Machine Learning 101

Selección de características

Felipe Alonso Atienza

Data Scientist @BBVA

Índice

1. **Motivación**
2. Métodos de filtrado
3. Métodos *wrapper*
4. Métodos *embedded*

Motivación selección características

1. Interpretabilidad
    - Eliminar variables irrelevantes
    - Entender mejor los datos

2. Reducir coste computacional del entrenamiento
    - Se busca solución dispersa

3. Evitar sobreajuste
    - Reducir la dimensionalidad del conjunto de entrada

Taxonomía

1. Métodos de filtrado
2. Métodos *wrapper*
3. Métodos *embedded*

Índice

1. Motivación
2. **Métodos de filtrado**
3. Métodos *wrapper*
4. Métodos *embedded*

Métodos de filtrado

- Se evalúa la **relevancia** de cada característica de forma individual
- Las variables se **ordenan de acuerdo a algún índice de relevancia**, de tal forma que las variables con valor más bajo son eliminadas
- Con el conjunto de variables seleccionadas entrenamos nuestro modelo de ML
  - Ejemplo (no de selección de características, pero sí de filtrado). Eliminar las variables con alta correlación.
- En scikit-learn se denomina Univariate feature selection.

Ventajas e inconvenientes

- Ventajas
  - Sencillos y rápidos de aplicar

- Desventajas:
  - No tienen en cuenta interacciones entre variables

- For regression: `f_regression`, `mutual_info_regression`
- For classification: `chi2`, `f_classif`, `mutual_info_classif`

The methods based on F-test estimate the degree of linear dependency between two random variables. On the other hand, mutual information methods can capture any kind of statistical dependency, but being nonparametric, they require more samples for accurate estimation.

Hora de practicar

Índice

1. Motivación
2. Métodos de filtrado
3. **Métodos *wrapper***
4. Métodos *embedded*

Métodos *wrapper*

- Se utiliza un algoritmo de ML como caja negra, para evaluar las prestaciones de distintos conjuntos de características

- Necesitan:
  - Un algoritmo de ML
  - Un criterio de relevancia
  - Un procedimiento de búsqueda de todos los posible subconjuntos de características (normalmente métodos heurísticos)

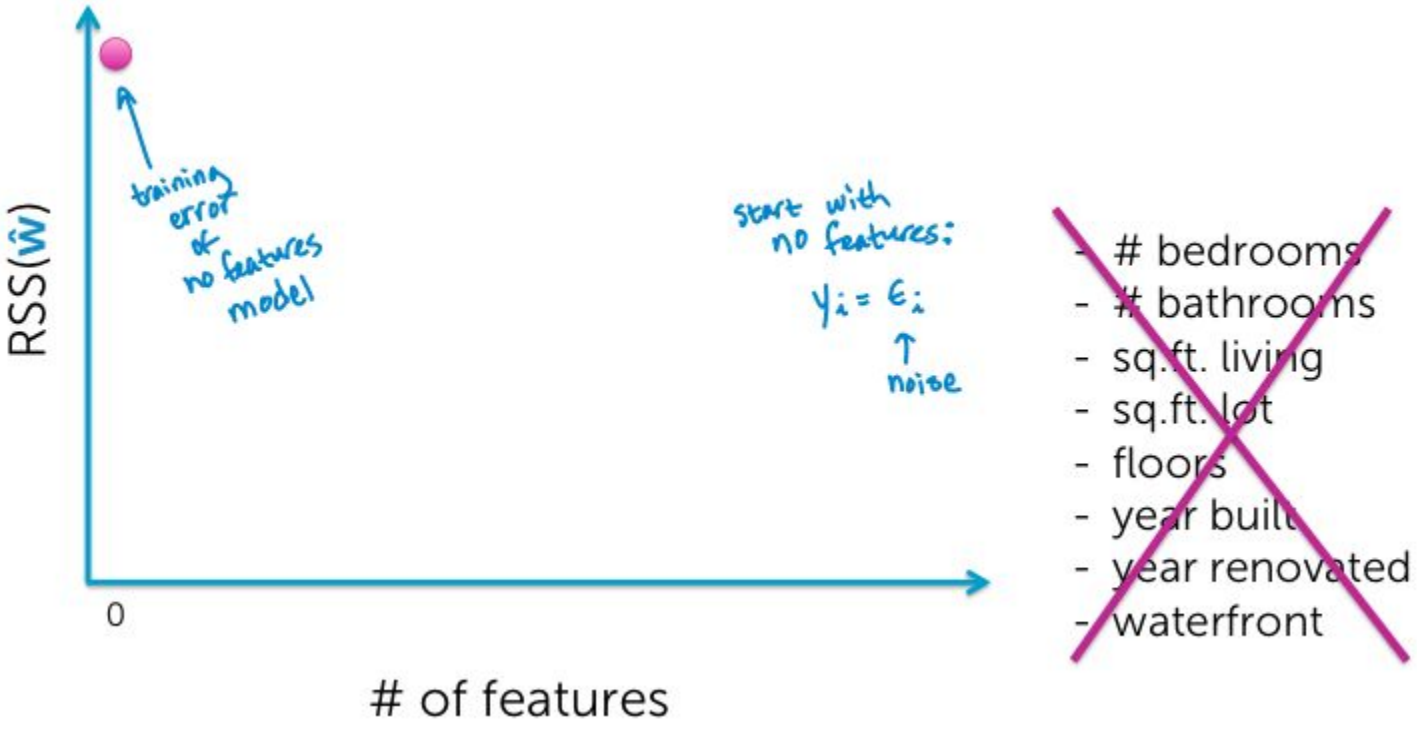
- Procedimientos de búsqueda
  - **Fuerza bruta**
  - Aleatorios: algoritmos genéticos, *simulating annealing*.
  - Estrategias greedy: selección **hacia delante** o **hacia atrás**.

