Primer Examen Parcial Modelado Predictivo - Equipo 3

Carpio Miranda Dan Benjamin
De La Cruz Martinez Jorge Luis
Elizalde Baez Regina
Maravilla Pérez Vianey
Roman Pardo Alejandro
Vázquez Portuguez José Antonio

Pregunta 1

Suponga que utiliza un perceptrón para detectar mensajes de spam. Digamos que cada correo electrónico es representado por la frecuencia de ocurrencia de palabras clave y la salida será +1 si el mensaje es considerado spam.

A).¿Cuáles son algunas palabras claves que podrían terminar con un peso positivo muy grande en el perceptrón?
Las palabras clave que tienen una alta frecuencia en los correos electrónicos de spam pueden terminar con un peso positivo muy grande. Estas palabras clave suelen aparecer con frecuencia en los mensajes de spam y, por lo tanto, es probable que tengan un peso positivo alto en el perceptrón. Cabe señalar que la selección de las palabras clave y el ajuste de los pesos dependen de la calidad y la cantidad de los datos de entrenamiento utilizados para el perceptrón. Algunos ejemplos de palabras son:

Suscríbete, Nuevo, oferta, Ganaste, hoy, tiempo, limitado, premio, pareja, dinero, felicidades,descubrir,creerás,Viagra,Oferta,Gratis,Dinero,Ganar,Sorpresa,Inversión,Descuento,Promoción,Urgente.

B). ¿Cuáles son algunas palabras claves que podrían terminar con un peso negativo en el perceptrón?
Las palabras clave que tienen una baja frecuencia en los correos electrónicos de spam y una alta frecuencia en los correos electrónicos legítimos (no spam) pueden terminar con un peso negativo en el perceptrón.

Trabajo, Carrera, Profesional, Reunión, Proyecto, Información, Ayuda, Consulta ,Pregunta, Respetuosamente, Horario, Tiempo, Recibo.

Pregunta 2

Considere un perceptrón en dos dimensiones:
$h(x) = sign(w^T x)$ $donde\ w = [w_0, w_1, w_2]^T\ y\ x = [1, x_1, x_2]^T$. Técnicamente, x tiene tres coordenadas pero se llama perceptron en dos dimensiones por que la primera coordenada está fijada a 1.

2.1.- Muestra que las regiones en el plano donde $h(x) = +1$ y $h(x) = -1$ están separadas por una línea.

$$h(x) = \text{sign}(\bar{w}^T \bar{x}) = \text{sign}(y(x)) \quad \text{donde} \quad h(x) \begin{cases} 1 & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x = 0 \\ -1 & \text{si} \quad x < 0 \end{cases}$$

$$x_2 = a x_1 + b$$

a es la pendiente para $a=0$, cuando $a>0$ la línea es horizontal.

b es el desplazamiento. x_2 queda en $x_2 = b$ para la línea que pasa por el origen si y sólo si $b=0$

Entonces igualamos $x_2 = 0$ ya que existe una línea que lo separa

$$x_2 = a x_1 + b \iff -a x_1 + x_2 + b = 0$$

$$x_2 \iff [a, -1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + b = 0$$

$$\iff [b, -a, 1]\begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

Por lo tanto, existe una línea donde x_2 es 0.

2.2- Si expresamos esta línea por la ecuación $x_2 = ax_1 + b$, ¿Cuál es la pendiente a y la intercepción b en términos de w_0, w_1 y w_2?

Entonces tenemos que la separación será una línea con la siguiente ecuación: $w^T x = w_0 + w_1 x_1 + w_2 x_2 = 0$

b) Si tenemos $x_2 = ax_1 + b$

Despejando x_2 de a

$$x_2 = \frac{-w_0 - w_1 x_1}{w_2} = \frac{-w_0}{w_2} - \frac{w_1 x_1}{w_2}$$

Si consideramos $x_2 = ax_1 + b$

$$\Rightarrow = \underbrace{\frac{-w_0}{w_2}}_{b} - \underbrace{\frac{w_1}{w_2}}_{a} x_1$$

$\therefore$ tendremos $\underbrace{a = -\frac{w_1}{w_2}}_{\text{pendiente}}$ y $\underbrace{b = \frac{-w_0}{w_2}}_{\text{Intercepción}}$

Pregunta 3

Una manera de descomponer el valor esperado del error fuera del entrenamiento respecto a cualquier dataset D está dada por:

$$E_D[E_{out}(g^{(D)})] = E_x[E_D[(g^{(D)}(x) - \overline{g}(x))^2] - (\overline{g}(x) - f(x))^2]$$

En donde

$$bias = E_X[E_D[(g^{(D)}(x) - \overline{g}(x))^2]]$$

y

$$varianza = E_x[\overline{g}(x) - f(x))^2]$$

Interprete esta descomposición y explique cuál sería su utilidad en la práctica

Descomponer el valor esperado del error en el entrenamiento respecto a cualquier conjunto del dataset, nos permite entender lo que contribuye relativamente las dos fuentes del error: El sesgo y la varianza, en el error total de nuestro modelo.

