Control 2 CC50Q
Teoría de la Información y Redes Neuronales

Pedro Ortega <peortega@dcc.uchile.cl>

16 de agosto de 2005

Tiempo: 2 horas. La nota depende del resultado y del procedimiento empleado para obtenerlo. *Argumente clara y rigurosamente.* Sin apuntes - con calculadora.

Pregunta 1

Parte a (2 pts): Se tiene un conjunto de datos $\mathcal{X} = \{\mathbf{x}^i\}_i \subset \mathbb{R}^2$ ilustrado en la figura 1 (la grilla es unitaria). Se desea ajustar un modelo que represente la densidad de probabilidad de estos puntos, dados por dos rectángulos $R_1(\mathbf{p}_1, \mathbf{q}_1)$ y $R_2(\mathbf{p}_2, \mathbf{q}_2)$, donde $\mathbf{p}_i$ y $\mathbf{q}_i$ son la esquina superior-izquierda y la esquina inferior-derecha del rectángulo i-ésimo respectivamente. La densidad de un punto sería:

$$P(\mathbf{x}|R_1, R_2) = \begin{cases} 1/A_1 & \text{si } \mathbf{x} \in R_1 \wedge \mathbf{x} \notin R_2 \\ 1/A_2 & \text{si } \mathbf{x} \notin R_1 \wedge \mathbf{x} \in R_2 \\ 1/A_1 + 1/A_2 & \text{si } \mathbf{x} \in R_1 \wedge \mathbf{x} \in R_2 \\ 0 & \text{en caso contrario} \end{cases}$$

donde A_i es el área cubierta por el rectángulo i-ésimo. Encuentre la estimación de máxima verosimilitud para los parámetros del modelo.

Parte b (4 pts): Un problema puede modelarse por medio de tres variables binarias A, B, C. Dada la tabla de probabilidades conjuntas, construya la red Bayesiana asociada.

x	ABC	$AB\overline{C}$	$A\overline{B}C$	$A\overline{BC}$	$\overline{A}BC$	$\overline{A}B\overline{C}$	$\overline{AB}C$	$\overline{ABC}$
$P(x)$	0	1/8	0	3/8	1/32	3/32	3/32	9/32

Pregunta 2

Considere que Ud. posee un conjunto de programas

$$P_n = \{p : p \text{ se detiene y es de largo } l(p) \leq n \text{ bits}\}$$

Estos programas escriben strings binarios que pueden ser interpretados como secuencias de acciones-estados $y_1 x_1 y_2 x_2 y_3 x_3 \cdots$. Ud. interpretará un bit y_t como una acción en el instante t y un bit x_t como el estado resultante tras haber ejecutado la acción y_t (hay dos acciones y dos estados).

Parte a: Construya la distribución de probabilidad $P(y_1 x_1 y_2 x_2 \cdots)$ de que la secuencia generada por un programa p escogido al azar de P_n comience con el string $y_1 x_1 y_2 x_2 \cdots y_m x_m$ $(m \in \mathbb{N})$.

Parte b: Use (a) para definir la probabilidad condicional $P(x_t | y_1 x_1 \cdots y_{t-1} x_{t-1} y_t)$. Explique en qué casos existe.

Parte c: Suponga que posee una función de utilidad U definida sobre los x_t, tal que $U(x_t) = x_t$ (es decir, los estados $x_t = 1$ son recompensas y los $x_t = 0$ no aportan). Indique la mejor decisión y_t en el instante t, dado que ha observado una historia $y_1 x_1 \cdots y_{t-1} x_{t-1}$, basada en la probabilidad $P(\cdot)$.

Parte d: ¿Qué relación tiene la *Navaja de Occam* con los conjuntos P_n?

Pregunta 3

Parte a (3 pts): Considere el problema multiclase de $\mathbb{R}^2$ en $\{a, b, c, d, e, f, g\}$ ilustrado en la figura 2. Construya una red neuronal que implemente esta clasificación.

Parte b (3 pts): Considere una neurona con función de activación *sigmoide logística* y criterio de error *entropía cruzada*, dado por

$$E = -\sum_{i=1}^{N} \left(d^i \ln y^i + (1 - d^i) \ln(1 - y^i) \right)$$

donde $y^i \equiv y(\vec{x}, \vec{w})$. Derive una regla de aprendizaje basada en el *descenso por el gradiente*. Entropía cruzada se utiliza como criterio de error para una *salida binaria*, i.e. cuando la salida y de la neurona modela la probabilidad condicional $P(d = 1 | \vec{x})$. ¿Qué ocurre si la red entrega un resultado *contrario al deseado* para una entrada $\vec{x}$?

Solución Pregunta 1

Parte 1: A pesar de que la figura parece sugerir el encierro de cada nube de puntos por un rectángulo, el problema no tiene solución definida. Por ejemplo, consideremos la disposición de los rectángulos de la siguiente forma:

El rectángulo pequeño R_2 puede hacerse tan pequeño uno quiere. Por lo tanto, la verosimilitud $P(\{\mathbf{x}\}, R_1, R_2) = 1/A_1 + 1/A_2 = \text{cte.} + 1/A_2$ para ese punto crece indefinidamente. Por lo tanto, el problema no tiene solución, pues no existe un máximo para la función verosimilitud.

