Selección de variables

Máster en Data Science y Big Data en Finanzas (MDS_F)
Máster en Data Science y Big Data (MDS)

José Ramón Sánchez Leo
rsanchez@afi.es

Febrero 2022
Afi Escuela

Índice

1. Objetivos

¿Qué es la selección de variables?

La **selección de variables** en aprendizaje automático es el proceso destinado a elegir subconjuntos de variables especialmente útiles para la construcción de un buen modelo predictivo. Los objetivos que se persiguen son los siguientes:

- Facilitar el entendimiento y la visualización de los datos.
- Reducir los requisitos de almacenamiento.
- Reducir tiempo de entrenamiento y predicción de los modelos.

¿Qué es la selección de variables?

¿Tienes conocimiento de negocio?

En este caso, construye, y considera las variables relevantes para el problema.

¿Qué es la selección de variables?

¿Tienes conocimiento de negocio?

En este caso, construye, y considera las variables relevantes para el problema.

¿Necesitas reducir el número de variables?

- Por coste.
- Por tiempo.
- Por interpretabilidad.

Pasos a seguir

- Descartar variables a partir de un análisis exploratorio.
- Utilización de métodos de filtrado a priori para eliminar variables.
- Construcción de modelos sencillos con pocas variables que den una intuición acerca del poder predictiva de las variables, descartando las peores.
- Construcción de modelos más complicados, empleando variables sugeridas por el proceso anterior.

2. Métodos de filtrado y rankeo de variables

2. Métodos de filtrado y rankeo de variables

Los métodos de filtrado de variables a priori tratan de eliminar variables irrelevantes o redundantes. Los métodos más comunes de filtrado para la selección de variables están basados en el **ranking de variables.**

Estos emplean algún criterio para evaluar cada variable, ordenándolas en función de su importancia. Finalmente, se seleccionarán subconjuntos de variables con una alta puntuación según el criterio utilizado.

Aunque no es el más "óptimo", es muy usado debido a su bajo coste computacional.

Metodología de rankeo

Dado un problema de aprendizaje supervisado del tipo "predecir la respuesta Y en base a los predictores $X_1, \dots, X_p$" el rankeo de variables consiste en:

1. Considerar una función de score S tal que para cada $j \in \{1,2, \dots, p\}$, $S(j)$ representa el valor de la variable X_j para predecir Y.
2. Ordenar las variables en sentido decreciente de S.
3. Utilizar las p' variables con mayor puntuación en el ranking.

Los diferentes métodos de filtrado vienen determinados por la función de score S y por la forma de elección de p' (normalmente por el criterio del codo).

Algunas funciones de score

- Para problemas de regresión:
  - $S(j) = \rho(X_j, Y)^2$, donde ρ puede ser un coeficiente de correlación (Pearson, Spearman, etc.).
  - $S(j) = \max\left\{\rho(X_j, Y)^2, \rho(\log(X_j), Y)^2, \rho(X_j^2, Y)^2\right\}$
- Para problemas de clasificación:
  - $S(j) = \rho(X_j, Y)^2$, donde ρ puede ser un coeficiente de correlación (Spearman) o χ^2, entre otros.
  - $S(j) = \kappa_{Y|X_j}$, para algún modelo $Y \sim X_j$,

    donde κ es una medida de Observed Accuracy/Predicted Accuracy

Correlación de Pearson

La existencia de correlación implica relación lineal, aunque no causalidad, es decir, pueden existir correlaciones que no indiquen relación "real". Sin embargo, la inexistencia de correlación lineal no implica independencia.

$$\rho_{X;Y} = \frac{cov(X;Y)}{\sigma(X)\sigma(Y)}$$

Interpretación

- 0.8-1.0 Correlación muy fuerte
- 0.6-0.8 Correlación fuerte
- 0.4-0.6 Correlación moderada
- 0.2-0.4 Correlación débil
- 0.0-0.2 Muy débil o sin correlación

$cov(X;Y)$: covarianza de X e Y.
$\sigma(X)$: desviación estándar de X.

Limitaciones de este método

El problema fundamental que presentan los métodos de filtrado basados en ranking es que pueden eliminar variables que por sí solas son malas, pero que resultan de gran utilidad al ser combinadas con otras.

