Control 1 CC50Q
Teoría de la Información y Redes Neuronales

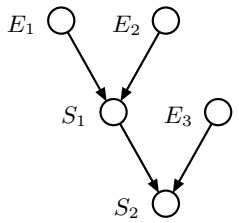
Prof.: Pedro Ortega <peortega@dcc.uchile.cl>
Aux.: Francisco Claude <fclaude@dcc.uchile.cl>

7 de septiembre 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

Una clínica ha modelado la dependencia entre tres enfermedades E_1, E_2 y E_3 y sus posibles síntomas S_1 y S_2 por medio de la red Bayesiana ilustrada a continuación:



Utilizando registros históricos, ha determinado las siguientes probabilidades:

$$P(E_1 = 1) = 0{,}001 \quad P(E_2 = 1) = 0{,}0001$$
$$P(E_3 = 1) = 0{,}0003 \quad P(S_1 = 1) = 0{,}0007$$

$$P(S_1 = 1|E_1 = 0, E_2 = 0) = 0{,}0001 \quad P(S_2 = 1|S_1 = 0, E_3 = 0) = 0{,}0001$$
$$P(S_1 = 1|E_1 = 0, E_2 = 1) = 0{,}2 \quad P(S_2 = 1|S_1 = 0, E_3 = 1) = 0{,}5$$
$$P(S_1 = 1|E_1 = 1, E_2 = 0) = 0{,}6 \quad P(S_2 = 1|S_1 = 1, E_3 = 0) = 0{,}3$$
$$P(S_1 = 1|E_1 = 1, E_2 = 1) = 0{,}7 \quad P(S_2 = 1|S_1 = 1, E_3 = 1) = 0{,}6$$

Un nuevo paciente llega y se observa que presenta ambos síntomas. Se descartó la posibilidad de existir la enfermedad E_3. Si se quiere determinar si el paciente presenta la enfermedad E_1 o E_2, indique la respuesta Bayesiana (la distribución a posteriori) y compárela contra la respuesta de máxima verosimilitud.

Solución

La solución de esta pregunta salía en forma directa si es que uno sabía cómo resolverla. El hecho de tener conocimiento de la variable S_1 restringía al problema a la

subred formada por E_1, E_2 y S_1. Derivemos este resultado de todas formas. Usemos la notación X cuando se trata de variables aleatorias libres, y $\dot{X}$ cuando se ha fijado un valor específico para la variable aleatoria (una variable "enlazada"). Entonces, la pregunta se formula como:

$$P(E_1, E_2|\dot{S}_1, \dot{S}_2, \dot{E}_3) = \frac{P(\dot{S}_1, \dot{S}_2, \dot{E}_3|E_1, E_2)P(E_1, E_2)}{P(\dot{S}_1, \dot{S}_2, \dot{E}_3)}$$

La red Bayesiana nos indica las dependencias entre las variables. Así,

$$P(E_1, E_2, E_3, S_1, S_2) = P(E_1)P(E_2)P(E_3)P(S_1|E_1, E_2)P(S_2|S_1, E_3)$$

Desarrollando la pregunta Bayesiana, se obtiene,

$$\begin{aligned}
P(E_1, E_2|\dot{S}_1, \dot{S}_2, \dot{E}_3) &=^1 \frac{P(\dot{S}_1, \dot{S}_2, \dot{E}_3|E_1, E_2)P(E_1, E_2)}{P(\dot{S}_1, \dot{S}_2, \dot{E}_3)} \\
&=^2 \frac{P(\dot{S}_1|E_1, E_2)P(\dot{E}_3|E_1, E_2, \dot{S}_1)P(\dot{S}_2|E_1, E_2, \dot{S}_1, \dot{E}_3)P(E_1)P(E_2)}{P(\dot{S}_1, \dot{S}_2, \dot{E}_3)} \\
&=^3 \frac{P(\dot{S}_1|E_1, E_2)P(\dot{E}_3)P(\dot{S}_2|\dot{S}_1, \dot{E}_3)P(E_1)P(E_2)}{P(\dot{S}_1, \dot{S}_2, \dot{E}_3)} \\
&=^4 \frac{P(\dot{S}_1|E_1, E_2)P(E_1)P(E_2)}{N}
\end{aligned}$$

El paso (1) se obtiene aplicando la regla de Bayes. El paso (2) aplicando la regla del producto sobre el primer término del numerador. El tercer paso se obtiene reconociendo que el numerador corresponde a $P(E_1, E_2, \dot{E}_3, \dot{S}_1, \dot{S}_2)$, por lo cual se pueden aplicar las simplificaciones detalladas por la red Bayesiana. En el último paso se agruparon los términos constantes, que no varían entre las cuatro posibles hipótesis.

