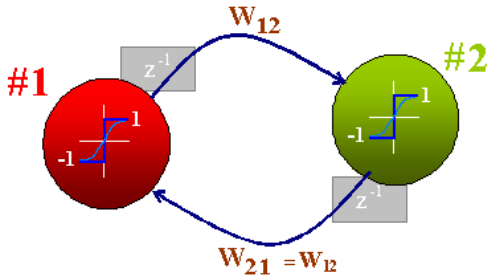
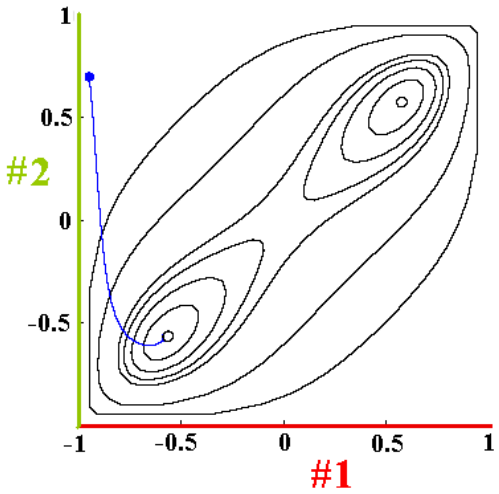
INTRODUCCION A LOS MODELOS COMPUTACIONALES 7 julio 2015

Alumno/a D.....

Cuestiones.- 1) (1.5 puntos) Dadas las siguiente figuras. Analice el tipo de red neuronal que representan. Comente su estructura y forme la matriz de pesos**.**





Solución.- Es una Red de Hopfield con 2-neuronas de estados continuos caracterizada por dos estados estables. Se puede pensar como un proceso de minimización de energía. El sistema dinámico correspondiente evoluciona hacia estados de baja energía. Se define una función de energía (análoga a la función de error del gradiente descendente). Supongamos que una unidad al azar ha sido actualizada: E siempre decrece!

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j + \sum_j s_j \Theta_j$$

En este caso i y j toman valores de 1 a 2, y la matriz de pesos es $\begin{pmatrix} 0 & w_{12} \\ w_{21} & 0 \end{pmatrix}$

Si si es inicialmente –1 y $\sum_j w_{ij} s_j > \Theta_i$ entonces si se convierte en +1

Hay un cambio en $E = -\frac{1}{2}\sum_{i,j}(w_{ij} s_j + w_{ji} s_j) + \Theta_i = -\sum_j w_{ij} s_j + \Theta_i < 0$!!

Si si es inicialmente +1 y $\sum_j w_{ij} s_j < \Theta_i$ entonces si se convierte en -1

Hay un cambio en $E = \frac{1}{2}\sum_{i,j}(w_{ij} s_j + w_{ji} s_j) - \Theta_i = \sum_j w_{ij} s_j - \Theta_i < 0$!!

2) (2 puntos) Analizar el siguiente algoritmo de gradiente descendente para un modelo de red neuronal de unidades producto, F(**x**). Construir el grafo de la red. Calcular para la función $F(\mathbf{x}) = 3\text{-}2x_1 + 2x_2 - x_1^2 x_2 - 4x_2^2$, dado el punto inicial $\mathbf{x_0} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ y α=0,3 el punto al cual nos llevara, y el valor de la función, al cabo de dos ejecuciones del citado algoritmo.

