Otros métodos de regresión lineal

UCO ONLINE

Otros métodos de regresión lineal

1. Introducción

En los temas anteriores hemos estudiado distintas curvas de ajuste para construir modelos de regresión lineal múltiple. Estos modelos permiten relacionar de distinta manera las variables de entrada (o predictores). Sin embargo, se ha mantenido constante la idea de que a cada predictor se le asocia un coeficiente que representa su implicación en la predicción de la variable de salida. A medida que el número de predictores aumenta, mantener todos los predictores no es recomendable ya que afecta a dos aspectos fundamentales en los modelos de aprendizaje supervisado:

- Precisión en la predicción: métodos como el de los mínimos cuadrados funcionan bien cuando el número de observaciones es muy superior al número de predictores. Esto es, el modelo estimado tendrá poca variabilidad y responderá bien en la partición de test. Pero si el número de observaciones es bajo, el modelo puede sufrir sobreajuste. Es más, si el número de predictores es superior al de observaciones, el método de los mínimos cuadrados no nos garantiza una estimación única.
- Interpretabilidad del modelo: es posible que algunos de los predictores no influyan realmente en la variable de salida, o lo hagan escasamente. Mantener variables de entrada no asociadas con la predicción hace que se generen modelos más complejos que pueden ser más difíciles de interpretar.

Por los dos motivos expuestos, han surgido otros métodos de regresión lineal que buscan reducir el número de predictores necesarios para explicar la respuesta de la variable de salida. Estos métodos pueden basarse en varias estrategias:

- Selección de un subconjunto de predictores: consiste en identificar los predictores más relevantes para la predicción, aplicando luego un método estándar (como el de los mínimos cuadrados) para construir el modelo lineal.
- Métodos de contracción (*shrinkage*): son métodos que estiman los coeficientes para todos los predictores tratando de reducir la variabilidad del modelo. Para ello, intentan "truncar" algunos de los coeficientes a valores cercanos a cero. Algunos métodos pueden llegar a asignar cero al coeficiente, actuando como métodos de selección de variables.
- Reducción de la dimensionalidad: son métodos que proyectan el conjunto de predictores en un espacio de menor dimensionalidad, calculando para ello combinaciones lineales entre ellos. A continuación, aplican un método de regresión lineal sobre las variables resultantes.

En lo restante de este tema, nos centraremos en dos métodos de contracción: *Ridge* y *LASSO*.

2. Concepto de regularización

Los métodos de contracción se fundamentan en el concepto de *regularización*. La regularización consiste en añadir información al problema, normalmente en forma de penalización ante el incremento de la complejidad. Las técnicas que aplican este concepto son útiles para prevenir el sobreajuste, pues tratan de buscar el modelo más simple que es capaz de seguir explicando adecuadamente los datos. Esto es, se trata de reducir el número de predictores o los valores de sus coeficientes sin que por ello se pierda la capacidad del modelo de representar la relación entre los predictores y la variable de salida.

Al contrario que los métodos de selección de predictores, que directamente ajustan el modelo con un subconjunto de los predictores seleccionado a priori, los métodos de contracción parten de todo el conjunto de predictores. Es durante el proceso de ajuste cuando se decide qué predictores se van "reteniendo", y cuáles se van "descartando" (si el método acepta que el coeficiente sea cero) debido a que su presencia aumentaría la variabilidad del modelo.

Para entender cómo funcionan los métodos de regresión, necesitamos retomar la ecuación de que nos calcula el error cuadrático medio de un modelo lineal múltiple:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad (1)$$

La regularización consiste en añadir un término de penalización a la expresión anterior, que únicamente actúa sobre los valores de los coeficientes asociados a los predictores ($\beta_j \; \forall j > 0$). Dicha penalización se añade al cálculo de RSS de la siguiente forma:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 - \lambda f(\beta) \quad (2)$$

El parámetro λ es el que *regula* la penalización que se aplica. Si $\lambda = 0$, el método no aplica regularización y, por tanto, se comporta como el método de los mínimos cuadrados. A medida que $\lambda \to \infty$, el impacto de la regularización es mayor y los coeficientes tenderán a ser cero. Por tanto, la elección del valor para el parámetro λ es de gran importancia. La segunda parte del término de penalización, $f(\beta)$, hace referencia a cómo se ven afectados los coeficientes, y es dependiente del método concreto que se aplica. A continuación, se explican los dos métodos más importantes:

- Ridge, que aplica la norma l_2 como regularización, busca reducir todos los coeficientes, especialmente los que tienen valores más grandes.
- LASSO, que aplica la norma l_1 como regularización, permite descartar algunos de los coeficientes generando modelos *dispersos*.

