Medicina Clínica Práctica 4 (2021) 100296

MEDICINA CLÍNICA PRÁCTICA

www.elsevier.es/medicinaclinicapractica

Comunicación breve

Valorar la predicción

Assessing the prediction

Erik Cobo*

Departamento de Estadística e Investigación Operativa, Universitat Politècnica de Catalunya/BarcelonaTech, Barcelona, España

Un objetivo distinto de investigación pretende analizar hasta qué punto es posible anticipar los valores de una variable «respuesta» a partir de los valores de ciertas variables «predictoras». En meteorología, por ejemplo, deseamos anticipar el tiempo, no modificarlo. Si tenemos que seleccionar un método de previsión del tiempo, conviene conocer con qué precisión podemos anticiparlo: qué método, o qué formula reduce más la incertidumbre previa sobre esa variable respuesta.

Coeficiente de determinación R^2

Supongamos que deseamos anticipar el peso de la próxima persona que se sentará en la silla de nuestro despacho. En ese entorno, admitamos que el peso corporal tiene una media de 70 kg y una desviación típica (SD) de 10 kg. Predeciremos 70 kg y esperaremos equivocarnos en 10^2 kg^2 (tiene su sentido utilizar el cuadrado de la escala para expresar error o variabilidad, pero no nos extenderemos aquí). Pero si una puerta translúcida nos permite averiguar que la persona que está a punto de entrar mide 1,9 metros, nos interesa saber las características del peso de las personas de 1,9 metros. Pongamos que tengan una media de 90 kg y una SD de 7 kg. Si cambiamos el valor predicho de 70 a 90 kg, reducimos el error de predicción de 10^2 a 7^2 kg^2, bajando la incertidumbre de nuestra predicción de 100 kg^2 a 49 kg^2: una reducción del 51%. De forma similar calculamos el R^2 o coeficiente de determinación, que cuantifica la capacidad de predecir en la regresión lineal. Ahora bien, un 51%, ¿es mucho o es poco? Depende del ejemplo y de las acciones que debamos emprender. O más simple, depende de las alternativas. Si, hasta el momento, el mejor modelo predecía un 48%, pues muy poco lo hemos mejorado. Sin embargo, si añadiendo la edad y el género el R^2 puede subir al 64%, por ejemplo, este nuevo modelo es más discriminativo y supone un modelo mejor.

Bondad del ajuste

Clásicamente, se hablaba de «bondad del ajuste». En cambio, Tripod[1] habla de 2 conceptos: discriminación y calibrado (Figura 1). Veámoslos para predecir un *outcome* o respuesta dicotómica: la aparición o no de un evento o desenlace de interés clínico, como podría ser la curación después de una intervención.

Discriminación

Discriminación es la capacidad del modelo para acertar el futuro, para anticipar, o discernir. Que el modelo prediga que el paciente curará y realmente cure. Por ejemplo, que el modelo acierte en el 90% de los que ha predicho que curarán; y en el 80% de los que ha dicho que no curarán. Si el resultado del modelo es una escala o índice de curación con distintos niveles o graduaciones, la medida más popular de discriminación es el área bajo la curva ROC, que valora el grado de separación entre las distribuciones de los casos que curan, y la de los que no lo hacen. Cuanto más separadas estén, mayor será su valor. Esta área bajo la curva ROC coincide con los estadísticos clásicos de Mann-Whitney y de Wilcoxon, y ahora es más conocido como estadístico «C». Compara todas las posibles parejas formadas por un paciente que cura y uno que no lo hace, obteniendo la proporción de parejas en las que el caso curado tenía un valor mayor en la escala de curación. Un estadístico «C» del 100% indicaría una predicción perfecta: todos los curados tendrían un valor predicho mayor que cualquiera de los no curados; y un valor del 50% que, en la mitad de las parejas, el curado tenía un valor mayor; y en la otra mitad, el no curado. Es decir, un indicador con un valor de 50% del estadístico «C» equivale a lanzar una moneda al aire: no anticipa nada. En resumen, la discriminación valora la utilidad del modelo para anticipar el valor de la variable predicha. Para una explicación más detallada, encontrará en Youtube un vídeo de varios autores de nuestra universidad[2].

Calibrado

Por su parte, el calibrado estudia la corrección del modelo, en el sentido de que los valores predichos se corresponden con los valores promedios observados. En el caso de predecir una dicotomía, que las probabilidades predichas por el modelo cuadren con las proporciones observadas. La tabla 1 muestra un calibrado muy bueno en una escala para predecir la aparición de una neumonía postoperatoria[3]. Proporciona la correspondencia entre: el resultado predicho por el modelo («probabilidad predicha», segunda fila) y las proporciones estimadas en las 2 muestras usadas, una para generar el modelo («aprendizaje», tercera fila), y otra para confirmar su rendimiento («validación», cuarta fila). Las 5 columnas muestran el riesgo predicho para distintos valores de su escala: a la izquierda, con bajo riesgo, entre 0 y 15 puntos; y a la derecha, con alto riesgo, más de 55 puntos. Observe la mayor frecuencia de casos con menor riesgo, a la izquierda: casi 70.000 casos; y la muy menor a la derecha, con menos de 100.