Entonces,

$$\begin{aligned}
E_1 = 0, E_2 = 0 &: 0{,}0001 \times 0{,}999 \times 0{,}9999 \times N = 0{,}0000999 \\
E_1 = 0, E_2 = 1 &: 0{,}2 \times 0{,}999 \times 0{,}0001 \times N = 0{,}00002 \\
E_1 = 1, E_2 = 0 &: 0{,}6 \times 0{,}001 \times 0{,}9999 \times N = 0{,}0005999 \\
E_1 = 1, E_2 = 1 &: 0{,}7 \times 0{,}001 \times 0{,}0001 \times N = 7 \times 10^{-}8
\end{aligned}$$

Luego, normalizando estos valores, obtenemos,

$$\begin{aligned}
P(E_1 = 0, E_2 = 0|S_1 = 1, S_2 = 1, E_3 = 0) &\approx 0{,}1388 \quad (13{,}88\,\%) \\
P(E_1 = 0, E_2 = 1|S_1 = 1, S_2 = 1, E_3 = 0) &\approx 0{,}0278 \quad (2{,}78\,\%) \\
P(E_1 = 1, E_2 = 0|S_1 = 1, S_2 = 1, E_3 = 0) &\approx 0{,}8333 \quad (83{,}33\,\%) \\
P(E_1 = 1, E_2 = 1|S_1 = 1, S_2 = 1, E_3 = 0) &\approx 0{,}0001 \quad (0{,}001\,\%)
\end{aligned}$$

Ahora, la verosimilitud está dada por el término $P(S_1 = 1, S_2 = 1, E_3 = 0|E_1, E_2)$ que vimos que equivalía a comparar simplemente $P(S_1 = 1|E_1, E_2)$. Este valor se maximiza para la hipótesis $E_1 = 1, E_2 = 1$, que es precisamente aquella menos probable según el método Bayesiano.

Pregunta 2

Considere un robot que explora su medio ambiente tratando de ganar recursos (medidos en puntaje). El robot percibe su medio ambiente por medio de sensores, siendo capaz de discriminar un total de $10'000$ estados distintos, es decir, $\mathcal{X} = \{x_1, \ldots, x_{10'000}\}$. El robot posee una función de utilidad $U : \mathcal{X} \to \mathbb{R}^+$ y un conjunto de 10 acciones $\mathcal{A} = \{a_1, \ldots, a_{10}\}$. La dinámica del robot es la siguiente: en cada iteración, el robot, estando en un estado x, selecciona una acción a y llega a un estado y con una probabilidad $P(y|a, x)$ *desconocida* y obtiene $U(y)$ puntos.

Como el robot no conoce la probabilidad real de transición de estado $P(y|a, x)$, su cerebro posee programas que modelan estas probabilidades. Estos 500 programas $p_1, \ldots, p_{500}$ son tales que $p_i(a, x, y) = P_i(y|a, x)$, donde $P_i(y|a, x)$ es la estimación del programa i-ésimo de la probabilidad $P(y|a, x)$. Por ejemplo, si el programa 2 es ejecutado con entrada $(10, 169, 13)$ y entrega 0,845, entonces esto se interpreta como 'estando en el estado 169 y ejecutando la acción 10 se llega al estado 13 con probabilidad 0.845 según el programa número 2'.

Con respecto a lo anterior, responda las siguientes preguntas:

Parte a (3 puntos): Identifique el conjunto de hipótesis $\mathcal{H}$ y el conjunto de datos $\mathcal{D}$ que debe utilizar el robot para aplicar inferencia Bayesiana. Formule el problema de inferencia Bayesiana especificando: la distribución a priori, la función verosimilitud, y encuentre la probabilidad a posteriori sobre el espacio de hipótesis.

Parte b (2 puntos): Suponga que el robot se halla en el estado 1 y debe realizar dos iteraciones. Éste debe decidir qué *primera acción* tomar en base a su estado de conocimiento, i.e. su distribución a priori sobre el estado de hipótesis. Encuentre una fórmula explícita para la mejor primera acción a^*.

