Ejercicios del libro

Realizado por: María Alejandra Perpiñán Barnés

1.2 Supongamos que usamos un perceptrón para detectar mensajes de spam. Digamos que cada mensaje de correo electrónico esta representado por la frecuencia de aparición de las palabras clave y el resultado es +1 el mensaje se considera spam.

a) ¿Puedes pensar en algunas palabras clave que terminarán con un gran valor positivo de peso en el perceptron?
- * Gratis
- * Oferta
- * Ganador
- * Descuento
- * Dinero
- * Urgente

b) ¿Cuáles palabras clave obtendrán un peso negativo?
- * Confirmación
- * Notificación
- * Factura
- * Soporte
- * Cancelación

c) ¿Qué parámetro en el perceptrón afecta directamente cuántos mensajes límite terminan siendo clasificados como spam?
El parámetro b, esto es porque determina el umbral para clasificar los correos en las categorías de spam y no spam

1.3 La regla de peso en la ecuación 1.3 tiene la buena interpretación de que se mueve en la dirección de clasificar $x(t)$ correctamente

a) Mostrar que $y(t)\, w^T(t)\, x(t) < 0$
Si asumimos que $x(t)$ es mal clasificado por w^T, entonces $w^T(t)x(t)$ tiene signos diferentes a $y(t)$, entonces
$$y(t)\, w^T(t)\, x(t) > 0$$

b) Mostrar que $y(t)w^T(t+1)x(t) > y(t)\, w^T(t)\, x(t)$
$$\begin{aligned} y(t)w^T(t+1)\, x(t) &= (y(t)\, w(t) + y(t)x(t))^T x(t) \\ &= y(t)\, w^T(t)\, x(t) + y(t)\, x^T(t)x(t) \\ &> y(t)\, w^T(t)x(t) \quad \text{ya que el término } y(t)\, x^T(t)\, x(t) \geq 0 \end{aligned}$$

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"
Del problema anterior sabemos que $y(t)\, w^T(t)\, x(t)$ está aumentando con cada actualización
Si $y(t)$ es positivo pero $w^T(t)x^T$ es negativo entonces movemos a $w^T(t)\, x(t)$ hacia una dirección positiva si lo aumentamos
Si en cambio $y(t)$ es negativo pero $w^T(t)\, x^T$ es positivo, $y(t)\, w^T(t)\, x(t)$ aumenta lo que significa que $w^T(t)\, x(t)$ esta disminuyendo. es decir moviendose en una dirección negativa.
Así el movimiento desde $w(t)$ a $w(t+1)$ es un movimiento en la "dirección correcta" ya que tiene como objetivo corregir la clasificación errónea ajustando el vector de peso de una manera que lo acerque a la clasificación correcta.

1.11 Tenemos un conjunto de datos $\mathcal{D}$ de 25 muestras de entrenamiento de una función objetivo desconocida $f: \mathcal{X} \longrightarrow \mathcal{Y}$ donde $\mathcal{X} = \mathbb{R}$ y $\mathcal{Y} = \{-1, +1\}$. Para aprender f, usamos un conjunto de hipótesis simple $\mathcal{H} = \{h_1, h_2\}$ donde h_2 es la función constante $+1$ y h_1 es la constante -1.
Consideremos dos algoritmos de aprendizaje, S (inteligente) y C (loco). S elige la hipótesis que más concuerda con $\mathcal{D}$ y C escoge la otra hipótesis deliberadamente. Veamos cómo funcionan estos algoritmos out of sample desde un punto de vista determinista y probabilístico
Suponga que en el punto de vista probabilístico hay una distribución de probabilidad en $\mathcal{X}$ y sea $\mathbb{P}[f(x) = +1] = p$

