Aprendizaje Automático - Cuestionario 3

Juan Luis Suárez Díaz

29 de mayo de 2017

Cuestión 1

Considere un modelo de red neuronal con dos capas totalmente conectadas: n_I unidades de entrada, n_H unidades ocultas y n_O unidades de salida. Considere la función de rror definida por $J(w) = \frac{1}{2}\sum_{k=1}^{n_O}(t_k - c_k)^2 = \frac{1}{2}\|t - c\|^2$, donde el vector t representa los valores de la etiqueta, c los valores calculados por la red y w los pesos de la red. Considere que las entradas a la segunda capa se calculan como $z_k = \sum_{j=0}^{N_H} y_j w_{kj}$, donde el vector y representa la salida de la capa oculta.

a) *Deducir con todo detalle la regla de adaptación de los pesos entre la capa oculta y la capa de salida.*

b) *Deducir con todo detalle la regla de adaptación de los pesos entre la capa de entrada y la capa oculta.*

Usar θ para notar la función de activación.

Solución

La regla de adaptación de pesos para w, de acuerdo con el gradiente descendiente viene dada por $w \leftarrow w - \eta \nabla E_{in}(w)$, donde η es una tasa de aprendizaje prefijada. Para calcular el gradiente debemos calcular por tanto las derivadas parciales respecto a cada elemento en el vector de pesos. Llamamos x_i a los datos de entrenamiento y t_i a las etiquetas. Llamamos z_i a las entradas a la capa de salida y c_i a las salidas de la capa de salida (y de la red neuronal). Para la capa oculta, llamamos n_i a las entradas a la capa e y_i a las salidas. Finalmente, llamamos r_i a las salidas de la capa de entrada. Consideramos los pesos como $w_{ji}^{(k)}$, donde j representa el índice de la conexión de salida, i el índice de la conexión de entrada, y k las capas entre las que se encuentra. Es decir, la distribución de la red es la que se muestra en la figura 1.

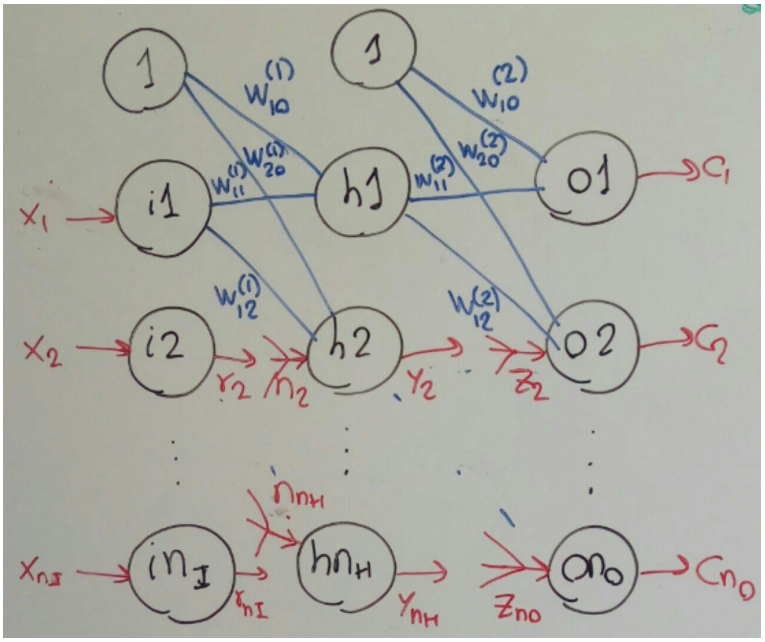


Figura 1: Distribución de la red neuronal.

Calculamos la regla de adaptación de pesos para los pesos entre la capa oculta y la de salida, $w_{ji} \equiv w_{ji}^{(2)}$. Es decir, buscamos el valor de $\frac{\partial J}{\partial w_{ji}}$. Aplicando la regla de la cadena, obtenemos:

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial c_j}\frac{\partial c_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ji}}$$

Calculamos las derivadas parciales anteriores:

$$\frac{\partial J}{\partial c_j} = \frac{\partial}{\partial c_j}\frac{1}{2}\sum_{k=1}^{n_O}(t_k - c_k)^2 = \frac{\partial}{\partial c_j}\frac{1}{2}(t_j - c_j)^2 = -(t_j - c_j)$$

Para derivar c_j respecto de z_j, utilizamos que $c_j = \theta(z_j)$:

$$\frac{\partial c_j}{\partial z_j} = \frac{\partial}{\partial z_j}\theta(z_j) = \theta'(z_j)$$

