Decisiones Óptimas a un Horizonte Finito: EXPECTIMAX
CC50Q - Teoría de la Información y Redes Neuronales

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

22 de agosto de 2005

Expectimax

¿Cómo tomar decisiones óptimas a un horizonte finito?

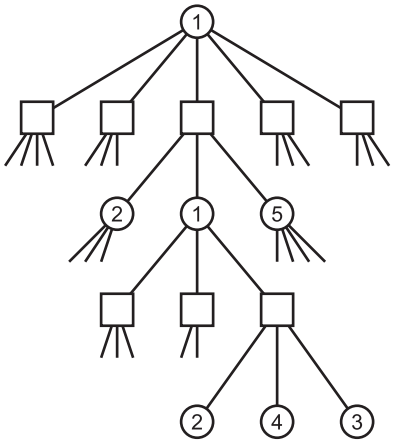
Sean:

- *Espacio de estados*: $\mathcal{X}$ es un conjunto (finito numerable) de estados. Para simplificar la notación, supondremos que un estado $x \in \mathcal{X}$ contiene toda la información necesaria para tomar decisiones futuras.
- *Espacio de acciones*: $\mathcal{A}$ es un conjunto (finito numerable) de acciones.
- *Probabilidades de transición*: Se dispone de las probabilidades $P(x_f|a, x_i)$ de transición. $P(x_f|a, x_i)$ es la probabilidad de llegar al estado x_f partiendo del estado x_i y tomando la decisión/acción a.
- *Función de utilidad*: La función de utilidad U se define en base a una función de utilidad parcial $u : \mathcal{X} \to \mathbb{R}^+$ que entrega un valor real positivo $u(x) \in \mathbb{R}$ para cada estado $x \in \mathcal{X}$. La función de utilidad total U está dada por la suma

$$U = \sum_{i=0}^{h} u(x_i)$$

de una cadena $x_0 x_1 \dots x_h$ de estados.

Gráficamente, uno puede imaginarse la situación anterior como un árbol de dos tipos de nodos: *nodos-estado* (○) y los *nodos-probabilísticos* □.



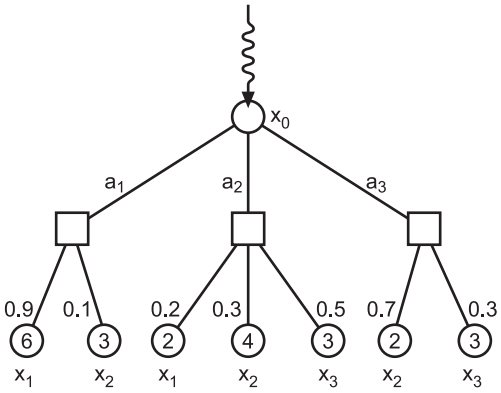
En el dibujo anterior se añadieron utilidades a los nodos-estados.

Las decisiones se toman a partir de un nodo-estado. Cada decisión lleva a un nodo-probabilístico diferente, a partir del cual se determina en forma probabilística el siguiente nodo-estado. Intuitivamente, los nodos-probabilísticos pueden verse como la acción de un agente externo.

La mejor acción $a^* \in \mathcal{A}$ que puede tomar a partir de un estado $x \in \mathcal{X}$ es aquella que *maximiza la esperanza de la utilidad futura*. Esta maximización, si bien es sencilla, requiere de ciertas precauciones. Analicémos este problema informalmente, primero el caso base, y luego el caso recursivo.

Caso Base

El caso base corresponde al caso en que falta tomar la última decisión. La figura a continuación ilustra un caso base de ejemplo.



Para este caso, la mejor decisión a^* corresponde a:

$$a^* = \arg\max_a \sum_{x'} u(x')P(x'|a,x)$$

donde $x \in \mathcal{X}$ es el estado actual y la suma corre sobre todos los estados futuros x' alcanzables tras ejecutar la acción hipotética $a \in \mathcal{A}$.

En el ejemplo de la figura, tendríamos:

$$E[U|a_1,x] = 0{,}9 \cdot 6 + 0{,}1 \cdot 3 = 5{,}4 + 0{,}3 = 5{,}7$$

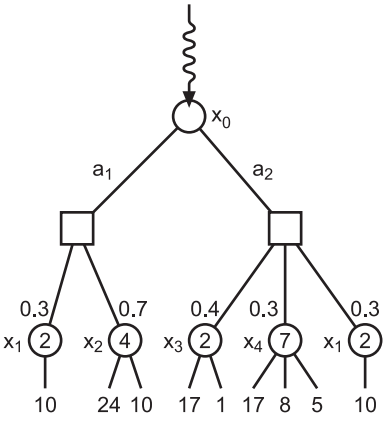
$$E[U|a_2,x] = 0{,}2{\cdot}2{+}0{,}3{\cdot}4{+}0{,}5{\cdot}3 = 0{,}4{+}1{,}2{+}1{,}5 = 3{,}1$$

$$E[U|a_3,x] = 0{,}7 \cdot 2 + 0{,}3 \cdot 3 = 1{,}4 + 0{,}9 = 2{,}3$$

y por lo tanto, la acción que maximiza la esperanza de la utilidad es $a^* = a_1$.

