Pauta Examen CC50Q
Teoría de la Información y Redes Neuronales

Prof. Pedro Ortega <peortega@dcc.uchile.cl>
Aux. Francisco Claude <fclaude@dcc.uchile.cl>

27 de noviembre de 2005

Tiempo: 1.5 horas. La nota depende del resultado y del procedimiento empleado para obtenerlo. *Argumente clara y rigurosamente*. Sin apuntes - con calculadora.

Pregunta 1

Considere un sistema de cancelación de ruido de dos micrófonos y un módulo de cancelación, como aquel indicado en la figura. El primer micrófono registra la señal de interés $s(t)$ más un ruido ambiental $n(t)$ ponderado (el factor de ponderación a es desconocido). El segundo micrófono, ubicado en otro lugar, registra sólamente al ruido ambiental $n(t)$. El cancelador de ruido F entrega dos resultados: una salida filtrada (idealmente $\hat{s}(t) \approx s(t)$), y una señal de error $e(t)$.

La función implementada por el cancelador de ruido es

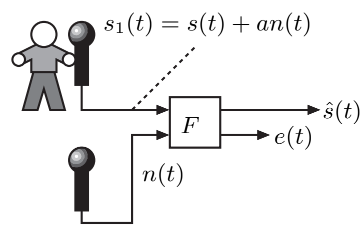
$$\hat{s}(t) = s_1(t) - wn(t)$$

$$e(t) = \hat{s}^2(t)$$

donde w es el único parámetro del módulo F.

Muestre que minimizando la esperanza $E[e(t)]$ se tiende a purificar la señal $s_1(t)$, i.e. cancelar su ruido ambiental. Proponga una regla de aprendizaje *on-line*, de manera que el parámetro w inicializado aleatoriamente converja a uno que produzca una señal purificada $\hat{s}(t)$.

Hint: $E[s(t)n(t)] = 0$ cuando $s(t)$ y $n(t)$ no están correlacionados. Recuerde que un algoritmo *on-line* optimiza una esperanza.



Solución Escribiendo $\hat{s}^2(t)$ tenemos:

$$\begin{aligned}
\hat{s}^2(t) &= (s(t) + (\alpha - w)n(t))^2 \\
&= s^2(t) + (\alpha - w)^2 n^2(t) + 2(\alpha - w)s(t)n(t)
\end{aligned}$$

La esperanza de esta expresión es:

$$\begin{aligned}
E[\hat{s}^2(t)] &= E[s^2(t) + (\alpha - w)^2 n^2(t) + 2(\alpha - w)s(t)n(t)] \\
&= E[s^2(t)] + E[(\alpha - w)^2 n^2(t)] + E[2(\alpha - w)s(t)n(t)] \\
&= E[s^2(t)] + (\alpha - w)^2 E[n^2(t)] + 2(\alpha - w)E[s(t)n(t)] \\
&= E[s^2(t)] + (\alpha - w)^2 E[n^2(t)]
\end{aligned}$$

Ya que $E[s(t)n(t)] = 0$ debido a que $s(t)$ y $n(t)$ son independientes y por lo tanto $E[s(t)n(t)] = E[s(t)]E[n(t)]$ donde $E[n(t)] = 0$ y $E[s(t)] = 0$.

Notamos además que $E[s^2(t)] \geq 0$, $E[n^2(t)] \geq 0$ y que $(\alpha - w)^2 \geq 0$, por lo que el mínimo se puede alcanzar igualando $\alpha = w$.

