Informe de actividades realizadas durante la primera mitad de la estancia

Sergi Elizalde

Durante los primeros meses de mi estancia en Dartmouth College he trabajado en varios de los problemas mencionados en mi plan de trabajo. El avance más significativo ha consistido en obtener una cota inferior en el número máximo de funciones de inferencia de cualquier modelo gráfico.

Como explicaba en mi memoria, los modelos estadísticos se usan para resolver algunos problemas en bioinformática, como determinar qué partes del genoma se traducen a proteínas, o cómo una secuencia de ADN se transformó en otra durante la evolución, a través de una serie de mutaciones, inserciones y delecciones. Cada posible respuesta tiene una cierta probabilidad que depende de los parámetros del modelo. Cuando éstos se conocen, la respuesta más probable, llamada *explicación*, se obtiene resolviendo un problema de optimización combinatoria. La aplicación que envía cada observación a su explicación correspondiente se llama *función de inferencia*.

El número de funciones de inferencia es importante porque indica cómo los parámetros de modelo pueden afectar la solución. Un caso particular muy importante en biología es el de alineación óptima de secuencias. En este caso, las funciones de inferencia indican, dado un par de secuencias de ADN, cuál es el proceso de evolución más probable que siguió una de ellas para transformarse en la otra. Otro caso particular de funciones de inferencia son las llamadas *funciones de búsqueda de genes*. Estas funciones determinan qué partes de una secuencia dada de ADN son *exones* y qué partes son *intrones*. Esta distinción es importante porque los exones son las partes que codifican proteínas, mientras que los intrones, que forman la mayor parte del genoma, son segmentos intermedios de los que no se conoce su función, pues nunca son traducidos a proteínas.

Antes de empezar mi estancia en Dartmouth College, uno de los resultados principales en este tema era una cota superior en el número de funciones de inferencia de cualquier modelo gráfico. Esta cota es polinómica en el tamaño del modelo. Una pregunta natural que se desprendía de este resultado era si la cota es ajustada (*"tight"*) o no. Si consiguiésemos determinar que es ajustada, esto implicaría que nuestra fórmula describe con precisión el comportamiento asintótico del número de funciones de inferencia de un modelo gráfico en el peor caso (*"worst case"*). Si, por el contrario, la cota no fuese ajustada, entonces sería posible encontrar una cota mejor.

Durante la primera mitad de mi estancia en Dartmouth he trabajado con Kevin Woods en este problema, y hemos conseguido obtener una cota inferior en el número máximo de funciones de inferencia de modelos gráficos. Esta cota coincide, salvo una constante multiplicativa, con la cota superior que se conocía. Esto demuestra que la cota es ajutada, y por tanto es óptima. En particular, la cota describe con precisión (salvo constante multiplicativa) el número funciones de inferencia de un modelo gráfico en términos del tamaño del modelo.

El método usado para conseguir la cota inferior consiste en constuir un modelo de Markov oculto (*"Hidden Markov model"*, o HMM) de tamaño n con d parámetros, para cualquier valores de n y d, cuyo número de funciones de inferencia coincide asintóticamente con la cota superior que se conocía. Una vez descrito el HMM, algunas de las herramientas que usamos para demostrar que tiene el número de funciones de inferencia deseado incluyen arreglos de hiperplanos (un tema de actualidad en combinatoria) y teoría de números.

Precisamente esta semana asistí a la conferenia internacional Formal Power Series and Algebraic Combinatorics 2006 que se celebró en San Diego del 18 al 23 de junio, donde presenté estos últimos resultados en una ponencia. Además, junto con Kevin Woods estamos escribiendo un

artículo con los resultados mencionados, titulado *Bounds on the number of inference functions of a graphical model*. El artículo lo enviaremos próximamente a la revista *Statistica Sinica*.