En la práctica, esta descomposición puede ayudar a entender mejor cómo mejorar los modelos. Si un modelo tiene un alto sesgo, podemos considerar aumentar la complejidad del modelo o utilizar un modelo más avanzado o distinto. Si un modelo tiene una alta varianza, podemos considerar reducir la complejidad del modelo o utilizar técnicas de regularización. Y también se usa para combatir o detectar que tenemos un overfitting o un underfitting en nuestro modelo, pues si tenemos bajo error en entrenamiento y alto en prueba, refleja una varianza alta. Y si tenemos errores altos en ambas pruebas, seguramente el modelo tendrá un alto sesgo.

Pregunta 4

Realice lo que se indica

1. Muestre que si $\mathcal{H}$ es cerrado bajo la combinación lineal (es decir, cualquier combinación lineal de hipótesis $g \in \mathcal{H}$ también es una hipótesis en $\mathcal{H}$), entonces $\bar{g} \in \mathcal{H}$, donde

$$\bar{g} = \frac{1}{K}\sum_{k=1}^{K} g_k(\mathbf{x})$$

2. De un modelo en donde la función promedio $\bar{g}$ no esté en el conjunto de hipótesis del modelo. [Tip: Use un modelo muy simple].

1.-

Suponemos $h_i \in$ *y* H *es finita.*

$h' = \alpha_1 h_1 + \alpha_2 h_2 + \dots + \alpha_{|H|} h_{|H|} \in$, *que es una combinación lineal*

$g_j \in H$

$\overline{g} = \frac{1}{N} g_1 + \frac{1}{N} g_2 + \dots + \frac{1}{N} g_N$

por lo tanto $\overline{g} \in H$

4.2

Supongamos un conjunto de hipótesis $H = \{g_1(x) = c, \quad g_2(x) = 0\}$

Por lo que:

$$\bar{g} = \frac{1}{2}(c + 0) = \frac{c}{2}$$

Esta $\bar{g}$ no pertenece a H.

Pregunta 5

5. Para funciones binarias, muestre que $P_r[[h(x) \neq f(x)]]$ puede ser escrita cómo el valor esperado de la métrica de error cuadrático medio en los siguientes casos:

5.1. La convención usada para la función binaria es 0 o 1 (en lugar de -1 o +1).

5.2. La convención usada para la función binaria es -1 o +1.

5 Recordemos que podemos escribir $P[h(x) \neq f(x)]$ de la siguiente forma para una función binaria

$$P[h(x) \neq f(x)] = P[h(x)=0, f(x)=1] + P[h(x)=1, f(x)=0]$$

Buscamos exhibir la probabilidad de error como el valor esperado del error cuadrático medio de la siguiente manera.

$$P[h(x) \neq f(x)] = E[(h(x)-f(x))^2]$$

a) Podemos probar los valores con la etiqueta 0, 1 con el error cuadrático medio

$$(h(x)-f(x))^2 \Rightarrow (0-0)^2 = 0 \qquad (1-1)^2 = 0 \rightarrow h(x)=f(x)$$

$$(1-0)^2 = 1 \qquad (0-1)^2 = 1 \rightarrow h(x) \neq f(x).$$

Observamos que se cumple entonces: $\frac{1}{N}\sum_{n=1}^{N}(h(x_n)-f(x_n))^2$

$f(x_n) \in \{0,1\}$

b) Con los valores de -1, +1 tenemos:

$$(h(x)-f(x))^2 \Rightarrow (-1-(-1))^2 = 0 \qquad (1-1)^2 = 0 \longrightarrow h(x)=f(x)$$

$$(1-(-1))^2 = 4 \qquad (-1-1)^2 = 4 \longrightarrow h(x) \neq f(x)$$

Observamos que ambas pueden escribirse entonces como el valor esperado, cambiando únicamente la escala.

Podemos descomponerlo también de la siguiente manera.

$$P[h(x) \neq f(x)] = P[h(x) \neq f(x)](h(x)-f(x))^2 + P[h(x)=f(x)](h(x)-f(x))^2$$

$$= P[h(x) \neq f(x)](1) + P[h(x)=f(x)](0)$$

$$\therefore P[h(x) \neq f(x)] = E[(h(x)-f(x))^2]$$