Fuerza bruta

Lot size	Dishwasher
Single Family	Garbage disposal
Year built	Microwave
Last sold price	Range / Oven
Last sale price/sqft	Refrigerator
Finished sqft	Washer
Unfinished sqft	Dryer
Finished basement sqft	Laundry location
# floors	Heating type
Flooring types	Jetted Tub
Parking type	Deck
Parking amount	Fenced Yard
Cooling	Lawn
Heating	Garden
Exterior materials	Sprinkler System
Roof type	⋮
Structure style	

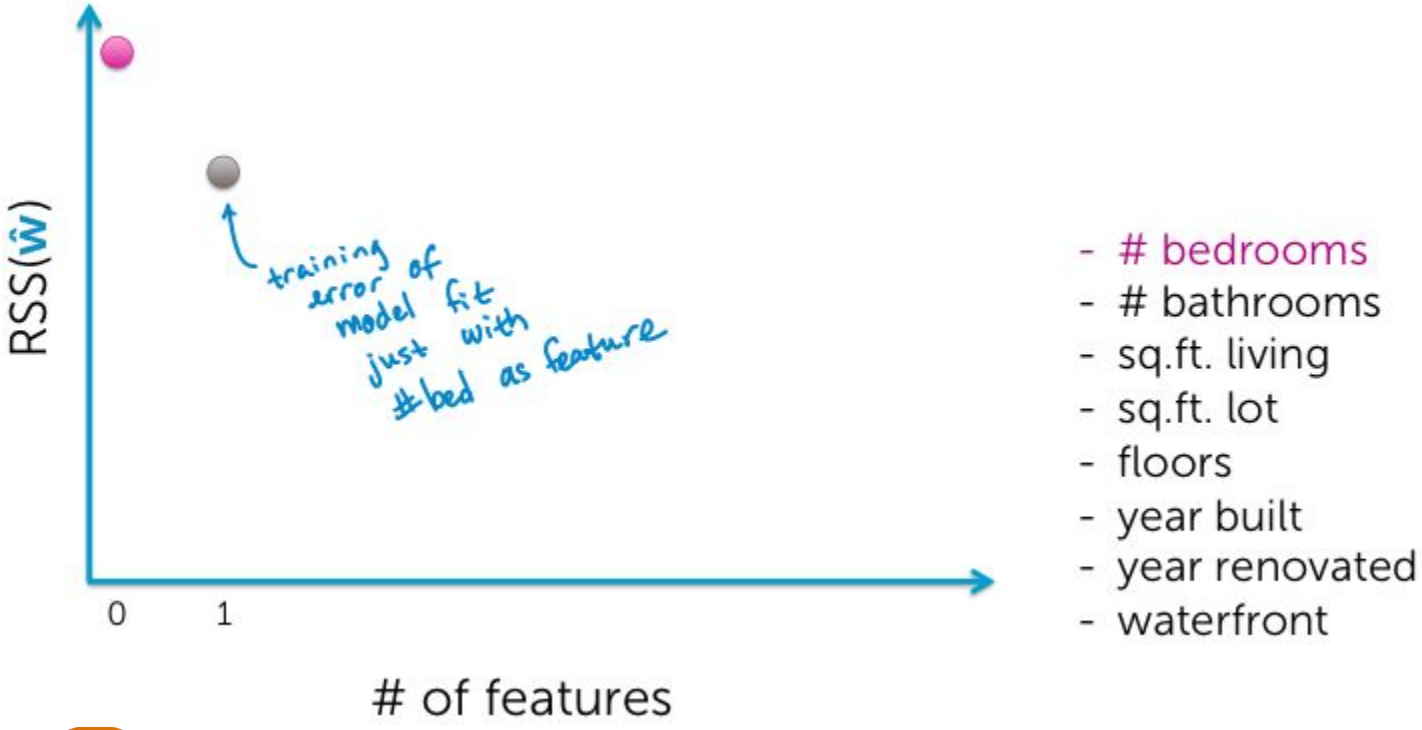
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



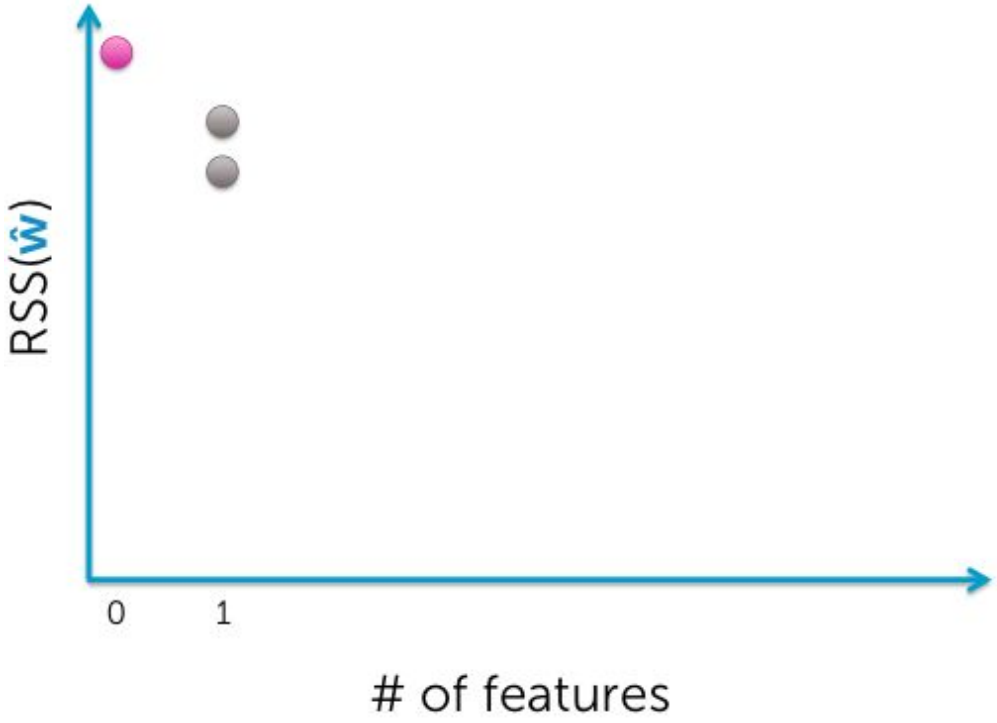
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



