Importancia de variables en el contexto de un análisis exploratorio

Gabriel Illanes - Centro de Matemática, Facultad de Ciencias, UdelaR.
Cecilia Aguerrebere - Centro de Estudios Fundación Ceibal, Uruguay.
Horacio Botti - Laboratorio de Biofísica Integrativa, Departamento de Biofísica, Facultad de Medicina, UdelaR. Unidad de Bioinformática, Institut Pasteur de Montevideo.
Flavio Pazos Obregón - Departamento de Biología del Neurodesarrollo, IIBCE, Motevideo, Uruguay. Instituto de Matemática y Estadística "Rafael Laguarda", Facultad de Ingeniería, UdelaR, Uruguay.
Gregory Randall - Instituto de Ingenieria Electrica, Facultad de Ingenieria, UdelaR.
Cameron MacPherson - Chief Data Scientist, bioStone Consulting.

Muchas veces, disponemos de un conjunto de datos del cual no sabemos demasiado, y mucho menos disponemos de un modelo matemático diseñado que lo explique. En otras palabras, es posible que deseemos encontrar las variables importantes de nuestro conjunto de datos, pero no sepamos exactamente qué significa que una variable es importante.

En general, nuestro conjunto de datos proviene de un problema enmarcado en un área de la ciencia (en economía podemos disponer de datos bancarios, en programación podemos disponer de datos sobre redes sociales, etc), lo cual nos permite generar cierta intuición, y diversos criterios sobre qué significa que una variable sea importante. Muchas veces, dichos criterios están pobremente definidos, e incluso pueden llegar a ser contradictorios. Estas características sobre el contexto de los datos suele derivar en que el conocimiento previo sobre los datos se use como criterio de validación del análisis, y no como datos que incorporamos en la exploración.

Considerando esta motivación, la idea es repasar el proceso del análisis exploratorio de un conjunto de datos genómicos, que consiste en conteos de RNA mensajero relativo a 587 genes en la sangre de 25 personas sanas al ser estimuladas (in vitro) con 28 estímulos de distinta naturaleza. Una descripción de los datos puede ser encontrada en http://www.cell.com/cell-reports/fulltext/S2211-1247%2816%2931057-9. Si bien uno de los objetivos del análisis exploratorio de los datos es identificar los genes más influyentes para cada uno de los estímulos, la definición de "influyente" no es clara, por lo cual hay que estudiar distintos modelos que apunten a la detección de distintos tipos de patrones en los datos, y qué herramientas se encuentran disponibles para la validación de los resultados.

Este trabajo se enmarca en los proyectos *Milieu Intérieur* (http://www.milieuinterieur.fr/en) y *Healthy Human Global Project* (https://research.pasteur.fr/en/program_project/the-healthy-human-global-project/), del Institut Pasteur, París.