Parte c (1 punto): ¿Cuál es el máximo número de bits que el robot puede reducir de su incertidumbre sobre el espacio de hipótesis tras ejecutar una acción? ¿En qué situación se daría este caso?

Solución

Parte a: El espacio de hipótesis corresponde a los modelos alternativos de los cuales debe hacer uso el robot, que en este caso, equivale al espacio de programas que dispone. El espacio de datos posible es una observación del robot, i.e. una tripleta "acción-estado inicial-estado final" (a, x, y). Como inicialmente no sabe cual es el programa correcto, aplicamos el principio de indiferencia. Por lo tanto, $P(p) = 1/500$. La función verosimilitud para este caso está dado entonces por $P(d|h) = P(a, x, y|p)$. Por lo tanto, la distribución a posteriori de una hipótesis (programa) es

$$\begin{aligned} P(p|a,x,y) &=^1 \frac{P(a,x,y|p)P(p)}{P(a,x,y)} =^2 \frac{P(a,x|p)P(y|a,x,p)P(p)}{\sum_{p'} P(a,x|p')P(y|a,x,p')P(p')} \\ &=^3 \frac{P(a,x)p(a,x,y)P(p)}{P(a,x)\sum_{p'} p'(a,x,y)P(p')} =^4 \frac{p(a,x,y)P(p)}{\sum_{p'} p'(a,x,y)P(p')} \end{aligned}$$

En el paso (1) aplicamos la regla de Bayes. En el paso (2) aplicamos la regla de la cadena sobre el término $P(a, x, y|p)$ obteniendo $P(a, x|p)P(y|a, x, p)$, y reescribimos

el denominador como una suma de las distintas hipótesis (término normalizador). En el paso (3) viene el gran truco: sabemos que $P(y|a,x,p)$ es la estimación de la probabilidad de que a partir del estado inicial x y aplicando una acción a, se llega al estado final y, según el programa p. Por lo tanto, esta probabilidad corresponde simplemente a $p(a,x,y)$. Además, notemos que la elección de (a,x) no depende del programa p (el programa no dictamina una probabilidad de ocurrencia de los "estados-acciones" iniciales). Aplicamos entonces este reemplazo tanto en el numerador como en el denominador. Para el paso (4), factorizamos y simplificamos $P(a,x)$.

Parte b: Desarrollando el árbol de alternativas en dos pasos y maximizando la utilidad, se obtiene

$$a^* = \arg\max_{a\in\mathcal{A}} \sum_{y\in\mathcal{X}} \left[U(y) + \max_{a\in\mathcal{A}} \sum_{z\in\mathcal{X}} U(z)P(z|y,a) \right] P(y|x,a)$$

pues el robot, al escoger una acción $a \in \mathcal{A}$, no sólo ganaría el puntaje asociado al segundo estado y, sino que también la utilidad esperada de la mejor acción a partir de este segundo estado y (es decir, la utilidad esperada de los posibles terceros estados $z \in \mathcal{X}$).

Parte c: Supongamos el caso en el cual el robot, a partir del estado x y tomando la acción a, llega al estado y. Ahora, él consulta la probabilidad de esta ocurrencia a todos sus 500 programas. Sin embargo, 499 de estos programas le contestan con $P_i(y|x,a) = 0$. En el fondo, estos 499 programas le están diciendo que esa posibilidad era *imposible*, lo cual contradice a su observación. Por lo tanto, el robot reconocería al programa restante como el único que contempló esa posibilidad, y concluye que es el único correcto. Por lo tanto, ganó $\log_2 500$ bits a partir de esta observación.

Pregunta 3

Parte a (3 puntos): Ud. quiere ordenar un conjunto de n palabras lexicográficamente realizando comparaciones (binarias). Demuestre que el mínimo número de comparaciones necesarias es $O(n \log n)$.

