Selección de características

Blanca Vázquez

16 de octubre de 2023

Introducción

El objetivo de los métodos de selección de características es **reducir** el número de variables de entrada:

- Eliminar variables repetidas o información redundante.
- Eliminar información no relevante.
- Ayuda a mejorar el rendimiento de los modelos.

La pregunta es...

¿Qué características deberían seleccionarse para garantizar un rendimiento óptimo?

Método 1: selección exhaustiva de características

Consiste en evaluar todas las posibles combinaciones de características.

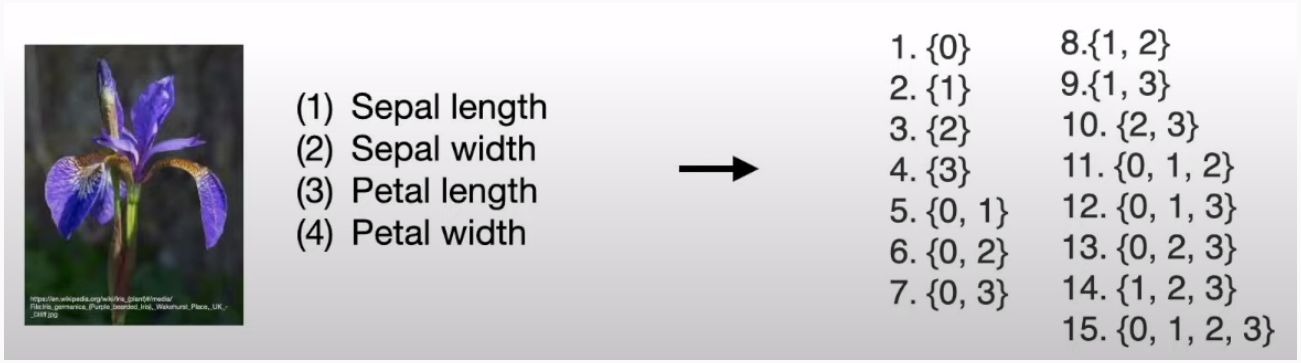

Imagen tomada de Raschka, "SequentialFeatureSelector", 2022.

Método 1: selección exhaustiva de características

Consiste en evaluar todas las posibles combinaciones de características.

$$\sum_{i=1}^{m} \binom{m}{i} \textit{Combinaciones}$$

$$\binom{4}{1} + \binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 15$$

Método 1: selección exhaustiva de características

Ahora, supongamos un conjunto de datos con 13 variables.

$$\sum_{i=1}^{m} \binom{m}{i} \textit{Combinaciones}$$

$$\binom{13}{1} + \binom{13}{2} + ... + \binom{13}{13} = 8191$$

Este método de selección es muy caro.

Método 2: Selección secuencial de características

Los métodos de *Sequential feature selection* son un conjunto de algoritmos voraces (*greedy*) que se usan para reducir d dimensiones hacia un espacio de tamaño k donde $k < d$.

- La idea es seleccionar automáticamente un subconjunto de características relevantes al problema.
- Eliminan características irrelevantes o con ruido.
- Son algoritmos eficientes.
- Disminuyen el error de generalización en los modelos.
- Los tipos más comunes son: *backward* y *forward*.

Método 2: Selección secuencial de características (backward)