* Autor para correspondencia.
Correo electrónico: erik.cobo@upc.edu.

https://doi.org/10.1016/j.mcpsp.2021.100296

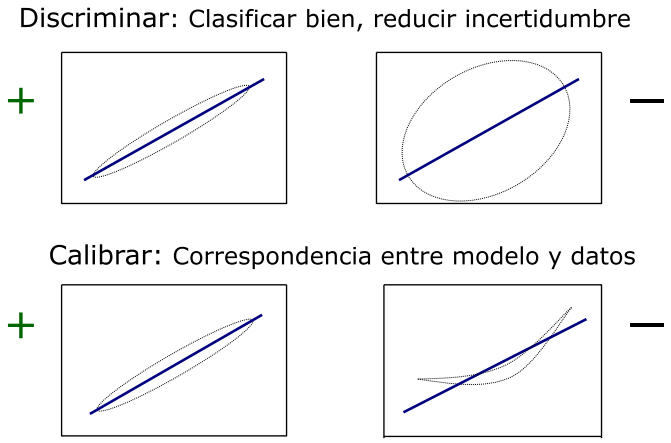

Figura 1. Los gráficos representan, para una respuesta numérica, en su eje horizontal de abscisas, el valor predicho; y en el vertical de ordenadas, el valor observado. El superior izquierda muestra buena discriminación; el derecho, mala; y los inferiores, buen y mal calibrado. El gráfico inferior derecho muestra un mal calibrado que, en cambio, permitiría una buena discriminación, ya que acertaría bastante en sus predicciones.

Tabla 1
Ejemplo de muy buen calibrado en la predicción del riesgo posquirúrgico de neumonía. En la segunda columna, para valores bajos de la escala (0 a 15 puntos), con bajo riesgo, el calibrado es perfecto: con 4 decimales coincide la probabilidad predicha con las proporciones observadas en ambas muestras. En la última columna, con menos de 100 casos, no es tan bueno: predice una probabilidad de 0,153 y observa una proporción de 0,158 en la muestra de aprendizaje y de 0,159 en la de validación

Grupo de riesgo Puntos en la escala original	1 0-15	2 16-25	3 26-40	4 41-55	5 > 55
Número de pacientes en muestra aprendizaje %	69.333 43%	44.757 35%	32.103 20%	3.517 2%	95 0,1%
Promedio de las **probabilidades predichas** por el modelo **Proporción** observada en muestra **aprendizaje** **Proporción** observada en muestra **validación**	0,0024 0,0024 0,0024	0,0120 0,0119 0,0118	0,040 0,040 0,046	0,094 0,094 0,108	0,153 0,158 0,159

Fuente: Adaptada de Arozullah AM et al[3].

Observe también la perfecta concordancia entre las probabilidades predichas y las proporciones observadas en la primera columna de bajo riesgo y alta frecuencia: un 0,0024. Y la buena concordancia en la columna de la derecha de alto riesgo y baja fecuencia: una probabilidad de 0,153 y proporciones del 15,8% y del 15,9% en las muestras de aprendizaje y de validación.

Digamos, para terminar, que Tripod desaconseja los valores de P. Por ejemplo, sobre la clásica prueba de Hosmer-Lemeshow, el documento explicativo de Tripod dice[4] que suele resultar casi siempre positiva («estadísticamente significativa») en muestras grandes; y casi nunca en muestras pequeñas.

En resumen, en un estudio de predicción, valore la capacidad de discriminación, posiblemente con el área bajo la curva ROC o con el coeficiente de determinación R^2; y también el calibrado, valorando si las predicciones medias predichas por el modelo coinciden con los resultados observados.

Financiación

PID2019-104830RB-I00 DOI (AEI): 10.13039 / 501100011033: STATISTICAL METHODOLOGIES FOR CLINICAL AND OMICS DATA AND THEIR APPLICATIONS IN HEALTH SCIENCES (SAMANTHA) del Ministerio de Ciencia e Innovación.

Responsabilidades éticas

No implica pacientes y no requiere permiso ético.

Bibliografía

1. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. J Clin Epidemiol. 2015 Feb;68(2):134–43 PMID: 25579640.
2. Curva ROC: aplicación al diagnóstico médico. Video en Youtube. [Consultado el 31/10/2019]. https://youtu.be/pA4E2uVHiYo
3. Arozullah AM, Khuri SF, Henderson WG, Daley J. Participants in the National Veterans Affairs Surgical Quality Improvement Program. Development and validation of a multifactorial risk index for predicting postoperative pneumonia after major noncardiac surgery. Ann Intern Med. 2001 Nov 20;135(10):847–57.
4. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162(1): W1–73. 25560730.