Fuente: Machine learning course. Emily Fox and Carlos Guestrin

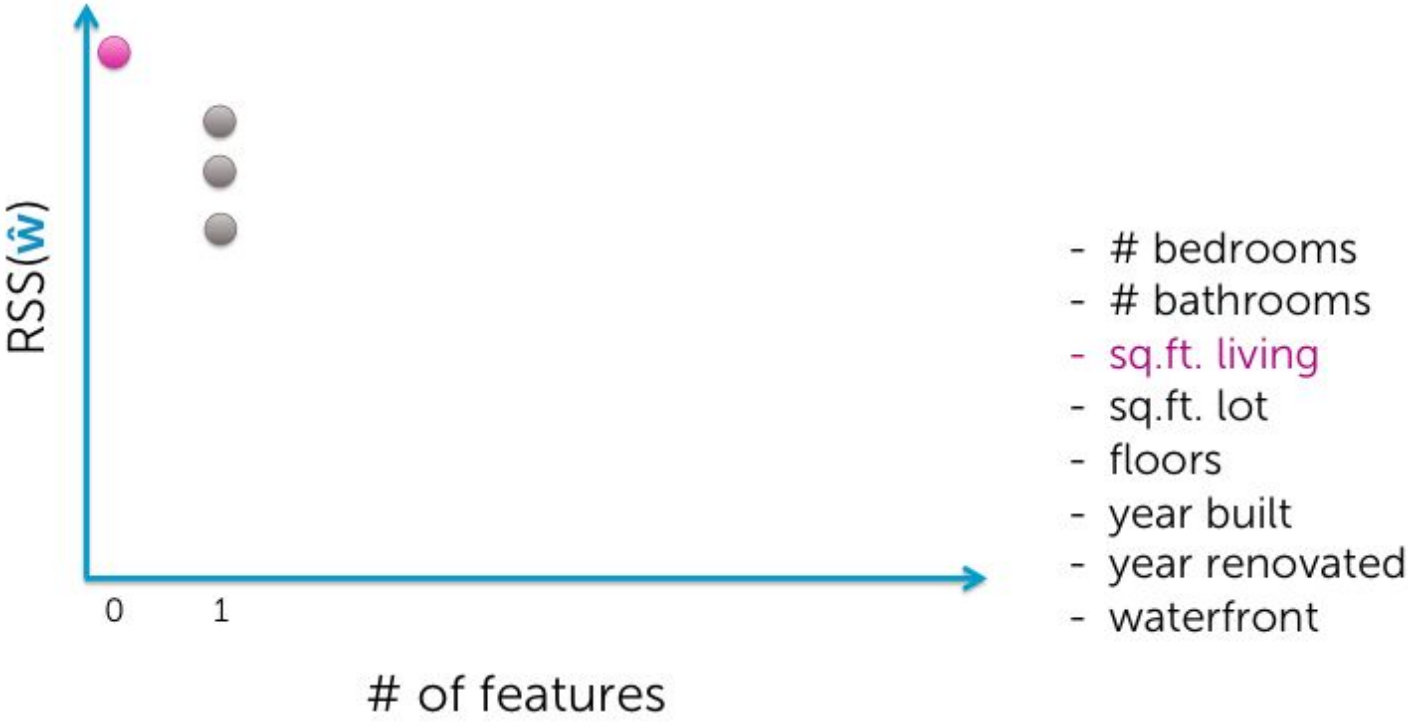
Fuerza bruta



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

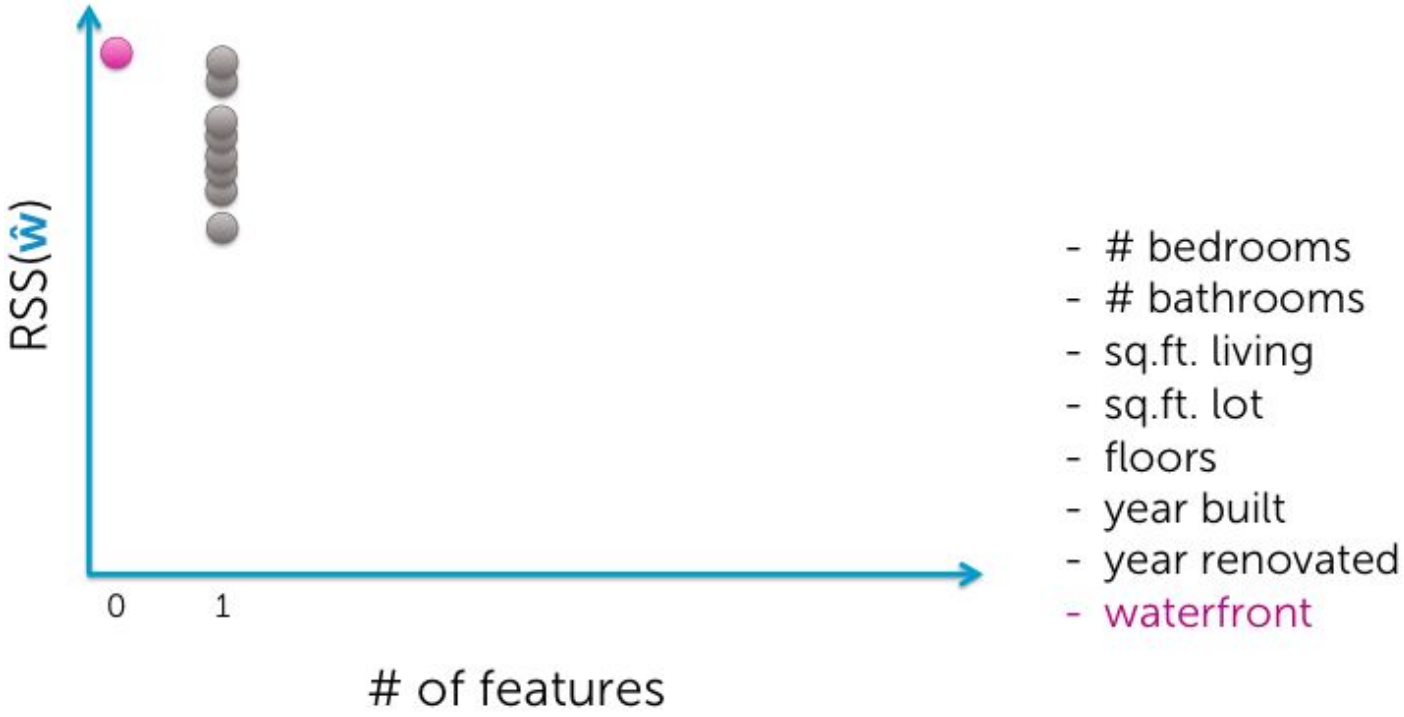
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



