Diplomatura Universitaria en Ciencia de Datos

https://exa.unne.edu.ar/diplomatura/

Módulo 3. Análisis Exploratorio de Datos
Selección de variables

Equipo Docente:

Dra. Sonia I. Mariño

Lic. Lucia del Valle Ledezma

Lic. Rafael Perez

2024

Temas

- Selección de características
- Métodos:
  - Filtros,
  - Envoltorios o Wrappers,
  - Envolventes,
  - Combinados o Híbridos
- Programación Phyton

Selección de Características

Definición:

Proceso de identificar y seleccionar un subconjunto relevante de características (o variables) para el modelo de ML / CD.

Dado d (conjunto de variables), ¿cuál es el mejor subconjunto de variables de tamaño p?

¿Porqué seleccionar características ?,

remover características poco relevantes (ruido o “distractores”)

- Ahorro de tiempo y memoria.
- Reducción de la complejidad del modelo,
- Mejora la performance del clasificador, una representación más estable disminuye la posibilidad de sobreajuste (overfitting).
- Facilita la visualización y la comprensión de los datos

Selección de Características

Método de fuerza bruta (método óptimo)

Si se evalúa el criterio de optimalidad para todas las posibles combinaciones de *p variables* seleccionadas a partir de **d variables**, implica evaluar un número de combinación igual a:

$$n_p = \frac{d!}{(d-p)!p!}$$

Número muy elevado incluso para valores moderados de d y p.

Se carece de un criterio para seleccionar p,

- el número de posibles combinaciones que deberían ser evaluadas puede crecer exponencialmente.

Se definen métodos subóptimos de selección de variables, sus componentes son:

- un criterio de selección y
- una estrategia de búsqueda.

Métodos de selección

Basados en filtro

Basados en envoltorios (wrappers)

Basados en embedding

Enfoques combinados

Objetivos:

- mejorar la precisión del modelo,
- reducir la complejidad del modelo,
- reducir el sobreajuste,
- reducir tiempo de entrenamiento.

Métodos de selección

Métodos de Filtros (filtering)

- selecciona las características en forma independiente del clasificador. Usa criterio de “relevancia”
- simples, rápidos, no considera interacciones

Métodos de Envoltorios (Wrapper / Encapsulado / wrapping):

- alta precisión, costoso computacionalmente en tiempo
- selecciona los subconjuntos de características en función del desempeño de un clasificador.
- necesita estrategia de búsqueda para explorar en forma eficiente el espacio de subconjuntos.

Métodos Embedding (intrínseco)

- la selección en el proceso de aprendizaje devuelve un subconjunto de características y el clasificador entrenado.
- se aplican n entrenamientos, se evalúa el costo de agregar o quitar característica pero no se reentrenan,
- balance entre precisión y eficiencia, dependiente del algoritmo

Enfoques combinados,

- En la selección se aplican métodos de “filtros” y para refinar métodos de envoltorios o “wrapper”

Métodos de selección. Basados en Filtros

Métodos de filtros

Métodos que constan de una función objetivo para evaluar el subconjunto de características a partir de su contenido de información.

(+) Simples, rápida ejecución, no consideran interacciones

- se aplican cálculos generalmente no iterativos relacionados con el conjunto de datos

(+) Generalidad.

- evalúan las propiedades de los datos, más que las interacciones con un modelo de aprendizaje particular,
- resultados más generalidad. Es decir, la solución “buena” para una familia más grande de modelos.

(-) Seleccionar subconjuntos de características grandes.

- El filtro tiende a seleccionar el conjunto completo de variables como el mejor.

Alternativas de criterios de decisión (función objetivo):

- medidas de correlación
- medidas de teoría de la información

Métodos de selección. Basados en filtros

Coeficiente de correlación, mide el nivel de relación “lineal” entre dos variables.

Criterio basado en la suposición de que los subconjuntos de características óptimos, contienen características altamente correlacionadas con la variable de salida y no correlacionadas con las demás variables de entrada.

Donde:

ρ, coeficiente de correlación entre las variables indicadas por los subíndices,

c, la variable de salida (variable a predecir).

$$J = \frac{\sum_{i=1}^{p} \rho_{ic}}{\sum_{i=1}^{p} \sum_{j=i+1}^{p} \rho_{ij}}$$

Métodos de selección. Basados en filtros

Información mutua

medida más robusta, donde

I(Xm;c), información mutua entre el subconjunto de variables Xm y la variable de salida c

H(c), entropía de la variable de salida c

H(c|Xm). entropía condicional de c, siendo conocido Xm

Información mutua corresponde a la reducción en la incertidumbre de la variable c, dado que se conocen las variables incluidas en el subconjunto Xm

Entropía es un funcional, una función que tiene como entrada otra función, la cual corresponde a la distribución de probabilidad de la variable bajo análisis.

La implementación de la Información mutua, depende del tipo de función de distribución que se asuma para cada una de las variables.

$$J = I(X_m; c) = H(c) - H(c|X_m)$$

Métodos de selección. Basados en filtros

Chi cuadrado

- Prueba de chi-cuadrado mide la dependencia entre variables estocásticas.
- Es decir, la función "elimina" las características que tienen más probabilidades de ser independientes de la clase e irrelevantes para la clasificación.
- Esta puntuación se puede utilizar para seleccionar las características n_features con los valores más altos para la prueba estadística de chi-cuadrado de X.
    - Debe contener solo características no negativas como valores booleanos o frecuencias en relación con las clases.
- En Phyton

Calcula la prueba chi-cuadrado entre cada característica y la clase no negativa.

scikit-learn - sklearn.feature_selection.chi2()

Métodos de selección. Envoltorios

Basados en Wrapper (Envoltorios / Encapsulado / wrapping):

Se define la **función objetivo** como un **modelo de aprendizaje**, en principio encuentra el subconjunto más “útil”.