Para el entrenamiento, se sabe que el entrenamiento $on - line$ mediante descenso por el gradiente minimiza la esperanza del criterio de error, en este caso $e(t)$. Por lo tanto, la regla $on - line$ es la siguiente:

$$\begin{aligned}
w(t+1) &= w(t) - \mu \frac{\partial e}{\partial w} \\
\frac{\partial e}{\partial w} &= 2(-n(t))(s_1(t) - wn(t)) \\
w(t+1) &= w(t) + 2\mu n(t)(s_1(t) - wn(t))
\end{aligned}$$

Con $\mu \in (0, 1)$. *Ojo: la regla de aprendizaje no puede contener los términos a ni* $s(t)$, *ya que estos son desconocidos. Este es un error conceptual grave.*

Pregunta 2

Se desea interpolar una función mediante un polinomio. Se tienen N datos de ejemplo ($\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$) y $m+1$ parámetros ajustables ($w_i, i \in \{0, 1, \ldots, m\}$). El modelo planteado es el siguiente:

$$\begin{aligned} y(\mathbf{w}, x) &= \sum_{i=0}^{m} w_i x^i \\ E(\mathbf{w}) &= \sum_{n=1}^{N} (y(\mathbf{w}, x^{(n)}) - t^{(n)})^2 \end{aligned}$$

donde $t^{(n)}$ es la salida esperada para el ejemplo número n.

1. Derive una regla de aprendizaje mediante descenso por el gradiente.
2. Plantee un sistema de ecuaciones que permita obtener $\mathbf{w}$ de la siguiente forma:

   $$\mathbf{A}\mathbf{w} = \mathbf{b}$$

   donde $\mathbf{A}$ es una matriz y $\mathbf{b}$ un vector.
3. Determine cuantos ejemplos son necesarios para que el entrenamiento se adapte de forma óptima.

Solución

1. Para plantear el método de decenso por el gradienta basta calcular el gradiente del error con respecto a los parámetros del sistema y luego actualizar los pesos de la forma $w_j(t+1) = w_j(t) - \frac{dE}{dw_j}$.

   La derivada del error es la siguiente:

   $$\begin{aligned} \frac{\partial E}{\partial w_j} &= \frac{\partial E}{\partial y} \frac{\partial y}{\partial w_j} \\ \frac{\partial E}{\partial y^{(n)}} &= 2(y^{(n)} - t^{(n)}) \\ \frac{\partial y^{(n)}}{\partial w_j} &= (x^{(n)})^j \end{aligned}$$

   Finalmente se obtiene:

   $$\frac{\partial E}{\partial w_j} = 2 \sum_{n=1}^{N} (y^{(n)} - t^{(n)})(x^{(n)})^j$$

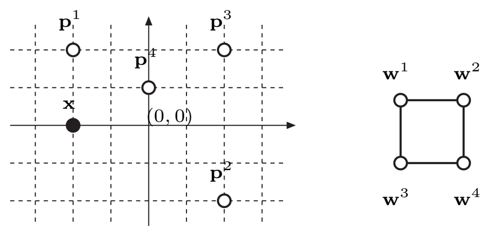
2. El sistema se construye en base a los ejemplos recibidos de la siguiente forma:

   $$\mathbf{A} = \begin{bmatrix} 1 & x^{(0)} & {x^{(0)}}^2 & \ldots & {x^{(0)}}^m \\ 1 & x^{(1)} & {x^{(1)}}^2 & \ldots & {x^{(1)}}^m \\ \vdots & & & & \vdots \\ 1 & x^{(N)} & {x^{(N)}}^2 & \ldots & {x^{(N)}}^m \end{bmatrix} \tag{1}$$

$$
\mathbf{b} \quad = \quad \begin{pmatrix} t^{(0)} \\ t^{(1)} \\ \vdots \\ t^{(N)} \end{pmatrix} \tag{2}
$$

3. Para que el entrenamiento se adapte de forma óptima, es decir $\mathbf{w}$ quede totalmente determinado, es necesario tener $N = m + 1$ ejemplos distintos (independientes). Con esto nos aseguramos que el sistema anterior tenga solución y única.

Pregunta 3



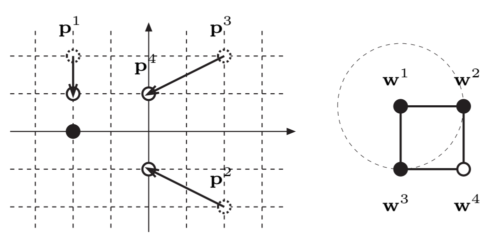
Considere el Mapa de Kohonen de 4 neuronas representado en la figura anterior. El dibujo a la izquierda ilustra la disposición de los prototipos en el espacio de entrada y el dibujo a la derecha la grilla de neuronas que constituye el espacio de salida. En ambos dibujos, las grillas son unitarias.

