Recap

13 May 2023 07:05

Feature selection

n cols → p cols

$\boxed{100}$ $\boxed{10 \text{ cols}}$

$p < n$

1) Curse of dimension
2) complexity
3) Interpret → predict, inference

f_1 f_2 y

lat long

1 feature → 2 feature

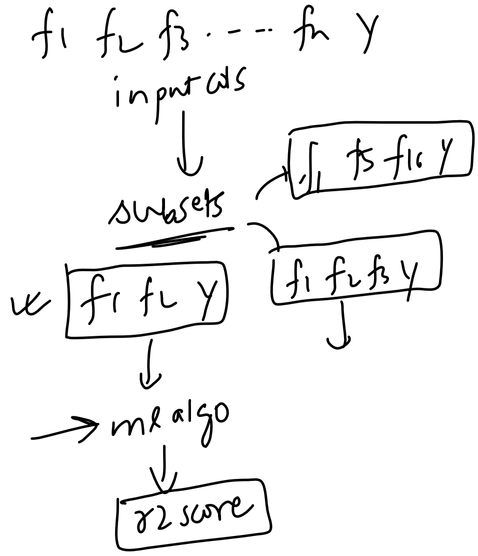
feature selection

- filter methods
  - variance
  - corr
  - chisquare
  - ANOVA
- Wrapper method
- Embedded methods

Wrapper Methods

13 May 2023 07:05

Wrapper methods for feature selection are a type of feature selection methods that involve using a predictive model to score the combination of features. They are called "wrapper" methods because they "wrap" this type of model-based evaluation around the feature selection process.

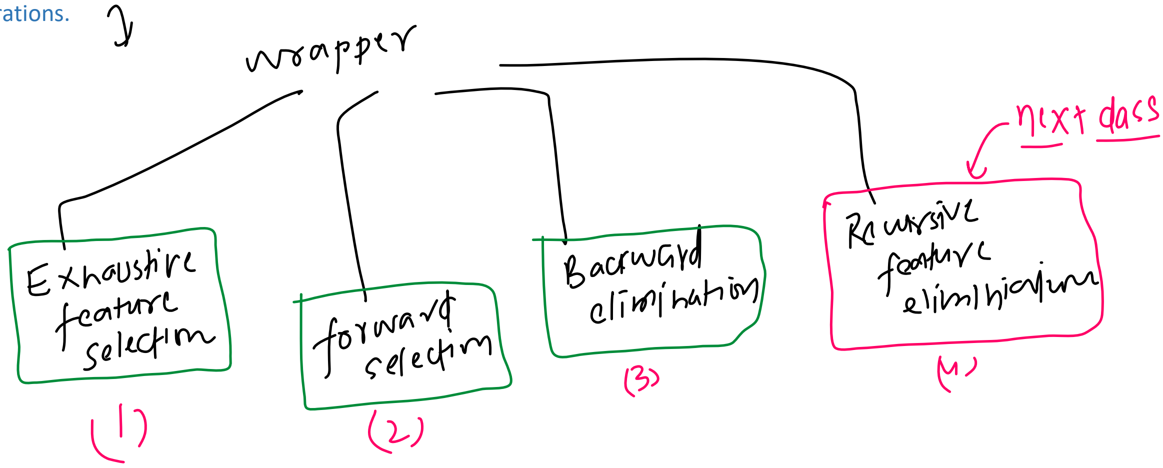


Here's how wrapper methods work in general:

1. Subset Generation: First, a subset of features is generated. This can be done in a variety of ways. For example, you might start with one feature and gradually add more, or start with all features and gradually remove them, or generate subsets of features randomly. The subset generation method depends on the specific type of wrapper method being used.

2. Subset Evaluation: After a subset of features has been generated, a model is trained on this subset of features, and the model's performance is evaluated, usually through cross-validation. The performance of the model gives an estimate of the quality of the features in the subset.

3. Stopping Criterion: This process is repeated, generating and evaluating different subsets of features, until some stopping criterion is met. This could be a certain number of subsets evaluated, a certain amount of time elapsed, or no improvement in model performance after a certain number of iterations.



Exhaustive Feature Selection/Best Subset Selection

13 May 2023 07:06

filter

f1 f2 f3 y

5 f1 f2 f3 f4 f5 y $\boxed{2^n} - 1$

try out all the subsets

f1 f2 y | f1 y | f2 y | f1 f2 y

ml 0.62 | ml 0.51 | ml 0.72

f1 f2 f3

1 f1 y 2 f2 y 3 f3 y

4 f1 f2 y 5 f2 f3 y 6 f1 f3 y 7

f1 f2

n → cols → 7

15

f1 f2 f3 y

$2 \to 3$

$3 \to 7$

$4 \to 15$

$\boxed{2^{563} - 1}$ ← $\boxed{n \to ?}$ $\boxed{2^n - 1}$ n features $\boxed{f = 563}$

10 → 8 cols

15 20

Disadvantage

1. Computational Complexity: The biggest drawback is its computational cost. If you have n features, the number of combinations to check is 2^n. So, as the number of features grows, the number of combinations grows exponentially, making this method computationally expensive and time-consuming. For datasets with a large number of features, it may not be practical.