Se evalúa y determina el subconjunto de características

- (+) Exactitud.
  - generalmente alcanzan mejores tasas de predicción que los filtros, se ajustan para reducir el error de validación.
- (+) Capacidad predictiva
  - usan una metodología de validación, tienen la capacidad de evitar el sobreajuste y proporcionar mejor capacidad de generalización.

Desventajas

- (-) Ejecución lenta.
  - alto costo computacional, debe entrenar un clasificador por cada subconjunto de variables (o varios si se usa validación cruzada),
- (-) Especificidad de la solución
  - subconjunto de variables seleccionadas puede ser especifico para el modelo de predicción elegido en el criterio wrapper. Por ello, puede no ser buen criterio para otros modelos.

Métodos de selección. Envoltorios

Basados en Wrapper (Envoltorios / Encapsulado / wrapping):

- Métodos evalúan la combinación de características y seleccionan el mejor subconjunto.

Ejemplos:

RFE (Recursive Feature Elimination):

- Librería scikit-learn, RecursiveFeatureSelector,
- **Elimina** recursivamente las N características menos importantes o que dan peores resultados

SFS (Sequential Feature Selector). **Selecciona** secuencialmente las características

- Librería scikit-learn, SequentialFeatureSelector,
  - requiere un estimador, ejemplo K-Nearest Neighbors o KNN
- Forward Selection (selección hacia adelante)
- Backward Selection (selección hacia atrás)

Métodos de selección. Envoltorios

Basados en Wrapper (Envoltorios / Encapsulado / wrapping):

SFS (**Sequential Feature Selector). Selecciona secuencialmente las características, hacia adelante o hacia atrás**

- Librería scikit-learn, **SequentialFeatureSelector**
- **Forward Selection (selección hacia adelante)**
    - se inicia con 0 características,
    - se agregan –iterativamente- la característica del conjunto hasta contar con el número deseado de características seleccionadas.

Métodos de selección. Envoltorios

Basados en Wrapper (Envoltorios / Encapsulado / wrapping):

SFS (**Sequential Feature Selector). Selecciona secuencialmente las características, hacia adelante o hacia atrás**

- Librería scikit-learn, **SequentialFeatureSelector**
- **Backward Selection (selección hacia atrás)**
    - se inicia con todas las características,
    - se elimina –iterativamente- una característica del conjunto hasta disponer del número deseado de características seleccionadas.

Métodos de filtro vs Métodos envoltorios

	Métodos basados en filtro	Métodos envoltorios
Modelo de Machine Learning		X
Rapidez	X (no require entrenar el modelo)	
Computacionalmente costosos		X
Localizar el major subconjunto		X (naturaleza exhaustiva)
Sobreajuste (overfitting)		X (la selección de características require entrenar el modelo)

Métodos de selección. Embedded

La determinación de las características se guía por el proceso de aprendizaje, Se establecer una medida de utilidad de subconjunto de características.

En el proceso de evaluación, se aplica validación cruzada.

Resultados.

- similares a métodos Wrappers, menos costosos y menor tendencia a generar overfitting

Ej. métodos de regresión LASSO y RIDGE, incorporan funciones de penalización para reducir el sobreajuste.

Sugerencias

- Elaborar el análisis exploratorio de datos
- Experimentar con distintos métodos y comparar resultados
- Documentar el proceso y resultados

- Separar Train y Test en una proporción de 80/20 o 70/30
- Aplicar Cross Validation
  - Preferencia de usar Stratified-K-folds.
  - Número de “folds” depende del tamaño del dataset definido. En general el número es 5 (similar al 80-20 o 70/30).
  - En problemas de series de tiempo, usar TimeSeriesSplit
- Si la métrica Accuracy (u otra) es similar en los conjuntos de Train (Cross Validation) y Test, se acepta como bueno el modelo

Scikit-learn

Algunas técnicas disponibles

- SelectKBest:
  - Selecciona las características según una función estadística de puntuación, como ANOVA. Permite especificar el número de características amantener. Funciones de puntuación: f_classif o f-test de ANOVA, para problemas de clasificación. chi2 para datos categóricos, o mutual_info_classif para la información mutua.
- Recursive Feature Elimination (RFE).
  - Elimina recursivamente las características menos importantes en función de un modelo específico. Ajusta el modelo en iteraciones para seleccionar las mejores características.
- SelectFromModel.
  - Selecciona características basadas en la importancia calculada por un modelo, como RandomForest o un modelo lineal con regularización.
- VarianceThreshold.
  - Elimina características que tienen una varianza por debajo de un umbral dado. Es decir, útil para eliminar características que no varían mucho y aportan poca información.

Documentación:

- 1.13. Feature selection — scikit-learn 1.5.2 documentation

Featurewiz

Biblioteca de Python utilizada para la selección automática de características en modelos de ML,

Métodos:

- Análisis de importancia de variables basado en modelos.
    - Utiliza modelos como Random Forests y XGBoost para calcular la importancia de cada característica. Modelos de árboles de decisión pueden capturar relaciones no lineales y son robustos frente a datos ruidosos.
- Eliminación de características redundantes:
    - Identifica y elimina características redundantes, es decir, las altamente correlacionadas con otras características.

Documentación:

- https://github.com/AutoViML/featurewiz#featurewiz