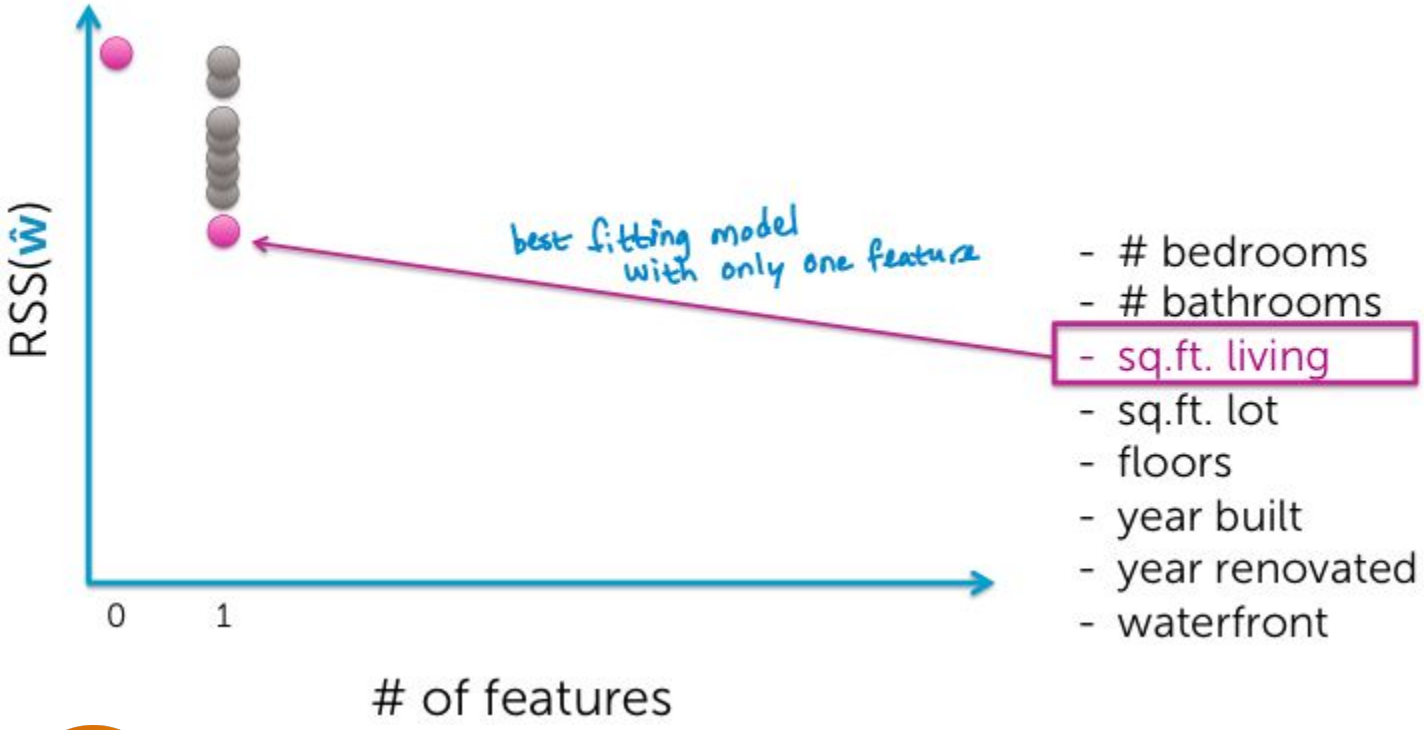
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



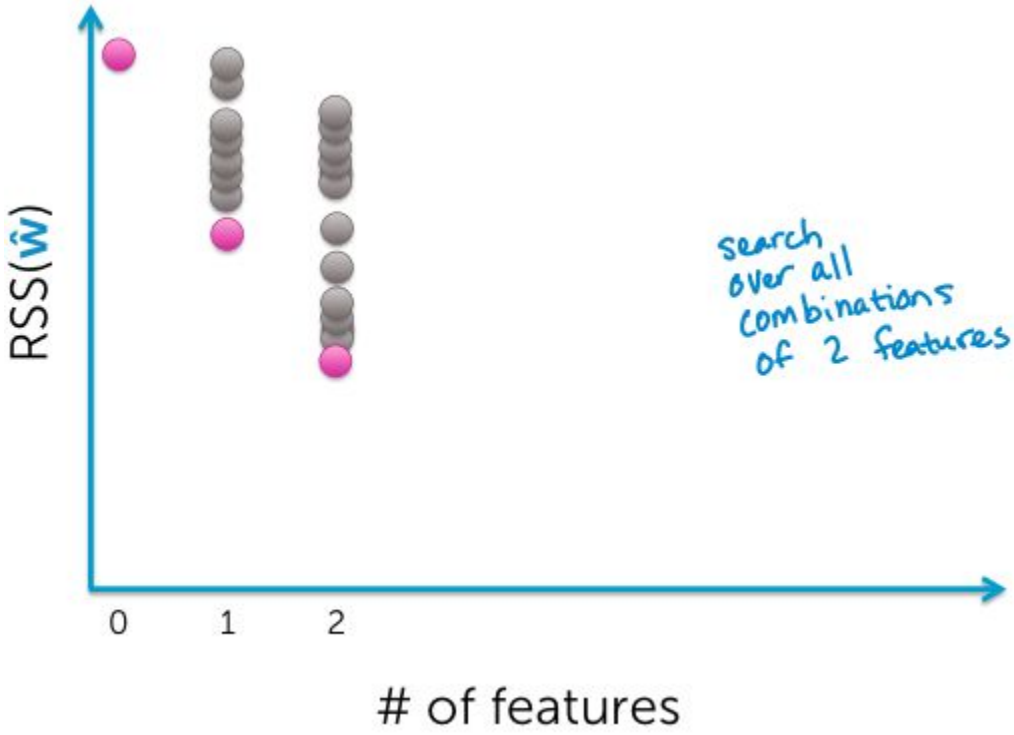
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