Hint: (Aproximación de *Stirling*):

$$n! \approx n^n e^{-n} \sqrt{2\pi n}$$

Parte b (3 puntos):

1. ¿Cuánta información gana Ud. al ver el resultado de una tirada simultánea de dos dados?
2. Estime la reducción en su incertidumbre sobre el tiempo del próximo día al ver el pronóstico.
3. Estime la incertidumbre sobre un fenómeno que produce puntos (x,y) que Ud. quiere representar por un modelo

$$y = ax + b, \qquad a,b \in [-1,1].$$

Solución

Parte a: Tenemos n cadenas, las cuales pueden ordenarse de $n!$ formas, de todas estas opciones, solo una corresponderá al conjunto de cadenas ordenadas lexicográficamente. Se puede ver que para determinar $n!$ opciones necesito $log_2(n!)$ bits. Como las comparaciones son binarias, es decir, solo pueden entregar dos resultados, con una comparación puedo ganar en el mejor caso 1 bit de información. Asumiendo que cada comparación efectivamente me entregará el máximo de información, necesitariamos $log_2(n!)$ comparaciones para adquirir todos los bits necesarios que nos permiten determinar el orden lexicográfico de las cadenas.

Usando la aproximación de Stirling tenemos que :

$$\begin{aligned} log_2(n!) &\approx log_2(n^n e^{-n}\sqrt{2\pi n}) \\ &\approx log_2(n^n) + log_2 e^{-n}) + log_2(\sqrt{2\pi n}) \\ &\approx nlog_2(n) + O(n) \\ &= O(nlog_2(n)) \end{aligned}$$

Parte b:

1. El arreglo de probabilidad asociado al problema es:

$$\begin{aligned} A_X &= \{\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\}, \\ &\quad \{2,2\},\{2,3\},\{2,4\},\{2,5\},\{2,6\},\{3,3\},\{3,4\}, \\ &\quad \{3,5\},\{3,6\},\{4,4\},\{4,5\},\{4,6\},\{5,5\},\{5,6\}, \\ &\quad \{6,6\}\} \\ P_X &= \{\frac{1}{36},\frac{2}{36},\frac{2}{36},\frac{2}{36},\frac{2}{36},\frac{2}{36}, \\ &\quad \frac{1}{36},\frac{2}{36},\frac{2}{36},\frac{2}{36},\frac{2}{36},\frac{1}{36},\frac{2}{36}, \\ &\quad \frac{2}{36},\frac{2}{36},\frac{1}{36},\frac{2}{36},\frac{2}{36},\frac{1}{36},\frac{2}{36}, \\ &\quad \frac{1}{36} \end{aligned}$$

   De esto vemos que podemos separar en dos casos:

   - Si $x = \{i,j\}, i = j$ entonces ganamos $log_2(36)$
   - Si $x = \{i,j\}, i \neq j$ entonces ganamos $log_2(18)$

2. Asumiendo que el tiempo puede tomar estados discretos $A_X = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ y que cada estado es equiprobable. Podemos platear el problema asumiendo que el pronóstico del tiempo acierta con probabilidad p y que con probabilidad $p-1$ es cualquiera de los estados restantes con la misma probabilidad.

   Antes de ver el pronóstico del tiempo, tenemos una incertidumbre de $H(X) = log_2(n)$ y luego de ver el reporte, tenemos que las probabilidades son ajustadas de forma que si se pronostica que estado i, $P(i) = p$ y para $j \neq i$ $P(j) = \frac{1-p}{n-1}$. Por

lo que mi desconocimiento es en este caso $H(X') = plog_2(\frac{1}{p}) + (n-1)log_2(\frac{n-1}{1-p})$. Para saber cuanto se ha reducido mi incertidumbre, basta calcular:

$$H(X) - H(X')$$

3. Para el modelo, tenemos que a y b son equiprobables en el espacio, debido a que no sabemos nada de ellos. De esto tenemos $f(a) = f(b) = \frac{1}{2}$. Utilizando la fórmula para entropía en el caso continuo y asumiendo independencia entre las variables, tenemos:

$$H(a,b) \quad = \quad H(a) + H(b)$$

Pero además notamos que $H(a) = H(b)$, por lo que $H(a,b) = 2H(a) = 2H(b)$, calculando $H(a)$:

$$\begin{aligned} H(a) &= \int_{-1}^{1} \frac{1}{2} log_2(2) da \\ &= \frac{1}{2} \int_{-1} 1 da \\ &= \frac{1}{2}(1 - -1) \\ &= \frac{1}{2} 2 \\ &= 1 \end{aligned}$$

Y finalmente tenemos $H(a,b) = 2$ bits.