Caso Recursivo

La figura a continuación ilustra un caso intermedio, en el cual no se está tomando la última decisión.



Para este caso existen 2 alternativas: acción a_1 y acción a_2. Los números al final corresponden a las sumas de las *utilidades futuras* si toma esos caminos. Por ejemplo, si tomará la acción a_1 y llegara al estado x_2, entonces a partir de este tendría dos nuevas alternativas: el camino izquierdo y derecho, que reportarían 24 y 10 puntos de utilidad futuras respectivamente. Ya veremos cómo se calculan estos últimos.

Nótese que para que el actor maximice su utilidad futura, siempre debe escoger la mejor alternativa. Es decir, intentará de optar por la mejor alternativa cada vez que tenga la oportunidad de tomar una decisión. Bajo este esquema, existen decisiones futuras que jamás tomaría.

Si el actor decide ejecutar la acción a_1 entonces puede terminar en el estado x_1 con probabilidad 0,3 o x_2 con probabilidad 0,7. Veamos qué ocurriría en cada caso.

- Si llega al estado x_1, entonces ganará un total de 12 puntos: 2 puntos de utilidad más 10 puntos debido a los estados futuros.
- Si llega al estado x_2, entonces ganará un total de 28 puntos: 4 puntos debido al estado x_2 en sí, más 24 puntos por la mejor rama a partir de x_2. La rama de 10 puntos futuros ni siquiera se considera, ya que estando en x_2, la mejor opción es la rama izquierda de 24 puntos.

Por lo tanto, la opción a_1 reportaría una utilidad esperada de

$$E[U|a_1,x] = 0{,}3 \cdot (2+10) + 0{,}7 \cdot (4+24) = 23{,}2$$

Haciendo el mismo análisis para la decisión a_2 tendríamos,

$$\begin{aligned} E[U|a_2, x] &= 0{,}4 \cdot (2 + 17) + 0{,}3 \cdot (7 + 17) \\ &\quad +0{,}3 \cdot (2 + 10) \\ &= 18{,}4 \end{aligned}$$

Por lo tanto, la mejor decisión a^* a partir del estado x_0 de la figura sería $a^* = a_1$.

La pregunta ahora es: ¿cómo se calcularon estas *utilidades futuras*? Como puede intuirse, se obtienen por medio de un cálculo recursivo. En la literatura, el término asociado al estado x se conoce como *el valor $V(x, h)$ de un estado x hasta un horizonte h*. El valor $V(x, h)$ está dado por la fórmula:

$$V(x, h) = \max_a \sum_{x'} P(x'|a, x)(u(x') + V(x', h - 1))$$

para el caso $h > 0$, y

$$V(x, 0) = u(x')$$

para el caso base de $h = 0$.

Algoritmo

Dicho lo anterior, podemos enunciar el algoritmo Expectimax en su totalidad. Dados un espacio de estados $\mathcal{X}$, un espacio de acciones $\mathcal{A}$, probabilidades de transición $P(x'|a, x)$, función de utilidad parcial u y un horizonte h:

1. Desarrollar el árbol completo hasta una profundidad h.
2. Calcular recursivamente los valores $V(x, h)$ de cada nodo-estado por medio de la fórmula

$$\begin{aligned} V(x, h) &= \max_a \sum_{x'} P(x'|a, x)(u(x') + V(x', h - 1)) \\ V(x, 0) &= u(x') \end{aligned}$$

3. Inicializar x en el estado inicial x_0 del árbol y m en h.
4. Tomar la alternativa óptima a^* dada por

$$a^* = \arg\max_a \sum_{x'} P(x'|a, x)(u(x') + V(x', m - 1))$$

5. Verificar el estado resultante y llamarlo x. Disminuir m en 1.
6. Si $m = 0$ terminar. Sino, volver al paso 4.

Notas

1. Si tenemos un problema para el cual debemos tomar la mejor decisión a un horizonte de $h + 1$ pasos, ¿serviría resolver primero el caso a h pasos? La respuesta es no, pues al aumentar la frontera del análisis en un paso adicional, pueden entrar nodos al análisis que podrían cambiar por completo la estrategia.
2. Hay casos en donde no se conoce en forma explícita la profundidad de árbol, pero se sabe que es finita. Para estos casos Expectimax también funciona. Sin embargo, como es un algoritmo de fuerza bruta, requiere desarrollar el *árbol completo* antes de tomar la primera decisión.
3. En casos especiales, como por ejemplo en juegos contra un jugador contrincante, las probabilidades de transición pueden desarrollarse en forma explícita. Por ejemplo, si se supone que el contrincante empleará una estrategia perfecta (p.ej. que esté a su vez utilizando Expectimax), entonces Expectimax se convierte en Max-Min: el contrincante en sus jugadas minimiza mis utilidades futuras y yo las maximizo.
4. Cabe preguntarse si existen formas más eficientes Expectimax para escoger la estrategia óptima. Existen aproximaciones sub-óptimas, la mayoría basadas en heurísticas, y soluciones óptimas para clases de problemas particulares, pero el caso general sólo se cubre por Expectimax.