1. Si se estimula la red con un estímulo $\mathbf{x}$ ilustrado en la figura izquierda, ejecute una iteración del algoritmo de aprendizaje, donde el factor de aprendizaje es $\mu = 1/2$ y el radio de vecindad $N = 1$.
2. Para esta nueva situación, considere el canal binario dado por dos posibles entradas a y b codificadas como $\mathbf{p}^1$ y $\mathbf{p}^2$. Las salidas $\mathbf{w}^1$ y $\mathbf{w}^2$ se interpretan como a y b respectivamente. Si el ruido del canal es tal que tanto $P(\mathbf{x}|a)$ como $P(\mathbf{x}|b)$ son uniformes dentro de cuadrados de lados $r < 4$ centrados en $\mathbf{p}^1$ y $\mathbf{p}^2$ respectivamente, calcule la capacidad del canal.

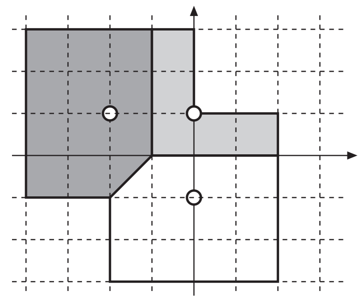
Solución El prototipo ganador es el prototipo $\mathbf{p}^1$, porque es aquel que está más cercano al estímulo. Por lo tanto, él y sus vecinos se acercan al estímulo $\mathbf{x}$ por medio de la regla:

$$\mathbf{p}^i \leftarrow \mathbf{p}^i + \mu(\mathbf{x} - \mathbf{p}^i)$$

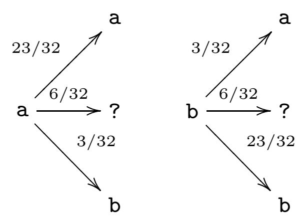
Como $\mu = 1/2$, entonces acortan su distancia hasta la mitad. Los prototipos que se mueven son aquellos que están dentro de un radio de vecindad menor o igual a 1 en el espacio de salida, i.e. los prototipos $\mathbf{p}^1$, $\mathbf{p}^2$ y $\mathbf{p}^3$.



Para la segunda parte se dibujan las regiones de impacto de un estimulo posible tras el envío de un bit. Notemos que la distorsión es tan grande que cualquiera de las cuatro neuronas puede resultar activada. Como la neurona $\mathbf{w}^1$ codifica a a y la neurona $\mathbf{w}^2$ codifica a b, entonces las neuronas $\mathbf{w}^3$ y $\mathbf{w}^4$ codifican a un mensaje desconocido que llamaremos **?**. La situación resultante se ilustra a continuación:



Lo interesante es que el canal resultante es el siguiente:



Por simetría, sabemos que la información mutua se maximiza cuando $P(x = a) = P(x = b) = 1/2$. Así podemos calcular la capacidad resultante:

$$\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= -\frac{23+3}{64}\log_2\frac{23+3}{64} - \frac{23+3}{64}\log_2\frac{23+3}{64} - \frac{6+6}{64}\log_2\frac{6+6}{64} \\
&\quad -\frac{1}{2}\left(-\frac{23}{32}\log_2\frac{23}{32} - \frac{6}{32}\log_2\frac{6}{32} - \frac{3}{32}\log_2\frac{3}{32}\right) \\
&\quad -\frac{1}{2}\left(-\frac{23}{32}\log_2\frac{23}{32} - \frac{6}{32}\log_2\frac{6}{32} - \frac{3}{32}\log_2\frac{3}{32}\right) \\
&= 0{,}5279 + 0{,}5279 + 0{,}4528 - (0{,}3424 + 0{,}4528 + 0{,}3202) \\
&= 0{,}3932 \text{ [bits]}
\end{aligned}$$