Finalmente, para derivar z_j respecto de w_{ji} usamos que $z_j = \sum_{k=0}^{N_H} y_k w_{jk}$:

$$\frac{\partial z_j}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}}\sum_{k=0}^{N_H} y_k w_{jk} = \frac{\partial}{\partial w_{ji}} y_i w_{ji} = y_i$$

Por tanto, la regla de adaptación de pesos para cada peso entre la capa oculta y la de salida es:

$$w_{ji} \leftarrow w_{ji} + \eta(t_j - c_j)\theta'(z_j)y_i$$

Ahora calculamos la regla de adaptación de pesos para cada pesos entre la capa de entrada y la oculta, $w_{ji} \equiv w_{ji}^{(1)}$. Buscamos de nuevo el valor de $\frac{\partial J}{\partial w_{ji}}$. Aplicando la regla de la cadena, obtenemos:

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j}\frac{\partial y_j}{\partial n_j}\frac{\partial n_j}{\partial w_{ji}}$$

Derivamos primero J respecto de y_j:

$$\frac{\partial J}{\partial y_j} = \frac{1}{2}\frac{\partial}{\partial y_j}\sum_{k=1}^{n_O}(t_k - c_k)^2 = \frac{1}{2}\sum_{k=1}^{n_O}\frac{\partial}{\partial y_j}(t_k - c_k)^2$$

Por la regla de la cadena, $\frac{\partial}{\partial y_j} = \frac{\partial}{z_k}\frac{\partial z_k}{y_j}$, para cada $k = 1, \ldots, n_O$. Aplicando esto a la expresión anterior, obtenemos:

$$\frac{\partial J}{\partial y_j} = \frac{1}{2}\sum_{k=1}^{n_O}\frac{\partial}{\partial z_k}(t_k - c_k)^2\frac{\partial z_k}{\partial y_j}$$

La expresión $\frac{\partial}{\partial z_k}(t_k - c_k)^2$ está calculada en el apartado anterior y vale $-(t_k - c_k)\theta'(z_k)$, mientras que $\frac{\partial z_k}{\partial y_j} = \frac{\partial}{\partial y_j}\sum_{i=0}^{N_O} y_i w_{ki}^{(2)} = \frac{\partial}{\partial y_j} y_j w_{kj}^{(2)} = w_{kj}^{(2)}$, es decir, el peso w_{kj} entre las capas oculta y de salida, que ya hemos calculado en el apartado anterior. Por tanto:

$$\frac{\partial J}{\partial y_j} = -\sum_{k=1}^{n_O}(t_k - c_k)\theta'(z_k)w_{kj}^{(2)}$$

Finalmente, calculamos el resto de derivadas parciales de forma análoga al apartado anterior:

$$\frac{\partial y_j}{\partial n_j} = \frac{\partial}{\partial n_j}\theta(n_j) = \theta'(n_j)$$

$$\frac{\partial n_j}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}}\sum_{k=0}^{N_I} r_k w_{jk} = \frac{\partial}{\partial w_{ji}} r_i w_{ji} = r_i$$

Por tanto, hemos obtenido que:

$$\frac{\partial J}{\partial w_{ji}} = -\left[\sum_{k=1}^{n_O}(t_k - c_k)\theta'(z_k)w_{kj}^{(2)}\right]\theta'(n_j)r_i$$

Luego la regla de adaptación de pesos entre la capa de entrada y la oculta es:

$$w_{ji} \leftarrow w_{ji} + \eta\left[\sum_{k=1}^{n_O}(t_k - c_k)\theta'(z_k)w_{kj}^{(2)}\right]\theta'(n_j)r_i$$

Cuestión 2

Tanto "bagging" como validación cruzada cuando se aplican sobre una muestra de datos nos permiten dar una estimación del error de un modelo ajustado a partir de dicha muestra de datos. Discuta cuál de los dos métodos considera que obtendrá una mejor estimación del error. Especificar con precisión las razones.

Solución

Cada estimación proporciona mejores resultados según el criterio que se desee evaluar (sesgo o varianza).

La validación cruzada, en su versión Leave One Out presenta muy poco sesgo, pero sin embargo tiene una gran varianza ya que

Cuestión 3

Considere que dispone de un conjunto de datos linealmente separable. Recuerde que una vez establecido un orden sobre los datos, el algoritmo perceptron encuentra un hiperplano separador iterando sobre los datos y adaptando los pesos de acuerdo al algoritmo

Algorithm 1 Perceptron

1: **Entradas**: $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, $\mathbf{w} = 0$, $k = 0$
2: **repeat**
3: $k \leftarrow (k + 1) \bmod n$
4: **if** $\text{sign}(y_i) \neq \text{sign}(\mathbf{w}^T\mathbf{x}_i)$ **then**
5: $\mathbf{w} \leftarrow \mathbf{w} + y_i\mathbf{x}_i$
6: **end if**
7: **until** todos los puntos bien clasificados

Modificar este pseudo-código para adaptarlo a un algoritmo simple de SVM, considerando que en cada iteración adaptamos los pesos de acuerdo al caso peor clasificado de toda la muestra. Justificar adecuadamente el resultado, mostrando que al final del entrenamiento solo estaremos adaptando los vectores soporte.

