8{,}1 \times 10^{-8}$.

d) ¿Existe algún valor de p para el cual sea más probable que no que C produzca una hipótesis mejor que S?
Si, en el ejemplo anterior eso sucedía.

5. Una amiga se acerca a ti con un problema de aprendizaje. Dice que la función objetivo f es completamente desconocida, pero tiene 4,000 puntos de datos. Está dispuesta a pagarte para resolver su problema y producirle un g que aproxime a f. ¿Qué es lo mejor que puedes prometerle de las siguientes opciones?

   a) Después de aprender, le proporcionarás un g que garantizará una buena aproximación de f fuera de la muestra.

   b) Después de aprender, le proporcionarás un g, y con alta probabilidad el g que produzcas aproximará bien a f fuera de la muestra.

   c) Ocurrirá una de las dos cosas.

      1) Producirás una hipótesis g;
      2) Declararás que fracasaste.

      Si devuelves una hipótesis g, entonces con alta probabilidad el g que produzcas aproximará bien a f fuera de la muestra.

   La respuesta más sensata para mi amiga es responderle con la opción *c*) pues no tenemos mucha información a cerca de la complejidad de la función objetivo, por lo tanto, la cantidad datos que nos proporciona podrían o no asegurar aprendizaje con un error pequeño.