A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

Ejemplo

04_Seleccion_Generacion__Seleccion.ipnyb
04_Seleccion_Generacion__Rankeo.R

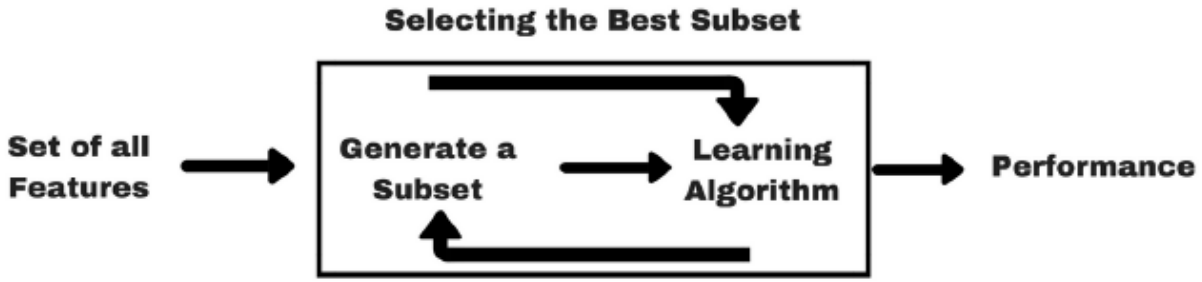
PYTHON

Es importante seleccionar subconjuntos de variables que, en conjunto, tienen un buen poder predictivo, en lugar de clasificar las variables según su poder predictivo individual.

Por ello, vamos a ver los métodos wrapper y embedded.

filter

wrapper



3. Métodos wrapper

Métodos wrapper

Los métodos *wrappers* utilizan un algoritmo de aprendizaje automática como caja negra para realizar la selección de variables:

1. Se considera una familia $\mathcal{F}$ de subconjuntos de predictores.
2. Para cada $S \in \mathcal{F}$, se entrena un modelo M_S de predicción (con el algoritmo fijado) utilizando S como conjunto de predictores.
3. Se elige el mejor conjunto de variables como aquel S que maximiza el rendimiento.

Los *métodos stepwise* son el caso más habitual de *wrappers* que se emplean en la práctica, aunque existen otros casos de uso frecuente, como *random-hill climbing algorithm* o *recursive feature elimination*.

Métodos wrapper

- ***Random-hill climbing algorithm***: modificación de los métodos *stepwise* "tradicionales": en cada paso, en lugar de determinar la mejor variables a añadir/eliminar, con una cierta probabilidad, la variable será determinada de forma aleatoria.

- ***Recursive feature elimination***: modificación de *backwards stepwise selection*: en cada paso, se ajusta un modelo, y se eliminará la variable o variables menos importantes para el modelo.

Ejemplo

04_Seleccion_Generacion__Seleccion.ipnyb
05_Seleccion_Generacion__Wrapper.R

PYTHON

4. Métodos embedded

Métodos embedded

Los métodos *embedded* son aquellos que realizan la selección de variables como parte del proceso de entrenamiento de algún modelo. Algunos ejemplos podrían ser:

- Árboles de decisión con poda.
- Regresión lasso:

Dado un problema de regresión, en lugar de encontrar β que minimice:

$$\sum_{i=1}^{n}(y_i - \beta^t x_i)^2 ,$$

Calcular el β que minimiza:

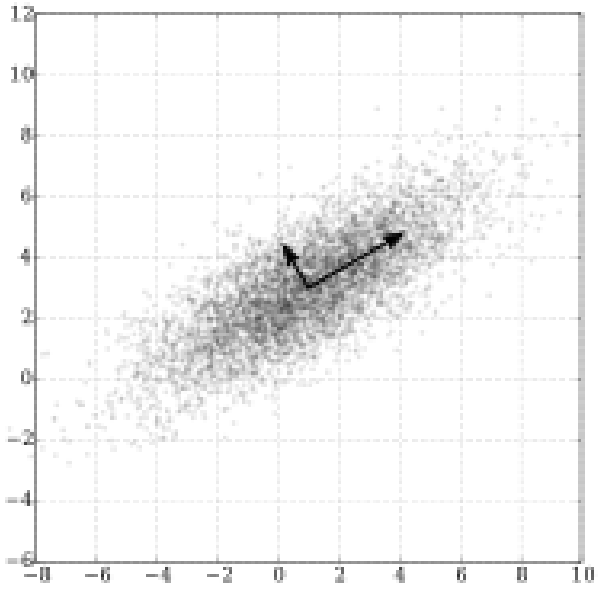
$$\sum_{i=1}^{n}(y_i - \beta^t x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