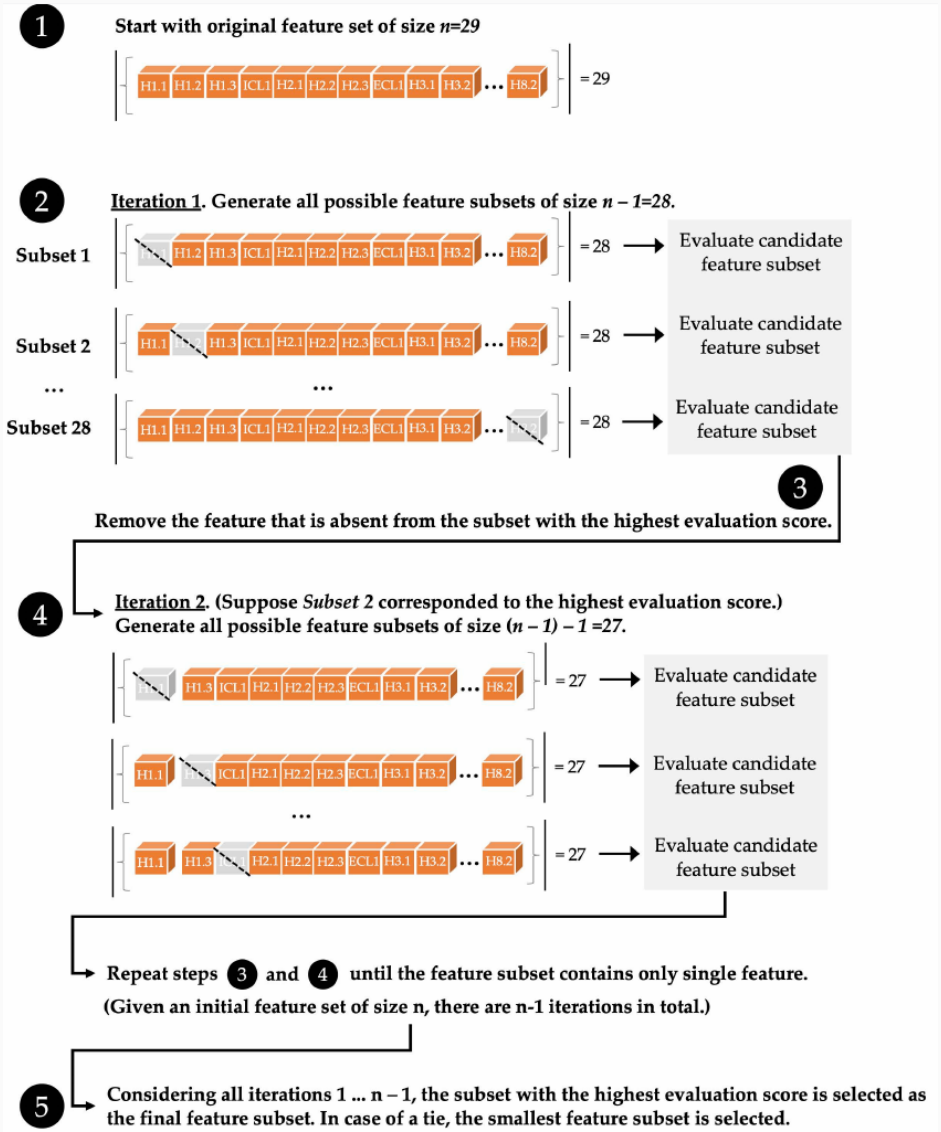

Imagen tomada de Joe Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn (2020) Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition Biomolecules 2020, 10, 454.

Método 2: Selección secuencial de características (backward)

