INTRODUCCION A LOS MODELOS COMPUTACIONALES 29 enero 2014

Alumno/a D.....

Cuestiones.- 1) (2 puntos) Explique línea por línea el significado del siguiente algoritmo

Algorithm Backpropagation;

Start with randomly chosen weights;

while MSE is unsatisfactory

and computational bounds are not exceeded, do

for each input pattern x_p, $1 \leq p \leq P$,

Compute hidden node inputs ($net_{p,j}^{(1)}$);

Compute hidden node outputs ($x_{p,j}^{(1)}$);

Compute inputs to the output nodes ($net_{p,k}^{(2)}$);

Compute the network outputs ($o_{p,k}$);

Modify outer layer weights:

$$\Delta w_{k,j}^{(2,1)} = \eta(d_{p,k} - o_{p,k})\mathcal{S}'(net_{p,k}^{(2)})x_{p,j}^{(1)}$$

2.- Se asignan pesos aleatorios a todas las conexiones de la red neuronal, por lo general en el intervalo [-1,1]

3,4.- Ejecutar mientras el error MSE no aumente, ni se haya alcanzado el criterio de parada asociado al número de iteraciones o épocas. Dado que la función objetivo es MSE se podría tratar de un modelo de regresión no lineal.

5.- Para cada patrón de entrenamiento hacer.

6.- Obtener para los nodos de la capa oculta, la suma ponderada de los pesos, de todas las conexiones de entrada al nodo.

7. Obtener la salida de los nodos de la capa oculta (calculando los valores de una función sigmoide del valor obtenido en 6

8. De igual manera que en 6, se calcula la suma ponderada de los pesos de las entradas a cada nodo de la capa de salida

9.- Al valor calculado, le aplicamos una función sigmoide o softmax y este será el valor de salida de la red, si es un problema de clasificación, en otro caso la salida es lineal

10, 11. Modificar los pesos de la red de capa oculta a capa de salida. Donde η es el coeficiente de aprendizaje, $(d_{p,k}-o_{p,k})$ es el error de salida del patrón p, S´ la derivada de la función sigmoide

2) Diseñe una red de base radial con dos nodos en la capa oculta y función Gaussiana como función de activación/transferencia para el problema XOR. Como se calculan los parámetros de la red neuronal asociados a los centroides y a los radios de las funciones de base?

Solución.- Los centroides se obtienen haciendo un análisis cluster de los patrones de entrenamiento , fijando el número de cluster lo que es lo mismo que fijar el número de funciones RBF.

Para obtener los radios de las Gaussianas se determina la distancia media del centroide al cual queremos calcular el radio a los demás centroides.

3) (1.5 puntos) Analice las reglas de decisión $\begin{array}{l} \text{Si } -\ln P(\mathbf{x}/C_1) + \ln P(\mathbf{x}/C_2) > 0 \text{ Entonces } \mathbf{x} \in C_1 \\ \text{Si } -\ln P(\mathbf{x}/C_1) + \ln P(\mathbf{x}/C_2) < 0 \text{ Entonces } \mathbf{x} \in C_2 \end{array}$,

¿Que denominación tienen? ¿Tienen en cuenta las probabilidades a priori de pertenencia a cada clase? ¿Bajo que hipótesis de las probabilidades $P(x/C_1)$ y $P(X/C_2)$ las reglas presentan clasificadores lineales?

Es una función discriminante lineal, construida con un modelo Bayesiano. No se tienen en cuenta las probabilidades de pertenencia a priori; porque en ese caso la función discriminante sería calcular el – logaritmo de la razón de verosimilitudes, H(X),

$$H(\mathbf{x}) = -L(\mathbf{x}) = -\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2)$$

supere el umbral dado por el cociente de las probabilidades a priori, y de esta forma la regla de decisión ahora es

$$H(\mathbf{x}) = -L(\mathbf{x}) = -\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2) > \ln \frac{P(C_2)}{P(C_1)}$$

x pertenece a la clase C_1, en otro caso a la clase C_2

Si consideramos que la distribución de $P(\mathbf{x} / C_i),$ para i=1,2 es normal con vector de medias $\mathbf{m_i}$ y matriz de varianzas-covarianzas $\mathbf{\Sigma_i}$; entonces si la función discriminante viene dada por: Una ecuación cuadrática si $\Sigma_1 \neq \Sigma_2$, Una ecuación lineal si $\Sigma_1 = \Sigma_2 = \Sigma$, esto es, si las dos matrices de varianza-covarianza son iguales.