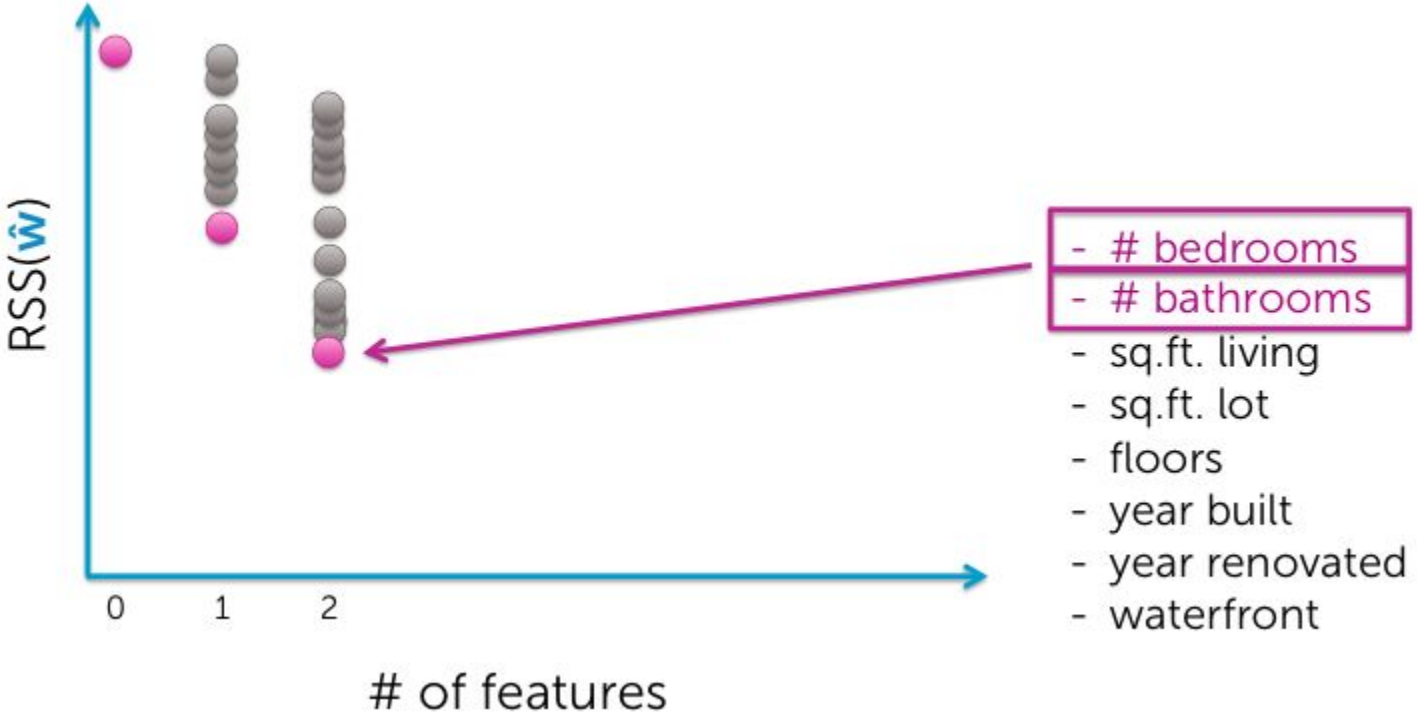
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



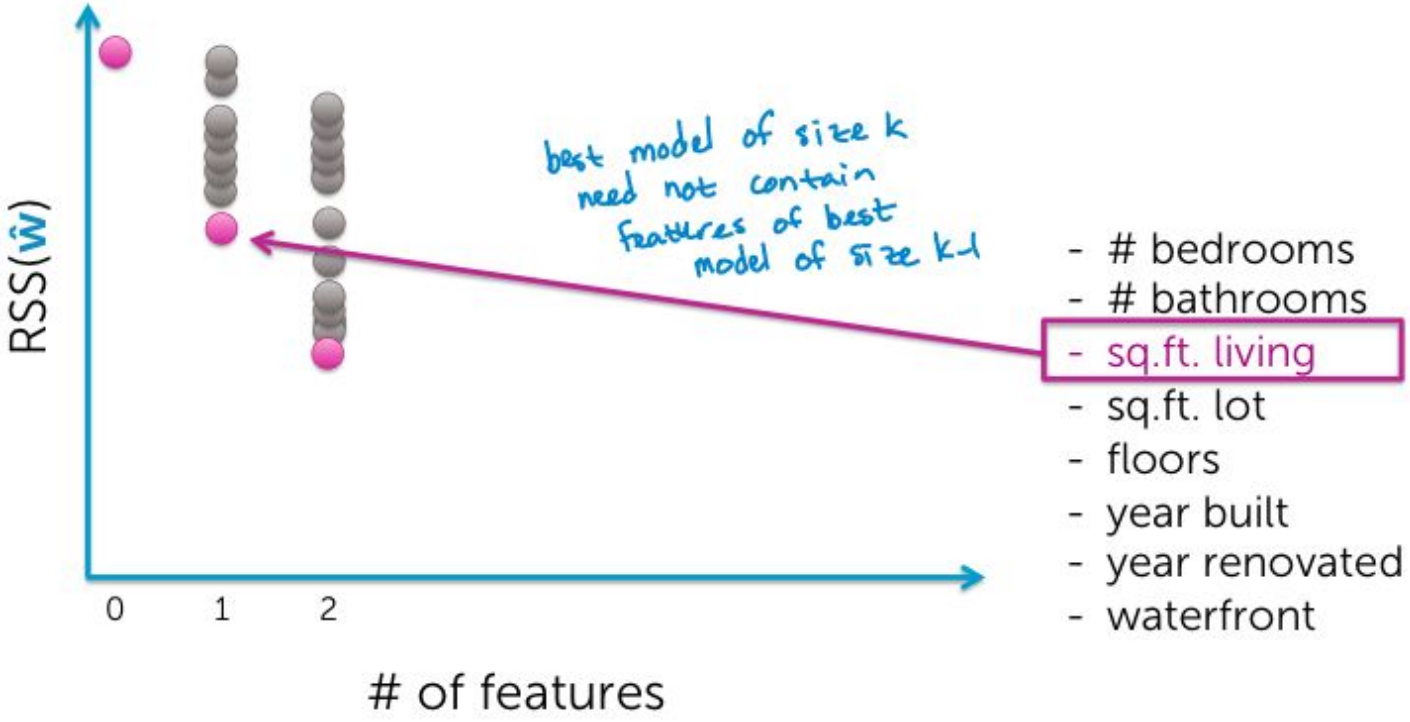
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



