EL PAÍS, miércoles 16 de abril de 2008

TREVOR HASTIE Matemático

"Con las estadísticas siempre hay formas de hacer trampas"

CLEMENTE ÁLVAREZ
Madrid

Cuando el profesor de estadística y bioestadística Trevor Hastie (Suráfrica, 1953) imparte una conferencia, entre el público pueden estar sentados matemáticos, médicos, financieros e incluso aquellos que propagan por Internet el llamado *spam* o correo basura. "Vienen para aprender las últimas técnicas desarrolladas para detectarlos", cuenta el director del Departamento de Estadística de la Universidad de Stanford (EE UU). Este investigador, que estuvo la semana pasada en Madrid invitado por la Fundación BBVA, centra su trabajo en el campo del *data mining*, es decir, modelos estadísticos intensamente computerizados que se ocupan de enormes cantidades de información.

Pregunta. ¿Qué hace la estadística en campos tan diversos?

Respuesta. La mayor parte de las cosas que medimos en la vida comportan una incertidumbre, son fuentes aleatorias de error. Por eso solemos realizar más de una medición. En el pasado existían limitaciones, pero con los adelantos de la tecnología podemos llevar a cabo muchas más mediciones. Esto hace que cada vez haya más datos, pero también que se necesiten herramientas para sacar conclusiones. Ahí es donde la estadística desempeña su papel.

P. ¿Cómo de grandes son los conjuntos de datos con los que trabaja?

R. Hace 30 o 40 años, cuando hablábamos de muchos datos nos referíamos a cientos de observaciones y decenas de variables. Ahora tenemos experimentos de física con conjuntos de millones de valores o análisis financieros con decenas de millones de observaciones.

P. ¿En qué casos trabaja con más información?

R. El mayor generador de datos hoy en día es Internet. El número de usuarios crece cada día y ya son cientos de millones en el mundo. Esto hace que la cantidad de información crezca de forma exponencial, y que tengamos un número infinito de datos. Probablemente, la mejor forma de hacer frente a este enorme volumen de información sea algo como el buscador Google, por los algoritmos que ha desarrollado.

P. Una de las áreas de aplicación es la medicina. ¿Hasta qué punto son eficaces las estadísticas en este campo?

R. Veamos un ejemplo. Yo he trabajado durante cinco años en la Universidad de Stanford con especialistas en cáncer de mama. Para realizar los diagnósticos, los médicos utilizan generalmente factores pronóstico, mediciones como el tamaño del tumor, el grado del tumor, si los ganglios linfáticos están implicados... Los oncólogos suelen ser muy conservadores y tratan con quimioterapia a la mayor parte de las pacientes. Ahora podemos efectuar mediciones a partir de la genómica, que abarcan cientos de genes y que permiten mejorar la capacidad de pronóstico de los oncólogos en un 30%. ¿Qué significa esto? Por medio de lo que se conoce como firma genética se consigue que haya un 30% de estas mujeres que ya no tengan que someterse a quimioterapia. Podemos perfilar mucho mejor quién debe recibir qué tipo de tratamiento.

P. ¿Sus investigaciones también se utilizan para detectar el correo basura?

R. Sí, se puede predecir si un correo electrónico es *spam* basándose en determinadas palabras del mensaje, como, por ejemplo, "tú", unos signos de admiración o el símbolo del dólar. Hoy los filtros funcionan a la medida del usuario. Hay algoritmos que, tras un periodo de aprendizaje, pueden predecir qué será considerado correo basura. También debo decir que la industria del *spam* ha crecido mucho. Ahora imparto conferencias sobre algoritmos para detección del correo basura. ¿Y quién cree que acude a mis conferencias? Los que crean el *spam*, para aprender las últimas técnicas desarrolladas para detectarlos. Así, tres semanas después aparecen nuevos sistemas de correo basura.

> "¿Quién acude a mis conferencias? Los que crean el 'spam', para aprender"

> "El mayor generador de datos hoy en día es Internet"

P. Las estadísticas se emplean igualmente para la publicidad de la Red. ¿No es así?

R. Las estadísticas son fundamentales en Internet. Hoy en día hay mucha actividad en torno a la publicidad. Si abres una página te sale un anuncio y lo que se intenta es que esa publicidad esté hecha a medida de cada usuario, según los sitios que ha visitado con anterioridad. Al principio no parece agradable, pues es como si te vigilaran. Pero en el fondo, si yo tuviese que hacer publicidad en la Red... pienso que tiene un sentido.

P. ¿Todo esto aumenta mucho el poder de los estadísticos?

R. Los estadísticos han ido adquiriendo más poder. En la Universidad de Stanford, cuantos más estadísticos formamos, más nos pide el mercado. Para genómica, bioinformática, farmacéuticas, financieras... Todos los fondos de cobertura (los *hedge funds*) exitosos cuentan con un equipo de estadísticos. Y no digo que sea un uso muy noble de la estadística, pero es así.

P. ¿Nos podemos fiar de las estadísticas?

R. Hay que ser cautos con los resultados de las estadísticas, pues con las estadísticas siempre hay formas de hacer trampas. Ahora se puede ver con facilidad si esto ocurre, pero no deja de existir la amenaza, especialmente en las ciencias médicas. Hay investigadores que sobreinterpretan los resultados para poder publicar sus trabajos. A medida que crece el uso de las estadísticas en las ciencias médicas y biológicas, aumenta también su abuso.