Ejercicio 1.-En que se basa el algoritmo de las Maquinas de Vectores Soporte? Dados los datos etiquetados como positivos definidos como vectores traspuestos $(1.5,1.5)^T$, $(1.5,-1.5)^T$, $(-1.5,-1.5)^T$, $(-1.5, 1.5)^T$ y los datos etiquetados como negativos $(1,1)^T$, $(1,-1)^T$, $(-1,-1)^T$ y $(-1,1)^T$ y la función de trasformación del espacio de características de entrada

$$\Phi(x_1, x_2)^T = \begin{cases} (4 - x_2 + |x_1 - x_2|,\ 4 - x_1 + |x_1 - x_2|)^T & \text{si } \sqrt{x_1^2 + x_2^2} > 2 \\ (x_1, x_2)^T & \text{en otro caso} \end{cases}$$

Calcular los vectores soporte, la ecuación del hiperplano de separación y utilizar el algoritmo SVM para clasificar el patrón de coordenadas $(2,5)^T$

Solución

El algoritmo SVM se basa en crear un clasificador biclase lineal, en una dimensión mayor a la dada en el espacio de variables independientes iniciales del problema, Conforme aumentamos la dimensionalidad, la probabilidad de que las clases en este nuevo espacio sean linealmente separable aumenta.

La idea por tanto es minimizar el margen, por una parte, y por la otra minimizar el número de errores de clasificación

Dados $p_1=(1.5,1.5)^T$, $p_2=(1.5,-1.5)^T$, $p_3=(-1.5,-1.5)^T$, $p_4=(-1.5, 1.5)^T$ patrones de la clase C+ y $p_5=(1,1)^T$, $p_6=(1,-1)^T$, $p_7=(-1,-1)^T$ y $p_8=(-1,1)^T$ patrones de la clase C-

Transformamos los patrones según la función de transformación nolineal del enunciado. De esta forma

$$p_1' = \phi(p_1) = \begin{pmatrix} 4 - 1,5 + |1,5 - 1,5| \\ 4 - 1,5 + |1,5 - 1,5| \end{pmatrix} = \begin{pmatrix} 2,5 \\ 2,5 \end{pmatrix};$$

puesto que para p_1 $\sqrt{(1,5)^2 + (1,5)^2} = 2,12 > 2;$

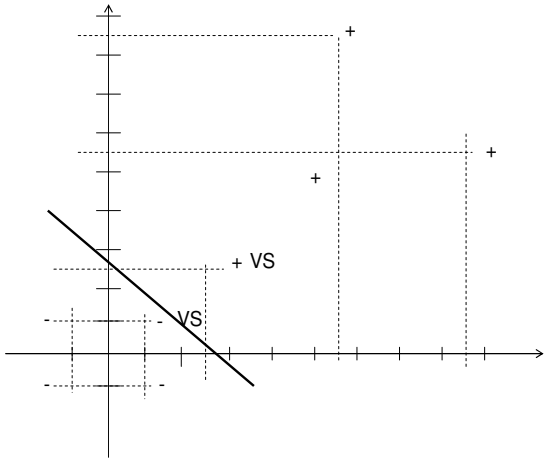
de forma similar $p_2' = \phi(p_2) = (8,5 \quad 5,5)^T$

$p_3' = \phi(p_3) = (5,5 \quad 5,5)^T$; $p_4' = \phi(p_4) = (5,5 \quad 8,5)^T$

Mientras que dado que p_5 es tal que $\sqrt{1^2+1^2}=\sqrt{2}=1,41<2$. Entonces no cambia de valor

$p'_5=(1,1)^T$, y de forma similar $p'_6=(1,-1)^T$, $p'_7=(-1,-1)^T$ y $p'_8=(-1,1)^T$

Si dibujamos los 8 nuevos patrones



Los puntos más cercanos de las dos clases son el $(2,5\ 2,5)^T$ y el $(1,\ 1)^T$

De esta forma $\overline{S}_1=(2,5\ 2,5\ 1)^T$ para C^+; mientras que $\overline{S}_2=(1\ 1\ 1)^T$ para C^-, de esta forma hemos ampliado el número de componentes de los vectores al añadirles un 1 para contemplar el sesgo