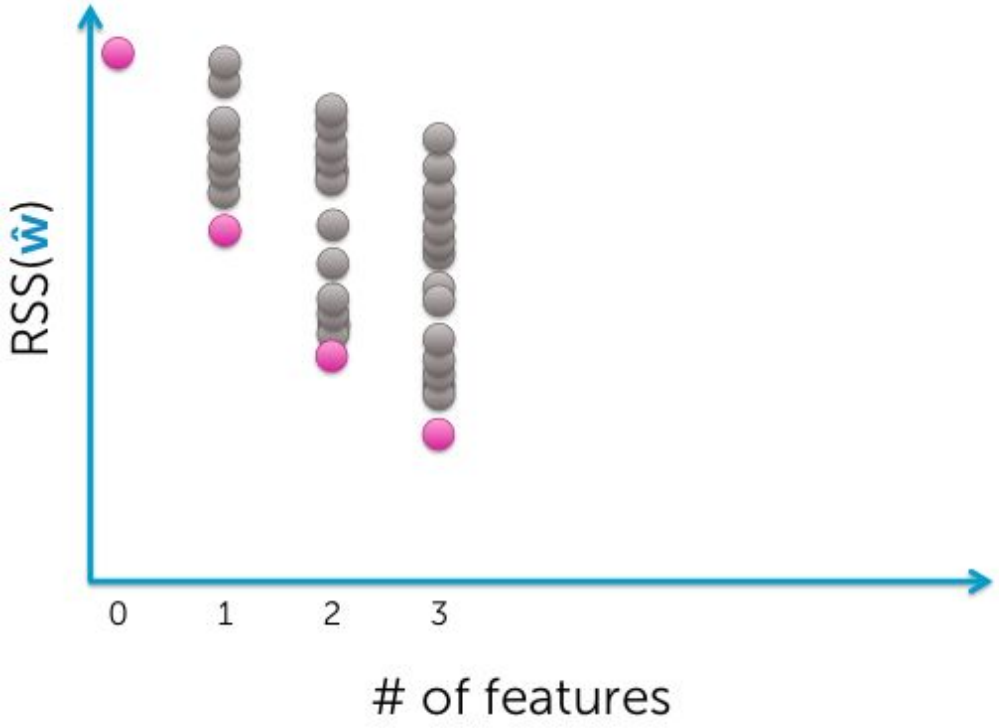
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Fuerza bruta