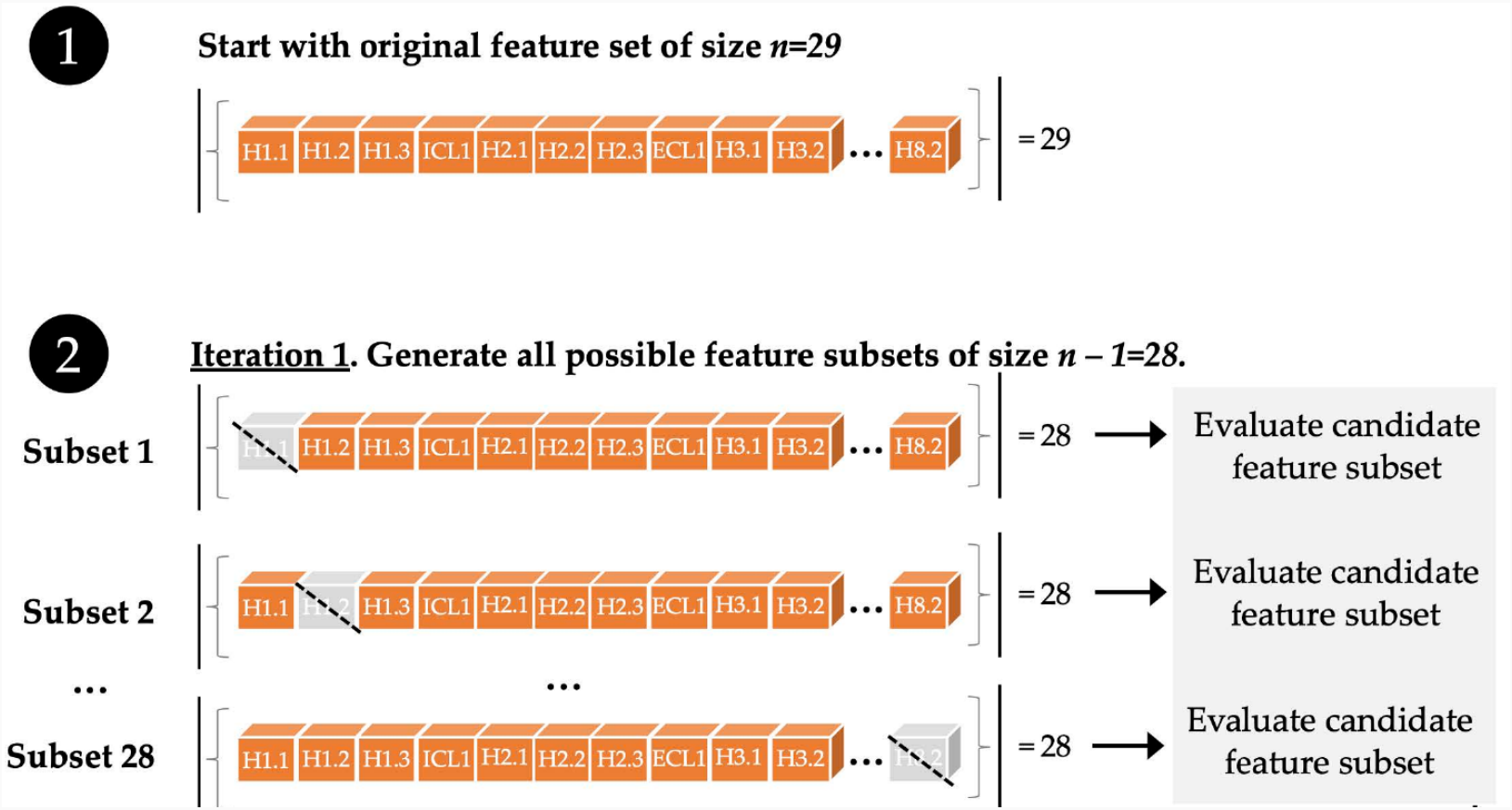

Imagen tomada de Joe Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn (2020) Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition Biomolecules 2020, 10, 454.

Método 2: Selección secuencial de características (backward)