Solución

Cuestión 4

Considerar un modelo SVM y los siguientes datos de entrenamiento: Clase-1: $\{(1,1),(2,2),(2,0)\}$, *Clase-2:* $\{(0,0),(1,0),(0,1)\}$

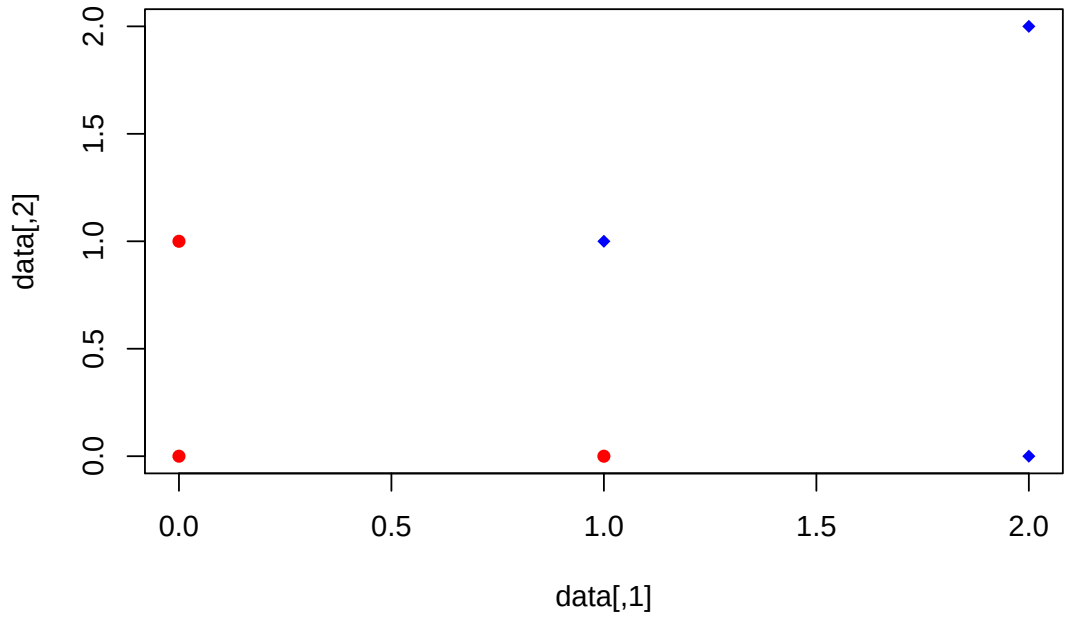
a) *Dibujar los puntos y construir por inspección el vector de pesos para el hiperplano óptimo y el margen óptimo.*

b) *¿Cuáles son los vectores soporte?*

c) *Construir la solución en el espacio dual. Comparar la solución con la del apartado (a)*

Solución

Asignamos la etiqueta 1 a los datos de la clase 1 y la etiqueta -1 a los datos de la clase 2. Los puntos se distribuyen de la siguiente forma:



Buscamos el hiperplano óptimo que separa los datos, y para ello buscamos un vector de pesos (bw_1w_2), con $w = (w_1w_2)$ que minimice la función $\frac{1}{2}w^Tw = \frac{1}{2}(w_1^2 + w_2^2)$ sujeta a las restricciones $y_n(w^Tx_n + b) \geq 1$, donde cada x_n representa a un dato y cada y_n su etiqueta asociada.