   → 10 cols

2. Risk of Overfitting: By checking all possible combinations of features, there's a risk of overfitting the model to the training data. The feature combination that performs best on the training data may not necessarily perform well on unseen data.

   → overfitting

3. Requires a Good Evaluation Metric: The effectiveness of exhaustive feature selection depends on the quality of the evaluation metric used to assess the goodness of a feature subset. If a poor metric is used, the feature selection may not yield optimal results.

adjusted r2 penalise

→ $\boxed{r2}$ → add feature

f1 f2 f3 f4 y

f1 f2 y → LR → train test split

r2 score with high cols

adjusted r2

baseline model train

1 before applying

avg score

13 cols

before applying
any fs

LR → r2
score

r2 → 0.65

— | | — → 13 cols

t | train → 10 cols → best subset

EFS → f1 f2 f3

✓ $\boxed{f1 \quad f2}$ → 97

f1 f2 f3 → 97

Sequential Backward Selection/Elimination

13 May 2023 07:08

features

1. f1 f2 f3 f4 → model → 0.89

2.
- f2 f3 f4 → model → 0.81
- f1 f3 f4 → m → 0.71
- f1 f2 f4 → m → 0.91
- f1 f2 f3 → m → 0.65

3.
- f1 f4 → 0.79
- f1 f4 → 0.81
- f1 f2 → 0.83

f3 f4 → best

4.
- f2 → 0.63
- f1 → 0.53

f1 f2 f3 f4	0.89
f1 f2 f4	0.91
f1 f2	0.83
f2	0.63

f1 f2 f4 → best feature subset

model train → result

faster →

$n(n+1)$ → $1+2+3+\dots n$

Ex → 2^n-1

$\frac{13 \times 14}{2} = 13 \times 7$

81

- 4 → 4
- 3 → 3
- 2 → 2
- 1 → 1

$n \to n$

$n-1$

$n-2$

⋮

1

13 cols

subset per itera

$1+2+3+\dots+n = \boxed{\frac{n(n+1)}{2}}$

$\boxed{100}$ cols

↓

$2^{100}-1$

n^2 →

$\frac{100 \times 101}{2} = 50 \times 101$ $\boxed{5050}$

disadvantage

Exhaust $\boxed{\text{all}}$

iteration → best → local selection

miss the best

Sequential Forward Selection

13 May 2023 07:07

$\boxed{1 \to 2 \to 3 \to 4} \to$

$1 \to 4$
$2 \to 3$
$3 \to 2$
$4 \to 1$

$1 \to n$
$2 \to n-1$
$3 \to n-2$

$\boxed{\frac{n(n+1)}{2}}$

$\boxed{f_1}$ $\boxed{f_2}$ $\boxed{f_3}$ $\boxed{f_4}$ $\to 0 \to$ model $\to$ y mean

add

local best

$\boxed{f_1} \to 0.63$
$\boxed{f_2} \to 0.51$
$\boxed{f_3} \to 0.43$
$\boxed{f_4} \to 0.49$

tol $\to \boxed{0.5}$

$\boxed{f_1}$ $\boxed{f_2} \to 0.63$
$\boxed{f_1}$ $\boxed{f_3} \to 0.71$
$\boxed{f_1}$ $\boxed{f_4} \to 0.80$

$\to f_2 f_4$
$f_1 \to 0.63$
$f_1 f_4 \to 0.80$
$\to f_1 f_4 f_3 \to 0.85$
$f_1 f_2 f_3 f_4 \to 0.83$

$\boxed{f_1}$ $\boxed{f_4}$ $\boxed{f_2} \to 0.81$
$\boxed{f_1}$ $\boxed{f_4}$ $\boxed{f_3} \to 0.85$

$\boxed{f_1}$ $\boxed{f_4}$ $\boxed{f_3}$ $\boxed{f_2} \to 0.83$

$\boxed{f_1 f_4 f_3} \to$ model

Backward $\to$ forward

$\boxed{563} \leftarrow$ best 550 backward

50 best $\to$ forwar

$1 \to 50$

best $\to$ f, b

n - features

Exhaustive $\to 2^n - 1$

Backward eli $\to \frac{n(n+1)}{2}$

Forward sel $\to \frac{n(n+1)}{2}$

100 feature

90 best $\to$ back

10 best $\to$ forw

10 best

100 best

Advantages and Disadvantages

13 May 2023 09:06

filter

Advantages

1. Accuracy: Wrapper methods usually provide the best performing feature subset for a given machine learning algorithm because they use the predictive power of the algorithm itself for feature selection.

2. Interaction of Features: They consider the interaction of features. While filter methods consider each feature independently, wrapper methods evaluate subsets of features together. This means that they can find groups of features that together improve the performance of the model, even if individually these features are not strong predictors.

Disadvantages

1. Computational Complexity: The main downside of wrapper methods is their computational cost. As they work by generating and evaluating many different subsets of features, they can be very time-consuming, especially for datasets with a large number of features.

2. Risk of Overfitting: Because wrapper methods optimize the feature subset to maximize the performance of a specific machine learning model, they might select a feature subset that performs well on the training data but not as well on unseen data, leading to overfitting.

3. Model Specific: The selected feature subset is tailored to maximize the performance of the specific model used in the feature selection process. Therefore, this subset might not perform as well with a different type of model.