3. Método Ridge

El método Ridge, también conocido como regularización de Tikhonov, penaliza los coeficientes que toman valores más altos. Para ello, aplica la norma l_2 sobre los coeficientes, por lo que la función de coste a minimizar se formula como sigue:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 - \lambda\sum_{j=1}^{p}\beta_j^2 \quad (3)$$

Para minimizar esta expresión, donde los valores de los coeficientes van elevados al cuadrado, es preferible que estos coeficientes tomen valores bajos. Ridge aplica el mismo factor de reducción a todos los coeficientes, de forma que todos ellos se mantienen en el modelo de regresión final. No obstante, el valor de los coeficientes varía en función del valor asignado al parámetro λ.
La estimación de coeficientes mediante el método Ridge puede reescribirse de la siguiente manera:

$$\hat{\beta}^{ridge} = min\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2,\ sujeto\ a \sum_{j=1}^{p}\beta_j^2 \leq t \quad (4)$$

El parámetro t tiene una correspondencia directa con λ, y es el parámetro que restringe el tamaño que pueden alcanzar los coeficientes. El método Ridge es especialmente adecuado cuando el número de predictores es superior al número de observaciones, o cuando existen dependencias entre el conjunto de predictores (colinealidad). Imaginemos una variable que tiene asociado un coeficiente positivo muy alto, pero está correlacionado con otra variable que tiene asociado un coeficiente negativo muy bajo. Ambas tienen una alta influencia en el modelo frente a otros predictores, pero en realidad, se anulan mutuamente. Añadiendo la restricción de que los coeficientes no tomen valores muy altos, este problema se alivia.
Un aspecto importante es que el método Ridge requiere que todos los predictores varíen en la misma escala de valores, algo no necesario, aunque sí recomendable, cuando se aplica el método de los mínimos cuadrados.

4. Método LASSO

El método LASSO (*Least Absolute Shrinkage and Selection Operator*) intenta reducir el número de coeficientes. Para ello, añade la norma l_1 al cálculo de la expresión de coste:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 - \lambda\sum_{j=1}^{p}|\beta_j| \quad (5)$$

De forma equivalente a lo explicado para Ridge, la estimación de coeficientes mediante el método LASSO puede reescribirse de la siguiente manera:

$$\hat{\beta}^{lasso} = min\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2,\ sujeto\ a\sum_{j=1}^{p}|\beta_j| \leq t \quad (6)$$

A diferencia de Ridge, LASSO puede producir modelos donde algunos de los coeficientes son descartados, esto es, se le asigna un coeficiente igual a cero. Esto no es posible con Ridge salvo que $\lambda = \infty$. Eliminar predictores interesa especialmente entre las variables que están muy correlacionadas, de forma que LASSO las detecta y elimina alguna de ellas del modelo. La norma l_1 tiene el efecto de permitir forzar a que los coeficientes tomen el valor cero cuando el valor del parámetro λ es lo suficientemente largo. Por otra parte, LASSO también es interesante en aquellas situaciones en las que el número de predictores es elevado, pues permite obtener modelos basados en un subconjunto de predictores que son más fáciles de interpretar. Al igual que Ridge, la selección del parámetro λ es determinante, y diferentes valores producirán diferentes modelos de regresión.

Cabría pensar que LASSO tenderá a generar modelos menos precisos que Ridge, ya que previsiblemente descartará predictores que sí estarán en el modelo creado por Ridge. Por otro lado, si el conjunto de predictores que realmente tienen relación con la variable de salida es un subconjunto del total de predictores disponibles, entonces LASSO parece tener una ventaja frente a Ridge, pues este último forzará a que todos los predictores aparezcan en el modelo de regresión, aunque tengan escasa influencia. En general, no es posible saber a priori cuántos predictores están realmente relacionados con la variable de salida, por lo que predecir de antemano qué método es más adecuado no es trivial y dependerá en gran medida de la distribución de los datos y la configuración de ambos métodos.

Referencias

G. Hackeling. “Mastering Machine Learning with Scikit-Learn”. Pactkt Publishing, 221 páginas. 2014.

T. Hastie, R. Tibshirani, J. Friedman. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. Springer Series in Statistics, 2ª edición, 745 páginas. 2017.

G. James, D. Witten, R. Tibshirani, T. Hastie. “An Introduction to Statistical Learning with Applications in R”. Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/