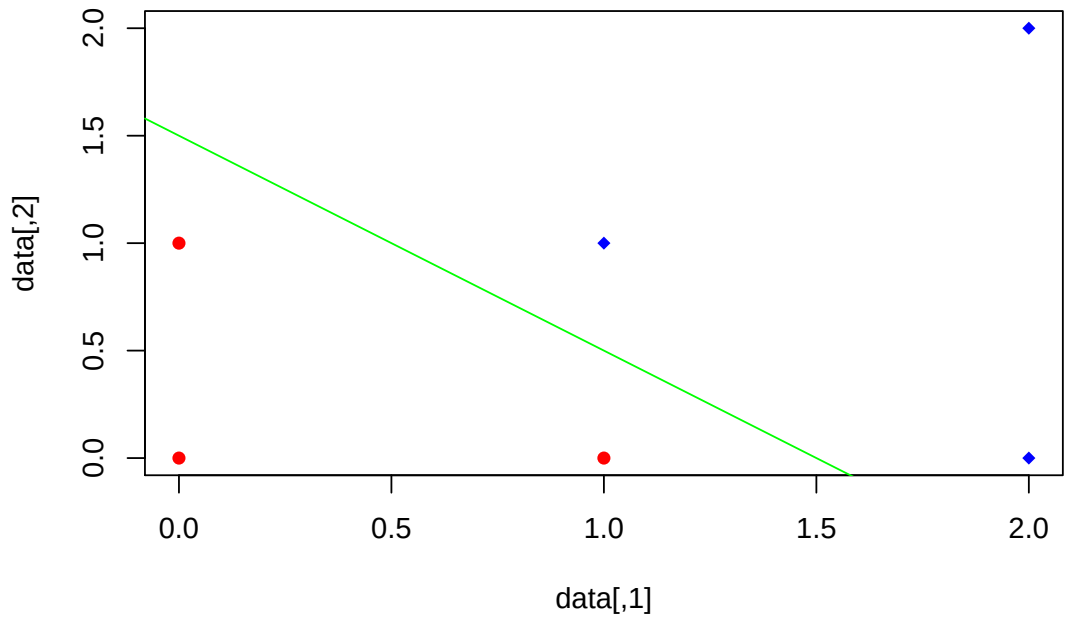
Las restricciones que obtenemos para los puntos dados son las siguientes:

$$
\begin{aligned}
(1,1) &\rightarrow w_1 + w_2 + b \geq 1 && (1)\\
(2,2) &\rightarrow 2w_1 + 2w_2 + b \geq 1 && (2)\\
(2,0) &\rightarrow 2w_1 + b \geq 1 && (3)\\
(0,0) &\rightarrow -b \geq 1 && (4)\\
(1,0) &\rightarrow -w_1 - b \geq 1 && (5)\\
(0,1) &\rightarrow -w_2 - b \geq 1 && (6)
\end{aligned}
$$

Sumando las inecuaciones (1) y (5), obtenemos que $w_2 \geq 2$, y sumando las inecuaciones (1) y (6) $w_1 \geq 2$. Además se verifica que $\frac{1}{2}w^T w$ sujeta a que $w_1, w_2 \geq 2$ alcanza su mínimo cuando $w_1 = w_2 = 2$. Tomando estos valores para w las desigualdades anteriores quedan:

$$
\begin{aligned}
(1,1) &\rightarrow 4 + b \geq 1 && (7)\\
(2,2) &\rightarrow 8 + b \geq 1 && (8)\\
(2,0) &\rightarrow 4 + b \geq 1 && (9)\\
(0,0) &\rightarrow b \leq -1 && (10)\\
(1,0) &\rightarrow b \leq -3 && (11)\\
(0,1) &\rightarrow b \leq -3 && (12)
\end{aligned}
$$

Por tanto, observamos que tomando $b = -3$ se satisfacen todas las restricciones, y además se minimiza $\frac{1}{2}w^T w$ sobre estas restricciones. Por tanto, el vector de pesos asociados al hiperplano es $(bw_1w_2) = (-322)$ y por tanto el hiperplano óptimo que separa los datos viene dado por la ecuación $2u + 2v = 3$. En la siguiente gráfica se muestra la recta obtenida:



Finalmente, el margen óptimo viene dado por la fórmula $M = \frac{1}{\|w\|} = \frac{1}{\sqrt{w_1^2 + w_2^2}} = \frac{1}{\sqrt{8}}$

b)

Los vectores soporte son aquellos para los que la distancia al hiperplano coincide con el margen, o equivalentemente, aquellos para los que se da la igualdad en la restricción asociada. Para ello, para cada dato y con el vector de pesos obtenido $(bw_1w_2) = (-322)$, calculamos $y_n(w^T x_n + b)$ y vemos si es igual a 1.

$$
\begin{aligned}
(1,1) &\rightarrow w_1 + w_2 + b = 1 && (13)\\
(2,2) &\rightarrow 2w_1 + 2w_2 + b = 5 && (14)\\
(2,0) &\rightarrow 2w_1 + b = 1 && (15)\\
(0,0) &\rightarrow -b = 3 && (16)\\
(1,0) &\rightarrow -w_1 - b = 1 && (17)\\
(0,1) &\rightarrow -w_2 - b = 1 && (18)
\end{aligned}
$$