Algoritmo

Elegir la siguiente etapa de forma tal que la función decrezca $F(\mathbf{x}_{k+1}) < F(\mathbf{x}_k)$

Para pequeños cambios en **x** podemos aproximar F(**x**) mediante la expresión

$$F(\mathbf{x}_{k+1}) = F(\mathbf{x}_k + \Delta\mathbf{x}_k) \approx F(\mathbf{x}_k) + \mathbf{g}_k^T \Delta\mathbf{x}_k \quad donde\ \mathbf{g}_k \equiv \nabla F(\mathbf{x}) \,|\, \mathbf{x} = \mathbf{x}_k$$

Si queremos que la función decrezca definimos $\mathbf{g}_k^T \Delta\mathbf{x}_k = \alpha_k \mathbf{g}_k^T \mathbf{p}_k < 0$

Podemos aumentar la velocidad de decrecimiento si definimos $\mathbf{p}_k = -\mathbf{g}_k$, de forma tal que

$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$

Solución.- $F(\mathbf{x}) =$ 3-$2x_1 + 2x_2 - x_1^2 x_2 - 4x_2^2$

F($\mathbf{x_0}$)= 3-2+2-1-4= -2

$$\frac{\partial F}{x_1} = -2 - 2x_1 x_2 \mid \mathbf{x} = \mathbf{x_0} = -4$$

$$\frac{\partial F}{x_2} = 2 - x_1^2 - 8x_2 \mid \mathbf{x} = \mathbf{x_0} = -7 \text{, luego } g_1 = \begin{pmatrix} -4 \\ -7 \end{pmatrix}$$

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 0,3 \begin{pmatrix} -4 \\ -7 \end{pmatrix} = \begin{pmatrix} 2,2 \\ 3,1 \end{pmatrix},$$

F($\mathbf{x_1}$)= 3-2(2,2)+2(3,1)-$(2,2)^2$(3,1)-4$(3,1)^2$= 48,64 con lo que la función se ha minimizado. El segundo paso sería similar

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 qué hipótesis de las probabilidades P($\mathbf{x}$/C_1) y P($\mathbf{x}$/C_2) las reglas presentan clasificadores lineales? ¿Y cuadráticos?

Solución.- 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

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

$\mathbf{x}$ 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.

Ejercicios

1.- (2.5 puntos) ¿En que se basa el algoritmo de las Maquinas de Vectores Soporte? Dados los datos etiquetados como positivos definidos como vectores traspuestos $(2,1.5)^T$, $(2,-1.5)^T$, $(-1.5,-2)^T$, $(-1.5, 2)^T$ 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

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

Calcular los vectores soporte, la ecuación del hiperplano de separación y utilizar el algoritmo SVM para clasificar el patrón de coordenadas $(-2,4)^T$

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 separables aumenta.

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

Dados $p_1=(2,1.5)^T$, $p_2=(2,-1.5)^T$, $p_3=(-1.5,-2)^T$, $p_4=(-1.5, 2)^T$ 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 no lineal del enunciado. De esta forma

$$p_1^{'}=\phi(p_1)=\begin{pmatrix}2-1,5+|2-1,5|\\ 2-2+|2-1,5|\end{pmatrix}=\begin{pmatrix}1\\ 0,5\end{pmatrix};$$

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

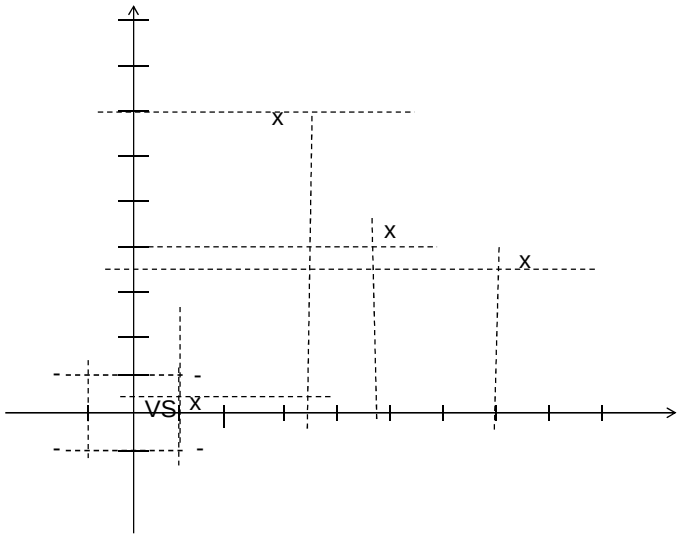
de forma similar $p_2^{'}=\phi(p_2)=\begin{pmatrix}7 & 3,5\end{pmatrix}^T$