Fuente: Machine learning course. Emily Fox and Carlos Guestrin

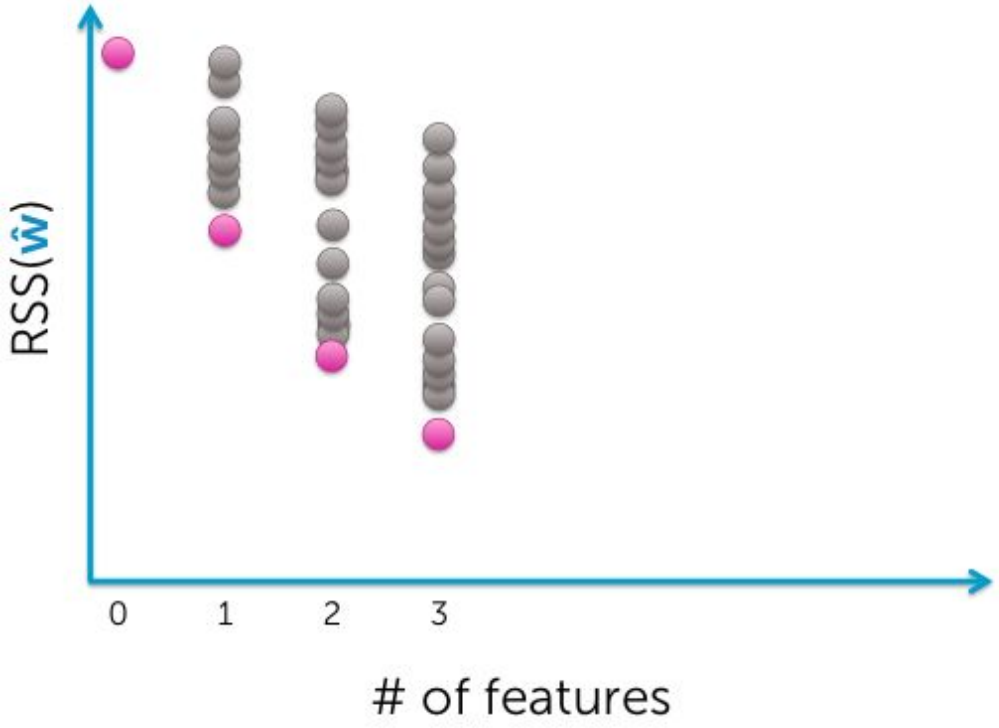
Fuerza bruta



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

Fuente: Machine learning course. Emily Fox and Carlos Guestrin

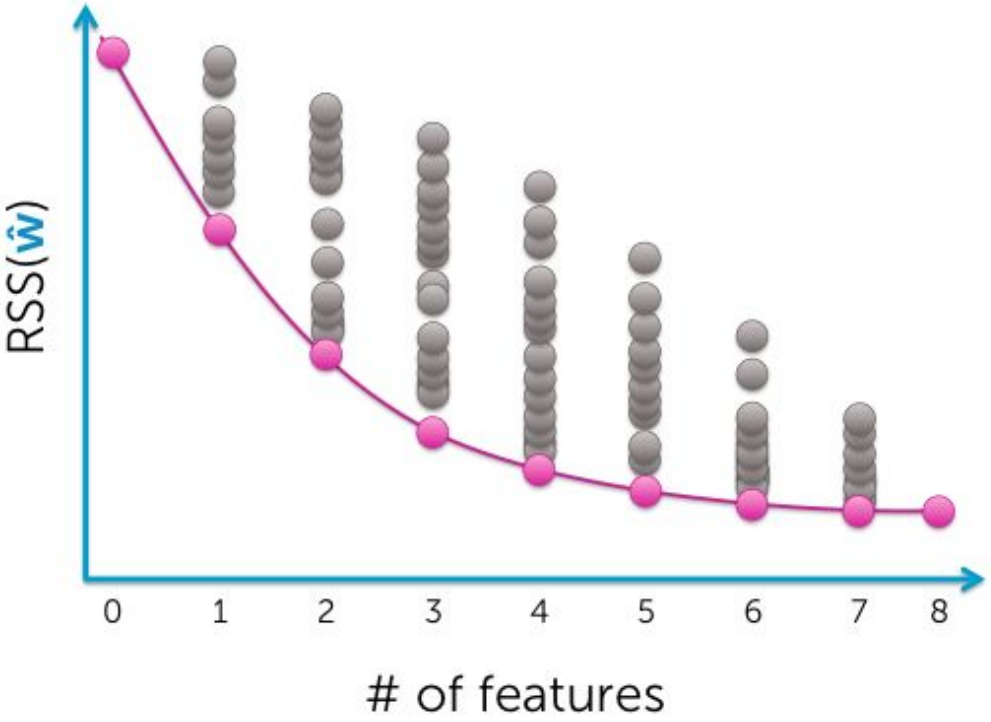
Fuerza bruta



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

Fuente: Machine learning course. Emily Fox and Carlos Guestrin

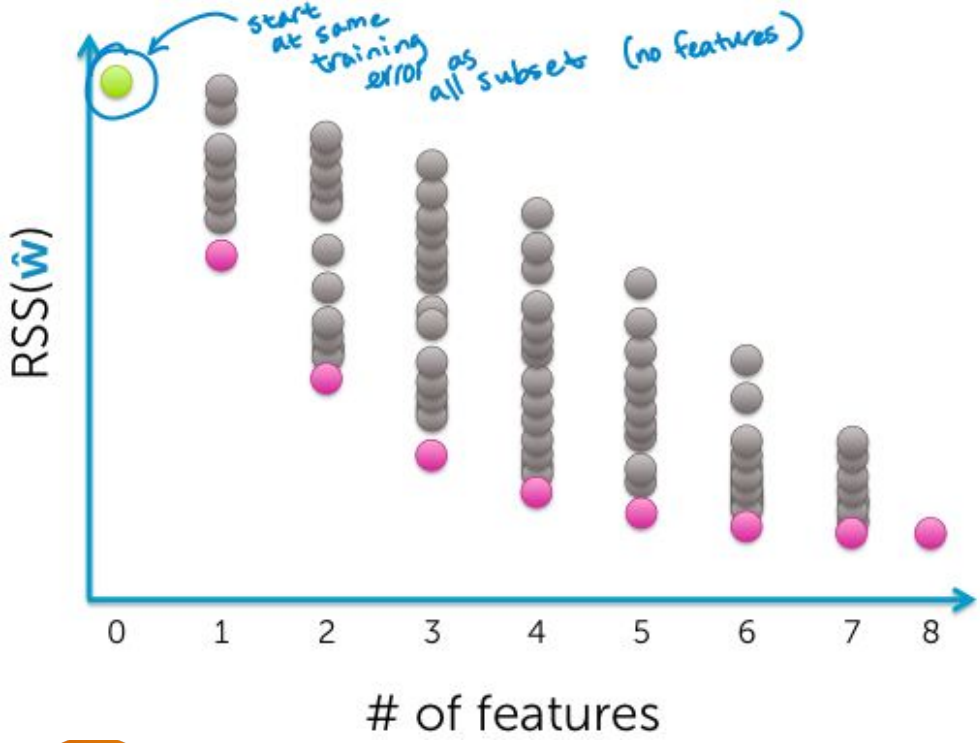
Fuerza bruta



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

Fuente: Machine learning course. Emily Fox and Carlos Guestrin

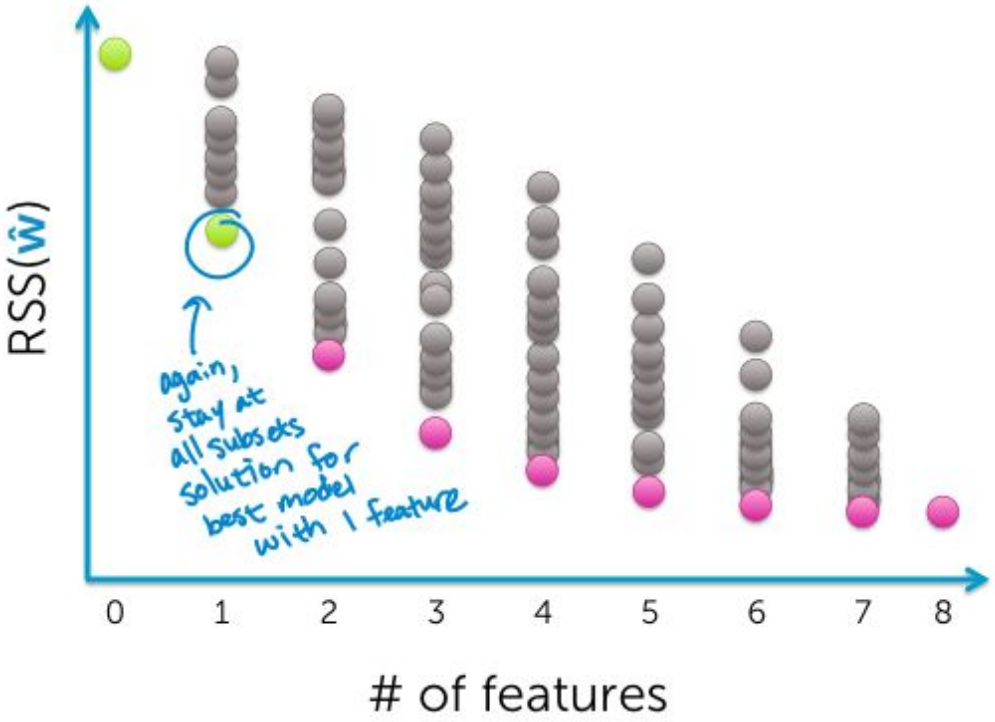
Selección hacia delante



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

Fuente: Machine learning course. Emily Fox and Carlos Guestrin

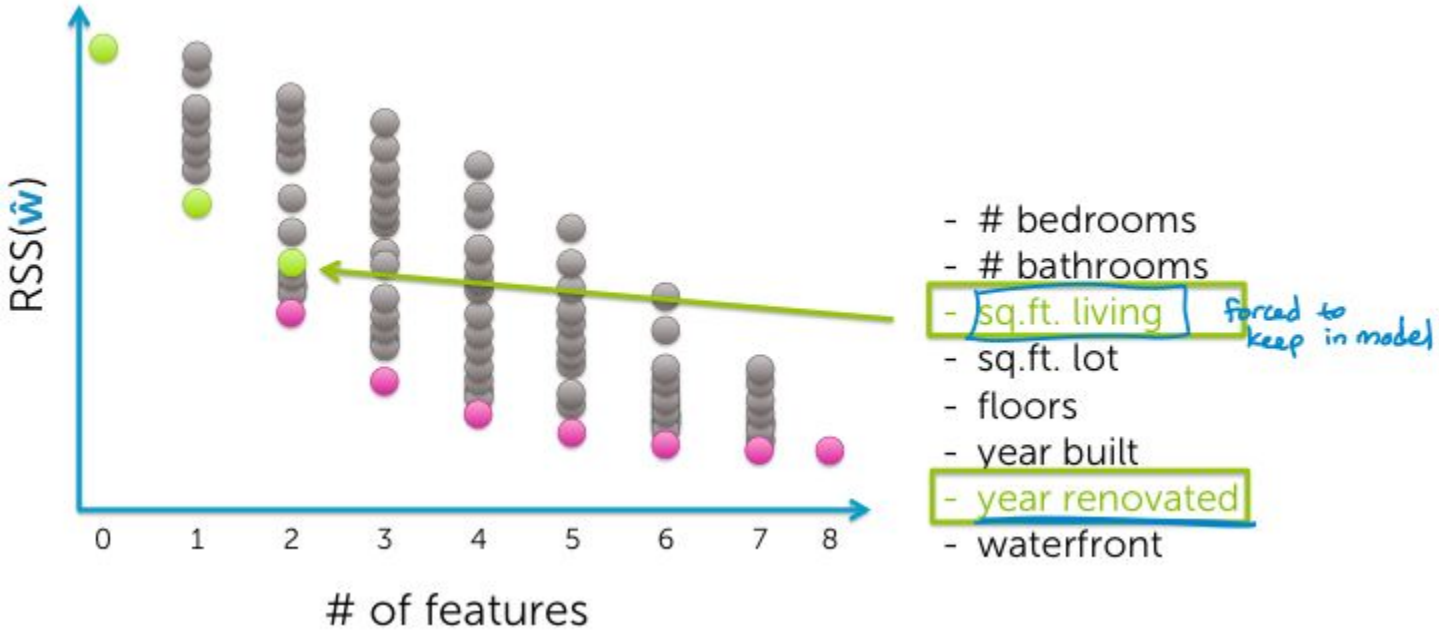
Selección hacia delante



- # bedrooms
- # bathrooms
- sq.ft. living
- sq.ft. lot
- floors
- year built
- year renovated
- waterfront

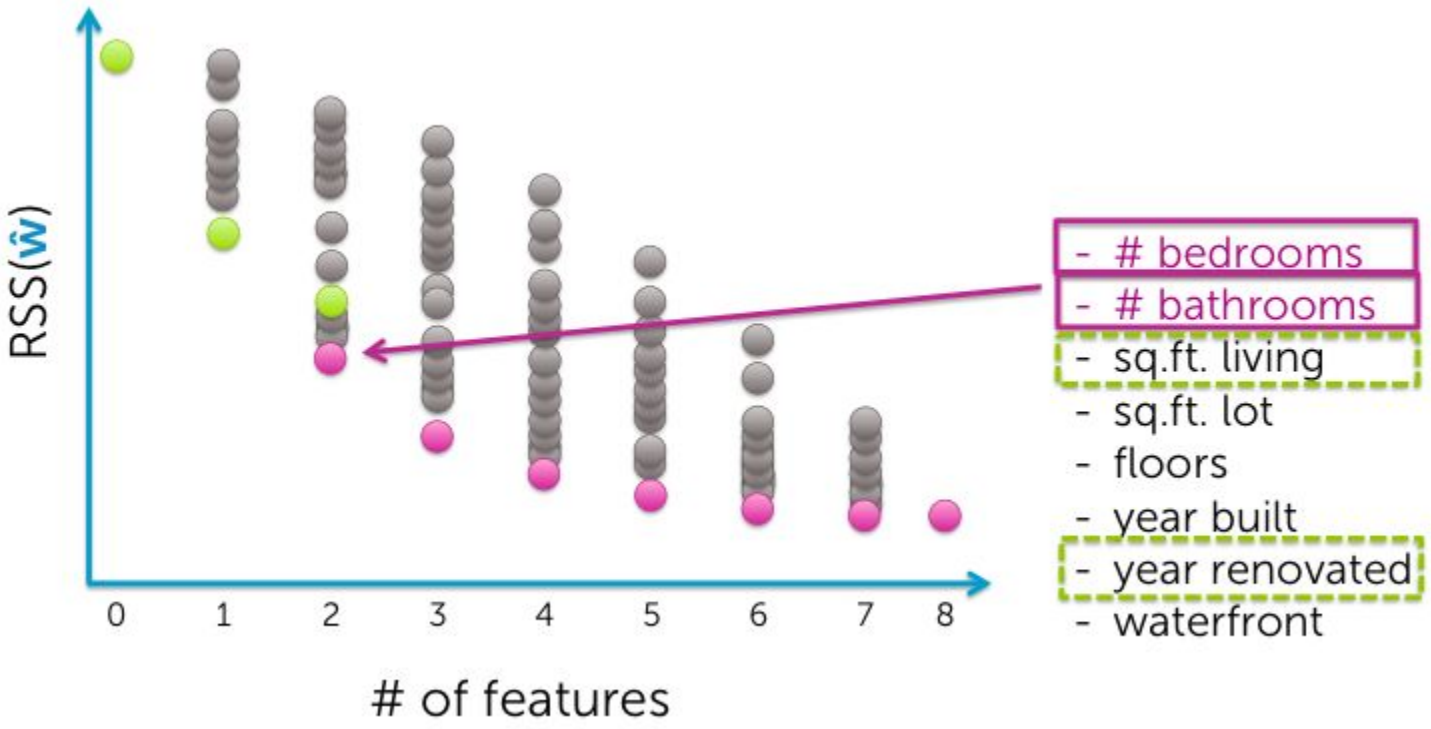
Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Selección hacia delante



Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Selección hacia delante



Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Selección hacia delante

Fuente: Machine learning course. Emily Fox and Carlos Guestrin