Por tanto, obtenemos que los vectores soporte son $(1,1), (2,0), (1,0)$ y $(0,1)$. En la gráfica anterior se comprueba que son los más cercanos al hiperplano.

c)

Cuestión 5

Una empresa está valorando cambiar su sistema de proceso de datos, para ello dispone de dos opciones, la primera es adquirir dos nuevos sistemas idénticos al actual a 200.000 euros cada uno, y la segunda consiste en adquirir un sistema integrado por 800.000 euros. Las ventas que la empresa estima que tendrá a lo largo de la vida útil de cualquiera de sus equipos son de 5.000.000 de euros en el caso de positivo, a lo que la empresa le asigna una probabilidad de que suceda del 30 %, en caso contrario, las ventas esperadas son de 3.500.000 euros. ¿Qué opción debería de tomar la empresa?

Solución

Tenemos dos alternativas a elegir:

- Alternativa **A**: Comprar los dos sistemas idénticos a 200.000 € cada uno.
- Alternativa **B**: Comprar un sistema integrado a 800.000 €.

Interpretando el enunciado como que las ventas estimadas son por cada dispositivo comprado, para la alternativa B tendríamos dos posibilidades: positivo (+), con probabilidad 0.3 y ganancia de $5 - 08 = 42$ millones, y negativo (-), con probabilidad 0.7 con ganancia de 2,7 millones.

Para la alternativa A, tendríamos 4 opciones, y como que se dé (+) en cada uno de los dispositivos son sucesos independientes, las opciones con sus probabilidades serían:

- $(+,+)$, con probabilidad 0,9, y ganancias de 9,6 millones ($5 + 5$ millones, menos el precio de los sistemas).
- $(+,-)$, con probabilidad 0,21 y ganancias de 8,1 millones.
- $(-,+)$, con probabilidad 0,21 y ganancias de 8,1 millones.
- $(-,-)$, con probabilidad 0.49 y ganancias de 6,6 millones.

El árbol de decisión sería el que se muestra en la figura (AÑADIR):

Finalmente, las ganancias esperadas para cada alternativa serían:

- Alternativa A: $0,09 \times 9,6 + 0,21 \times 8,1 + 0,21 \times 8,1 + 0,49 \times 6,6 = 7,5$ millones.
- Alternativa B: $0,3 \times 4,2 + 0,7 \times 2,7 = 3,15$ millones.

Claramente, con esta interpretación, la alternativa a escoger sería la A.

Interpretamos ahora el enunciado como que las ventas estimadas son siempre 5 millones en caso de éxito, y 3,5 millones en caso de no éxito, y que el éxito se da solo cuando todos los dispositivos tienen éxito (bajo la probabilidad de éxito por dispositivo de 0,3).

En este caso, los casos para la alternativa B son los mismos que en la interpretación anterior, con las mismas probabilidades y ganancias, mientras que los casos para la alternativa A quedan de la siguiente forma:

- *Positivo*, solo si se da positivo en los dos sistemas, es decir, tiene probabilidad $0,3 \times 0,3 = 0,09$ y la ganancia sería de 5 millones $-0,4$ millones $= 4,6$ millones.
- *Negativo*, si falla alguno de los dos sistemas, es decir, se tendría una probabilidad de $1 - 0,09 = 0,91$ y la ganancia sería de $3,5$ millones $-0,4$ millones $= 3,1$ millones.

El árbol de decisión sería el que se muestra en la figura (AÑADIR):

Finalmente, las ganancias esperadas para cada alternativa serían:

- Alternativa A: $0,09 \times 4,6 + 0,91 \times 3,1 = 3,235$ millones.
- Alternativa B: $3,15$ millones, como en la interpretación anterior.

Por tanto, con esta interpretación, elegiríamos también la alternativa A.

Cuestión 6

El método de Boosting representa una forma alternativa en la búsqueda del mejor clasificador respecto del enfoque tradicional implementado por los algoritmos PLA, SVM, NN, etc.

a) *Identifique de forma clara y concisa las novedades del enfoque.*
b) *Diga las razones profundas por las que la técnica funciona produciendo buenos ajustes (no ponga el algoritmo).*
c) *Identifique sus principales debilidades.*
d) *¿Cuál es su capacidad de generalización comparado con SVM?*

Solución

Cuestión 7

¿Cuál es a su criterio lo que permite a clasificadores como Random Forest basados en un conjunto de clasificadores simples aprender de forma más eficiente? ¿Cuáles son las mejoras que introduce frente a los clasificadores simples? ¿Es Random Forest óptimo en algún sentido? Justifique con precisión las contestaciones.

Solución