$p_3^{'}=\phi(p_3)=\begin{pmatrix}4,5 & 4\end{pmatrix}^T$; $p_4^{'}=\phi(p_4)=\begin{pmatrix}3,5 & 7\end{pmatrix}^T$

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 $(1\ \ 0{,}5)^T$ y el $(1,\ 1)^T$

De esta forma $\overline{S}_1 = (1\ 0{,}5\ 1)^T$ 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

$$\overline{S}_1.\overline{S}_1^T = 2{,}25\ ;\ \overline{S}_2.\overline{S}_1^T = 2{,}5\ ;\ \overline{S}_1.\overline{S}_2^T = 2{,}5\ y\ \overline{S}_2.\overline{S}_2^T = 3$$

Sustituyendo y resolviendo la ecuación tenemos que

α_1= 11 y α_2= -9,5; por lo que el vector de pesos es de la forma

$$w = 11\begin{pmatrix} 1 \\ 0{,}5 \\ 1 \end{pmatrix} - 9{,}5\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1{,}5 \\ -4 \\ 1{,}5 \end{pmatrix}$$

La ecuación de la recta es **w.x**+b,

$1{,}5x_2 = 4x_1 - 1{,}5$ o lo que es igual $x_2 = 2{,}6x_1 - 1$

Para clasificar el punto $(-2\ 4)^T$ tenemos que como

$\phi(-2\ 4) = \begin{pmatrix} 2-4+|-2-4| \\ 2+2+|-2-4| \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \end{pmatrix}$, si le añadimos la componente del sesgo, entonces

$$f\left(\begin{pmatrix}4\\10\end{pmatrix}\right)=\sigma\left[11.\begin{pmatrix}1\\0,5\\1\end{pmatrix}.\begin{pmatrix}4 & 10 & 1\end{pmatrix}-9,5\begin{pmatrix}1\\1\\1\end{pmatrix}.\begin{pmatrix}4 & 10 & 1\end{pmatrix}\right]=\sigma(-32,5)$$

por lo que el patrón pertenece a la clase C-, dado que la salida es negativa

2.- **(2.5 puntos)** ¿Qué hipótesis suponemos a la hora de aplicar un clasificador Naïve-Bayes? Utilice la 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 = cool; Humidity = high; Windy = true**

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=cool; H=high; W=true; Play=YES)= 0,5.0,16.0,33.0,83.0,6=

P(O=Rainy; T=cool; H=high; W=true; Play=NO)=0,25.0,5.0,75.0,5.0,4=

$$P\left(\text{Play=yes/O} = \text{Rainy};\ \text{T} = \text{hot};\ \text{H} = \text{high};\ \text{W} = \text{false}\right) =$$

$$= \frac{P(\text{O} = \text{Rainy};\ \text{T} = \text{hot};\ \text{H} = \text{high};\ \text{W} = \text{false};\ \text{Play} = yes)}{P(\text{O} = \text{Rainy};\ \text{T} = \text{hot};\ \text{H} = \text{high};\ \text{W} = \text{false})} =$$

$$= \frac{P(\text{Play=yes})P(\text{O} = \text{Rainy};\ \text{T} = \text{hot};\ \text{H} = \text{high};\ \text{W} = \text{false / Play=yes})}{P(\text{O} = \text{Rainy};\ \text{T} = \text{hot};\ \text{H} = \text{high};\ \text{W} = \text{false})}$$

$$= \frac{0,6 * (0,5*0,16*0,33*0,83)}{0,4 * 0,3 * 0,5 * 0,7} = \frac{0,0131}{0,042} = 0,311$$

- Un nodo es condicionalmente independiente de ascendientes que no sean sus padres
- Un nodo es condicionalmente independiente de todos los otros nodos de la red a excepción de sus padres, hijos y los padres de los hijos (también conocido como su capa de Markov)