Las ecuaciones del dual (antes hay que construir el primal con la función a optimizar y con las restricciones) son ahora

$$\begin{cases}\alpha_1.\overline{S}_1.\overline{S}_1^T+\alpha_2.\overline{S}_2.\overline{S}_1^T=+1\\ \alpha_1.\overline{S}_1.\overline{S}_2^T+\alpha_2.\overline{S}_2.\overline{S}_2^T=-1\end{cases}$$

pero

$\overline{S}_1.\overline{S}_1^T=13,5\ ;\ \overline{S}_2.\overline{S}_1^T=6\ ;\ \overline{S}_1.\overline{S}_2^T=6\ y\ \overline{S}_2.\overline{S}_2^T=3$

Sustituyendo y resolviendo la ecuación tenemos que

$\alpha_1=2$ y $\alpha_2=-4,3$; por lo que el vector de pesos es de la forma

$$w=2\begin{pmatrix}2,5\\2,5\\1\end{pmatrix}-4,3\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}0,7\\0,7\\-2,3\end{pmatrix}$$

La ecuación de la recta es **w.x**+b,

$0,7x_2=-0,7x_1+2,3$

Para clasificar el punto $(2\ 5)^T$ tenemos que como

$\phi(2\ 5)=\begin{pmatrix}4-5+|2-5|\\4-2+|2-5|\end{pmatrix}=\begin{pmatrix}2\\5\end{pmatrix}$, si le añadimos la componente del sesgo, entonces

$$f\left(\begin{pmatrix}2\\5\end{pmatrix}\right)=\sigma\left[2.\begin{pmatrix}2,5\\2,5\\1\end{pmatrix}.\begin{pmatrix}2 & 5 & 1\end{pmatrix}-4,3\begin{pmatrix}1\\1\\1\end{pmatrix}.\begin{pmatrix}2 & 5 & 1\end{pmatrix}\right]=\sigma(2,6)$$

por lo que el patrón pertenece a la clase C+, dado que la salida es positiva

2.- **(2.5 puntos)** ¿Qué hipótesis suponemos a la hora de aplicar un clasificador Naïve-Bayes?. Supongamos la siguiente base de datos de condiciones atmosféricas para jugar, o no, a tenis. Calcular la probabilidad de jugar, y de no jugar, bajo las siguientes condiciones atmosféricas
Outlook = rainy; Temperature = hot; Humidity = high; Windy = false

Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes

Solución.-
Tabla de frecuencias

Outlook	Play		Temp	Play		Hum	Play		Windy	Play		Play	
	Yes	No		Yes	No		Yes	No		Yes	No		
Sunny	1	3	Hot	1	2	High	2	3	False	5	2	Yes	6
Overcast	2	0	Mild	2	1	Normal	4	1	True	1	2	No	4
Rainy	3	1	Cool	3	1								
Total	6	4		6	4		6	4					10

Tabla de probabilidades

Outlook	Play		Temp	Play		Hum	Play		Windy	Play		Play	
	Yes	No		Yes	No		Yes	No		Yes	No		
Sunny	0,16	0,75	Hot	0,16	0,50	High	0,33	0,75	False	0,83	0,5	Yes	0,6
Overcast	0,34	0.00	Mild	0,34	0,25	Normal	0,66	0,25	True	0,16	0,5	No	0,4
Rainy	0,50	0,25	Cool	0,50	0,25								

P(O=Rainy; T=hot; H=high; W=false; Play=YES)= 0,5.0,16.0,33.0,83.0,6= 0,0131472

P(O=Rainy; T=hot; H=high; W=false; Play=NO)=0,25.0,5.0,75.0,5.0,4= 0,01875

$$\mathrm{P}(\text{jugar})=\frac{0,0131472}{0,0131472+0,01875}=0,4121$$

$$P(\text{no jugar})=\frac{0,01875}{0,0131472+0,01875}=0,5879$$

- Un nodo es condicionalmente independiente de sus no-descendentes dados sus padres
- Un nodo es condicionalmente independiente de todos los otros nodos de la red dados sus padres, hijos y los padres de los hijos (también conocido como su capa de Markov)