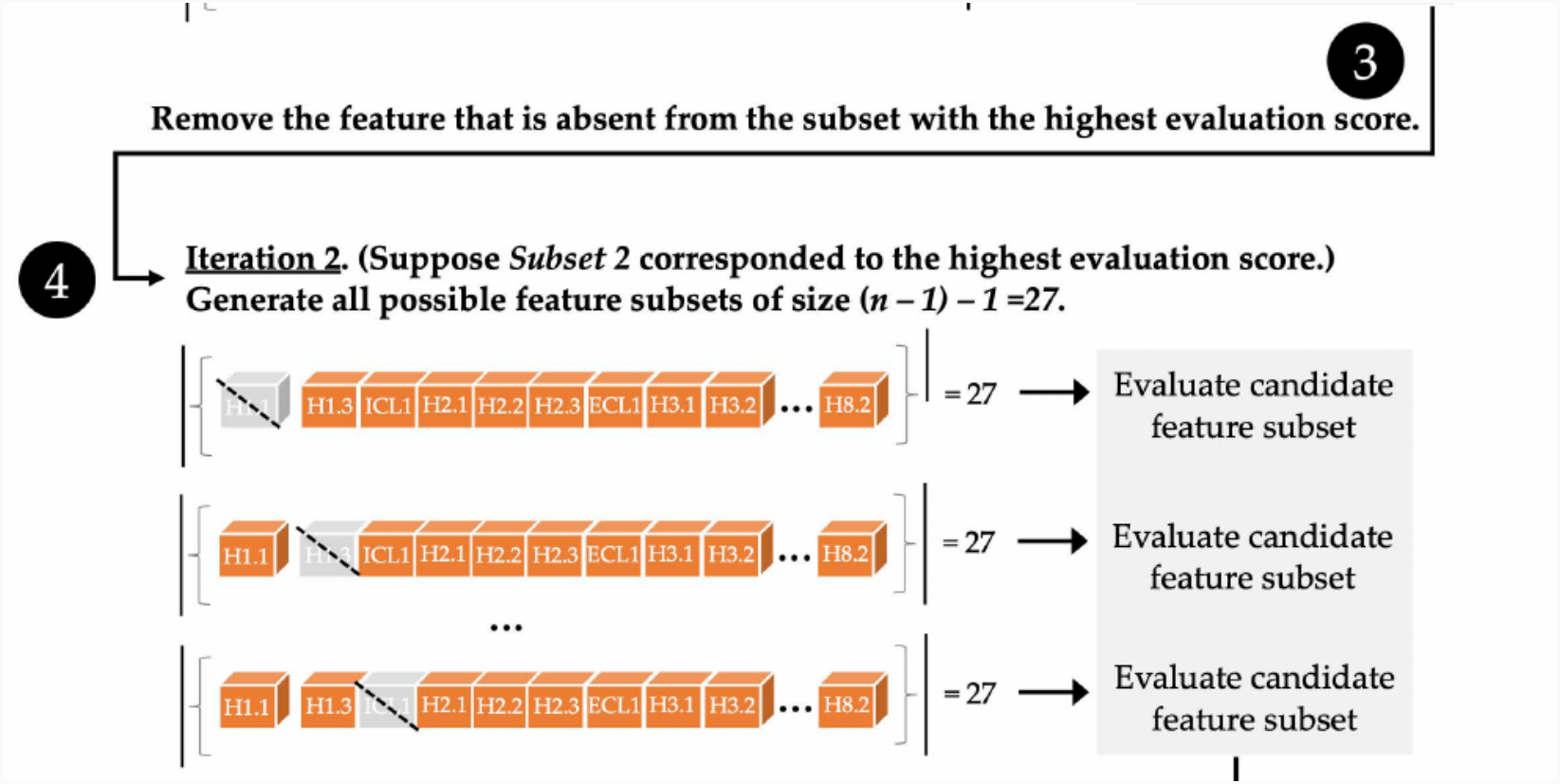

Imagen tomada de Joe Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn (2020) Machine Learning to Identify Flexibility Signatures of Class A GPCR Inhibition Biomolecules 2020, 10, 454.

Método 2: Selección secuencial de características (forward)

1. Create an empty set: $Y_k = \{\emptyset\},\ k = 0.$
2. Select best remaining feature:
   $x^+ = arg\ max_{x^+ \in Y_k}[J(Y_k + x^+)]$
3. If $J((Y_k + x^+) > J((Y_k)$
   a. Update $Y_{k+1} = Y_k + x^+$
   b. $k = k + 1$
   c. Go back to step 2.

Imagen tomada de Smith, Ashley; Mendoza-Schrock, Olga; Kangas, Scott;Dierking, Matthew; Shaw, Arnab. (2014). An end-to-end vechicle classification pipeline using vibrometry data. Proceedings of SPIE - The International Society for Optical Engineering.

¿Cuándo usar selección de tipo forward o backward?

Es importante comentar que ambos métodos no siempre generan los mismo resultados.

- Supongamos, un conjunto de datos de tamaño 100 y queremos seleccionar 5 características.
- Supongamos, un conjunto de datos de tamaño 100 y queremos seleccionar 95 características.

Método 3: Análisis de componentes principales (PCA)

Es un método para reducir dimensiones sin eliminar directamente características.

- **Feature selection**: selecciona un subconjunto de características que son relevantes a la predicción.
- **Feature extraction**: crea un conjunto nuevo de características que capturan la información más relevante.
  - Las nuevas características eliminan información redundante.
  - A estas características se les llama: componentes principales

FS preserva las características originales, mientras que FE crea un conjunto de características.

Intuición

Supongamos que tenemos una muestra de n individuos cada uno con p variables $(X_1, X_2, ...X_p)$, es decir, el espacio es de p dimensiones.

PCA tiene que encontrar un conjunto de factores $(z < p)$ que explique aproximadamente lo mismo que las variables p originales.

Por lo tanto, antes necesitábamos p variables para identificar a un individuo, ahora basta con z valores.

Cada componente principal (Z_i) se obtiene por combinación lineal de las variables originales que maximizan la varianza de las observaciones.

Cálculo de las componentes principales

Dado un conjunto de datos X con n observaciones y p variables, el proceso a seguir para calcular la primera componente principal es:

1. Centralización de las variables: se resta a cada valor la media de la variable a la que pertenece. Con esto se consigue que todas las variables tengan media cero.
2. Se resuelve un problema de optimización para encontrar el valor de los pesos con los que se maximiza la varianza.
    - Una forma de resolver esta optimización es mediante el cálculo de *eigenvector-eigenvalue* de la matriz de covarianzas.

Cálculo de las componentes principales

Una vez calculada la primera componente ($Z1$) se calcula la segunda ($Z2$) repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no pude estar correlacionada con la primera componente.

Esto equivale a decir que $Z1$ y $Z2$ tienen que ser perpendiculares.

Algunas consideraciones

El método de PCA es efectivo cuando:

- Los datos son numéricos
- Es sensible a la escala.
- Es sensible a los valores atípicos.