5. Otras posibilidades

Reducción de dimensionalidad

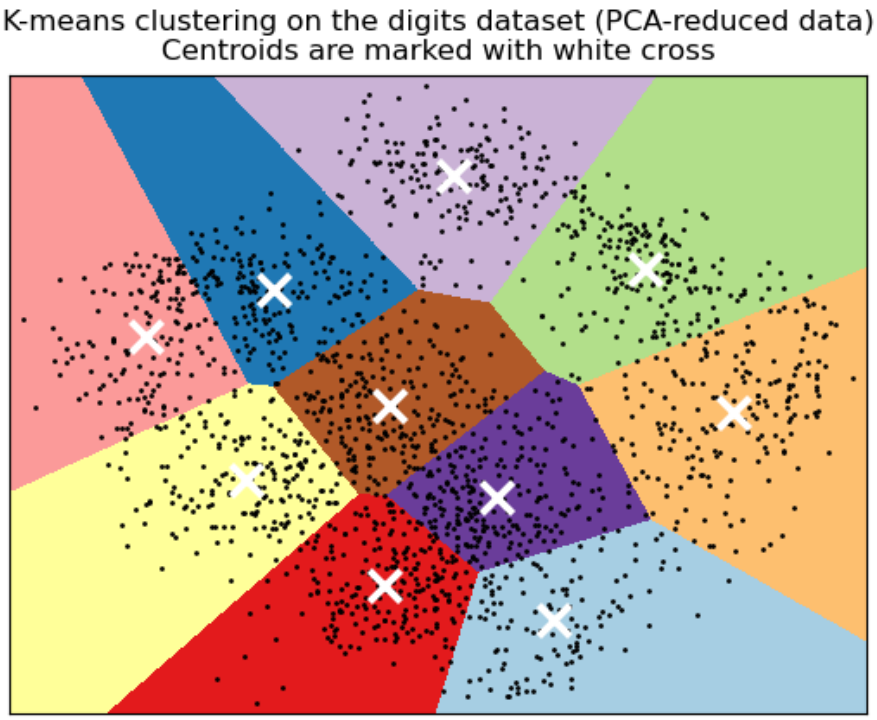
Cuando hay demasiadas variables o cuando el significado de las variables no es claro y las técnicas anteriores no dan buenos resultados, suelen emplearse técnicas de reducción de dimensiones para seleccionar variables.

La técnica más habitual en este sentido es el **análisis de componentes principales (PCA).**



Clustering

Otra posibilidad es aplicar **clustering** en el espacio de las variables y sustituir cada clúster de variables por un representante apropiado del clúster.



Evaluación de un método de selección

Para evaluar los beneficios de los métodos de selección de variables se debe realizar una comparación de resultados antes y después de realizar un filtrado de características, o comparar dos o más algoritmos de selección, para comprobar si uno es mejor que otro.

Para la selección final de las variables a tener en cuenta se deben contemplar los siguientes aspectos:

- Número de variables seleccionadas.
- Tiempo empleado.
- Rendimiento del modelo.

6. Paquetes interesantes

Paquetes interesantes

Algunos paquetes interesantes para la selección de variables:

- Selección de variables con caret.
- Selección de variables con scikit-learn.
- R FSelector package.
- IBM Python package: imdbpy.
- R leaps package for linear regression.

7. ¿Qué hacer cuando p es demasiado grande?

¿Qué hacer cuando p es demasiado grande?

- Hay que tratar de evitar usar todas las variables en el entrenamiento de un modelo (salvo que haya tiempo y potencia de cómputo ilimitados).
- Entrenar modelos con pocas variables para descartar las malas.
- Llegar a una hipótesis básica con pocas variables.
- Hacer *forward stepwise*.
- Remuestrear y repetir.

Referencias

Referencias

- Guyon, I., Elisseef, A., 2003. An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, pp. 1157-1182.
- Barzilay, O., Brailovsky, V.L., 1999. On domain knowledge and feature selection using a support vector machine.
- Groves, W., 2013. Using Domain Knowledge to Systematically Guide Feature Selection.
- Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection.
- Skalak, D., 1994. Prototype and Feature Selection by Random Mutation Hill Climbing Algorithms

Para ver:

- What makes a good feature? – Machine Learning Recipes #3
- Intro to feature engineering with TensorFlow – Machine Learning Recipes #9
- The 7 steps of machine learning
- Feature Selection and Data Visualization with Breast Cancer Wisconsin Datset

Afi Escuela