Métodos *wrapper*

- Complejidad
  - Fuerza bruta: $O(2^D)$
    - Si $D = 20 \Rightarrow 2^{20} \approx 1e6$ posibilidades
  - Complejidad selección hacia delante: $O(D^2)$

- OK: algoritmo de ML como caja negra, solución universal y sencilla.

- KO: para cada subconjunto se tiene que crear un nuevo modelo (entrenamiento/validación)
  - k-fold CV (hay que hacerlo bien!!)

Índice

1. Motivación
2. Métodos de filtrado
3. Métodos *wrapper*
4. **Métodos *embedded***

Métodos *embedded*

1. Incorporar la selección de características como parte del proceso de entrenamiento:
    - Utilizar algoritmos adecuados que permitan seleccionar características

2. Se pueden utilizar junto con técnicas hacia delante/atrás (métodos anidados): eficientes.

3. Ejemplos
    - Lasso
    - Árboles de decisión/regresión

Lo que puedes hacer ahora ...

- Técnicas de filtrado: filtrar (eliminar) las características poco relevantes en función de distintos criterios
- Técnicas wrapper: utilizar estrategias greedy para selección de características
- Técnicas embedded: utilizar las particularidades de un algoritmo para seleccionar las características relevantes

Referencias

- An Introduction to Statistical Learning. Capítulos 3, 6.

Hora de practicar