a) ¿Puede S producir una hipótesis que garantice un mejor desempeño que el aleatorio en cualquier punto fuera de $\mathcal{D}$?
Dado que el conjunto de hipótesis $\mathcal{H}$ para S consta de solo dos funciones constantes ($h_1 = +1$ y $h_2 = -1$) significa que S solo puede elegir una de estas dos funciones en función de los datos de entrenamiento $\mathcal{D}$. S seleccionara la hipótesis que más concuerda con $\mathcal{D}$, lo que significa que elegira la etiqueta mayoritaria en $\mathcal{D}$
Si S tiene 25 $+1$ en $\mathcal{D}$ pero -1 en los otros puntos en $\mathcal{X}$, S escogera la hipotesis h_1, lo cual no coincidira con $\mathcal{D}$ fuera de f.
Por otra parte una función aleatoria tendra $+1$ y -1 en una proporción 50/50 y coincidira con f la mitad del tiempo, lo que es mejor que la función producida por S

b) Supongamos que para el resto del ejercicio todas las muestras 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 todas las muestras $\mathcal{D}$ tienen $y_n = +1$ entonces S siempre escogera la hipótesis h_1 ($h_1 = +1$) y C escogera por defecto la hipotesis h_2.
En este caso como todas las muestras tienen $y_n = +1$ la hipótesis h_1 va a clasificar perfectamente todas las muestras en $\mathcal{D}$ resultando en un error de entrenamiento igual a cero. Sin embargo h_1 no va a generalizar bien para los puntos por fuera de $\mathcal{D}$ porque siempre los va a predecir como $+1$, por lo tanto el error out of example sera malo.

Por otro lado, la hipótesis h_2 escogida por C es la función constante -1, por lo cual el error de entrenamiento será 100%, sin embargo como nos interesa el error out of sample h_2 puede que funcione mejor que h_1 en los puntos de fuera de D ya que al ser desconocida la función objetiva por fuera de D entonces C tiene más probabilidad de predecir estos puntos

e) Si $p = 0.9$, ¿Cuál es la probabilidad de que S produzca una mejor hipótesis que C?

Como sabemos S siempre escogerá la hipótesis que mas concuerda con D. En este caso todas las muestras en D tiene $y_n = +1$. así que S siempre escogerá h_1. En este caso hay un 90% de probabilidad que una nueva muestra tenga $y_n = 1$ en ese caso h_1 siempre será hipótesis correcta. habrá solo un 10% de probabilidad de que una muestra y_n sea igual a -1.

Por lo tanto:

$P(S \text{ produzca una mejor hipótesis}) = P(\text{nuevas muestras } y_n = +1) * (S \text{ escoge } h_1) + P(\text{nuevas muestras } y_n = -1) * (S \text{ escoge } h_2)$

$= 0.9 * 1 + 0.1 * 0 = 0.9 = 90\%$

d) ¿Hay algún valor de p para el cual es más probable que C produzca una hipótesis mejor que S?

Si $p < 0.5$ entonce C podrá producir una hipótesis mejor que S, ya que C siempre produce h_2 que tendrá más probabilidad de pasar en caso de que $p < 0.5$.

1.12 Una amigo acude a ti con un problema de aprendizaje. Ella dice que la función objetivo f es completamente desconocida pero tiene 4000 muestras. Ella esta dispuesta a pagarte para que resuelvas el problema y crear un g que aproxime a f. ¿Entre lo siguiente que es lo mejor que le puedes prometer?

a) Después de aprender te daré un g que garantizara que se aproxima bien a f out of sample

b) Después de aprender le daras un g y con una probabilidad alta el g que le daras se aproximará bien a f out of sample

c) Una de dos cosas pasará:

I Produciras una hipótesis g

II Tu declararas que fallaste

Si entregas una hipótesis entonces lo haras con una probabilidad alta el g que le daras se aproximará bien a f out of sample

lo mejor que puedo prometer es b, no puedo garantizar que el g que produzca va a aproximar a f bien out of sample, esto es porque no conozco f. Sin embargo puedo decir que con una alta probabilidad el g que produzco va aproximar a f bien out of sample, esto es porque tengo 4000 muestras lo que es suficiente mente grande para entrenar un modelo que generalice bien sobre nuevos datos. la opcion no es realista porque no es posible declarar que he fallado, solo se puede decir que he fallado si soy incapaz de encontrar una hipótesis f que ajuste los datos y con 4000 muestras soy capaz de hacer esto.