Parte 2: En general, no hay un red Bayesiana única para un set de probabilidades conjuntas dadas, pero existen algunas que son más compactas que otras (tienen menos dependencias). Por lo tanto, para no complicarse, aplicaremos la regla de la cadena:

$$P(a, b, c) = P(a) P(b|a) P(c|a, b)$$

Lo cual sugiere la red Bayesiana siguiente:

Calculemos entonces estos tablas:

x	A	$\overline{A}$
$P(x)$	16/32	16/32

x	$B\|A$	$B\|\overline{A}$	$\overline{B}\|A$	$\overline{B}\|\overline{A}$
$P(x)$	4/32	4/32	12/32	12/32

x	$C\|AB$	$C\|A\overline{B}$	$C\|\overline{A}B$	$C\|\overline{AB}$	$\overline{C}\|AB$	$\overline{C}\|A\overline{B}$	$\overline{C}\|\overline{A}B$	$\overline{C}\|\overline{BC}$
$P(x)$	0	0	1/32	3/32	4/32	12/32	3/32	9/32

Solución Pregunta 2

Parte a: Esta respuesta es muy sencilla:

$$P(y_1x_1y_1x_1\ldots y_mx_m) = \begin{cases} \frac{|G_n|}{|P_n|} & \text{si } P_n \neq \emptyset \\ \text{indef.} & \text{si } P_n = \emptyset \end{cases}$$

con

$$G_n \quad := \quad \{p : p \in P_n, p \text{ genera } y_1x_1y_1x_1\ldots y_mx_m\ldots \in \{0,1\}^*\},$$

obviamente definida sobre strings que terminan con un y o un x.

Parte b: Aquí aplicamos la definición de probabilidad condicional,

$$P(x_t|y_1x_1y_1x_1\ldots y_t) = \frac{P(y_1x_1y_1x_1\ldots y_tx_t)}{P(y_1x_1y_1x_1\ldots y_t)}$$

donde ambos términos deben estar definidos y el denominador no puede valer cero.

Parte c: La respuesta es,

$$y_t^* = \arg\max_{y_t} \sum_{x_t} \left(U(x_t) + V(y_tx_t)\right) P(x_t|y_1x_1\ldots y_t)$$

donde el *valor* $V(\cdot)$ se define como

$$V(y_tx_t\ldots y_mx_m) = 0, \quad \text{si } y_tx_t\ldots y_mx_m\ldots \notin P_n,$$

$$V(y_tx_t\ldots y_{m-1}x_{m-1}) = \max_{y_m} \sum_{x_m} \left(U(x_m) + V(y_tx_t\ldots y_mx_m)\right) P(x_m|y_1x_1\ldots y_m).$$

Ésta es quizás la pregunta más difícil de control.

Parte d: Los programas en P_n pueden verse como modelos/descripciones de secuencias más largas, i.e. las secuencias generadas por los programas en P_n. Como las secuencias resultantes pueden ser de largo arbitrario, pero finito, entonces el conjunto P_n es una descripción comprimida o sencilla (requiere menos bits) de estas secuencias.

La probabilidad $P(\cdot)$ será mayor que cero sólo para aquellas secuencias que tienen una ley generadora sencilla/corta. Más aún, hay secuencias que son el prefijo de secuencias de muchos programas en P_n, por lo que puede decirse que éstas son 'más fáciles de generar' o existe una mayor probabilidad de encontrar al azar una ley que las genere.

Solución Pregunta 3

Parte a: Si codificamos las regiones como

entonces bastan 3 neuronas con activación escalón, que delimitan las regiones siguientes:

Observando que la frontera para la primera neurona está dada por una recta $x_2 = x_1$, es fácil ver que

$$w_{10} = 0, w_{11} = 1, w_{12} = 1.$$

Para la segunda neurona, observemos que $x_2 = 2$ en la frontera de decisión, por lo tanto,

$$w_{20} = -2, w_{21} = 0, w_{22} = 1.$$

La tercera neurona es equivalente a la segunda pero con los ejes intercambiados,

$$w_{30} = -2, w_{31} = 1, w_{32} = 0.$$

La red resultante de una capa de pesos tiene la forma,

Parte b: La regla de aprendizaje basada en el descenso por el gradiente corresponde a:

$$w \leftarrow w - \mu \frac{\partial E}{\partial \vec{w}}, \quad \text{o equivalentemente a} \quad w_j \leftarrow w_j - \mu \frac{\partial E}{\partial w_j}$$

donde μ es el factor de aprendizaje. Falta calcular cuánto vale $\partial E / \partial w_j$. Dado que una neurona calcula la función

$$\mathrm{sl}(a) := \frac{1}{1 + e^{-a}}, \quad \text{con} \quad a = \sum_j w_j x_j,$$

entonces,

$$\begin{aligned} \frac{\partial E}{\partial w_j} &= -\sum_{i=1}^{N} \left(\frac{d^i}{y^i} \frac{\partial y^i}{\partial w_j} - \frac{(1-d^i)}{(1-y^i)} \frac{\partial y^i}{\partial w_j} \right) = \sum_{i=1}^{N} \frac{y^i - d^i}{y^i(1-y^i)} \frac{\partial y^i}{\partial w_j} \\ &= \sum_{i=1}^{N} \frac{y^i - d^i}{y^i(1-y^i)} \cdot y^i(1-y^i) \cdot \frac{a}{\partial w_j} = \sum_{i=1}^{N} (y^i - d^i) x_j. \end{aligned}$$

Por lo tanto, la regla de aprendizaje es,

$$w_j \leftarrow w_j - \mu \sum_{i=1}^{N} (y^i - d^i) x_j$$

Si la red predice un resultado contrario al deseado, i.e. si $y^i = 0, d^i = 1$, o viceversa, se tiene que el término de la suma que constituye a la entropía cruzada vale

$$e^i = 1 \cdot \ln 0 + 0 \cdot \ln 1 \nearrow \infty$$