Feature Selection

Dr. Supaporn Erjongmanee

Department of Computer Engineering
Kasetsart University
fengspe@ku.ac.th

Outline

- Introduction
- Feature Selection Methods
  - Filter Methods
  - Wrapper Methods
  - Embedded Methods

Current Trend of Data

- Data nowadays come with large size.
- Variables continuously grow
  - No more 2-3 variables
- Question: If we have 1000 variables, shall we analyze them all?
- Answer: It is better to select subset of variables to study

Feature Selection

- Selecting subset of variables
  - Given n features, how to select m best features
- Also known as variable selection
- Requirement
  - Criterion to select the best
  - Algorithm to select features
- Common methods
  - Filter
  - Wrapper
  - Embedded

1. Performance stops increasing/decreasing
2. Predefined number of features is reached.

Feature Selection Benefit

- Why do we perform feature selection?
  - Less data storage
  - Less computation time
  - Easier in analysis (e.g., pattern recognition)
    - Removing redundant variables
    - Easier in visualizing
    - Improving performance

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Terminology

- Collinearity
  - Correlation between independent variables
  - When two independent variables are highly correlated,
    - Normally, we remove one of them
    - Concern: Sometimes they must be used together to predict output.
      - Remove one variable -> Increase in p-value (model is worse).
- Multicollinearity
  - When >=2 variables are highly linearly correlated

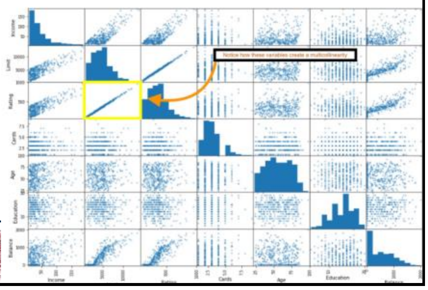


Image source: https://medium.com/future-vision/collinearity-what-it-means-why-its-bad-and-how-does-it-affect-other-models-94e1db984168

Supaporn Erjongmanee
fengspe@ku.ac.th

Outline

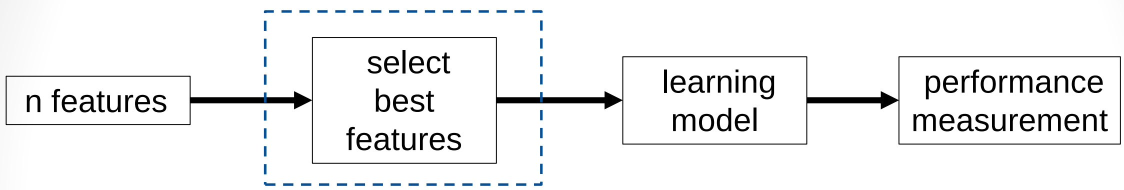
- Introduction
- Feature Selection Methods
  - Filter Methods
  - Wrapper Methods
  - Embedded Methods

Supaporn Erjongmanee
fengspe@ku.ac.th
Statistics in Computer Engineering
Slide 7
Department of Computer Engineering
Kasetsart University

7

Filter Method

- Independent of learning algorithm



- Generally used for pre-processing
- Given features $x_1, x_2, \ldots, x_n$, where $1 \le i \le n$, to predict output (dependent variable)
  - Use scoring function S(i) to assign ranks to each feature
  - Remove features with lowest ranks

Supaporn Erjongmanee
fengspe@ku.ac.th
Statistics in Computer Engineering
Slide 8
Department of Computer Engineering
Kasetsart University

8

Filter Method (cont.)

- Examples of scoring function
  - Missing Value
  - Variance
  - Correlation (R)
  - Mutual information : I(i)
  - Chi-squared test: measure dependency between features
  - Others: Markov blanket, Consistency-based filter.

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Missing Value Ratio

- Compute percent (or ratio) of missing data
- Acceptable threshold for %missing data
  - ~ 20-30%
- What to do if %missing data > threshold
  - Replace missing values with some other values
  - Drop such variable with large %missing values

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Variance Filtering

- Consideration:
  - Variables with low variance has small effect on target variable
- Solution
  - Remove variables with very small variance

Correlation Filtering

- When any two variables have high correlation, they are likely to have the same information
  - Normally, one of them can be dropped.
  - May need to consider overall performance after drop one of them
- Acceptable “high” correlation coefficient
  - $> 0.5 - 0.6$

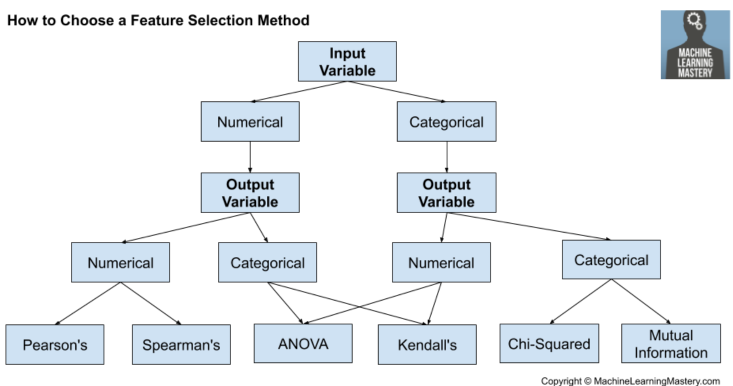
Filter Method (cont.)

- Mutual information: I(i)
  - Measure dependency between x_i and y
  - $I(i) = \int_{x_i} \int_y p(x_i, y) \log \frac{p(x_i,y)}{p(x_i)p(y)} dxdy$
  - Can also be computed from
  - $I(i) = \sum_x \sum_y P(X = x, Y = y) \log \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}$
  - If x_i and y are independent I(i) = 0

13

Filter Method (cont.)

- **Goal**: Drop highly correlated variables + Keep independent variables
- Methods depends on data types



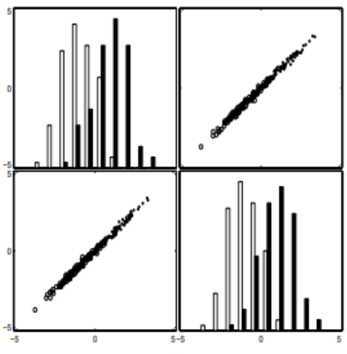
Linear Non-linear Linear Non-linear

Source: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

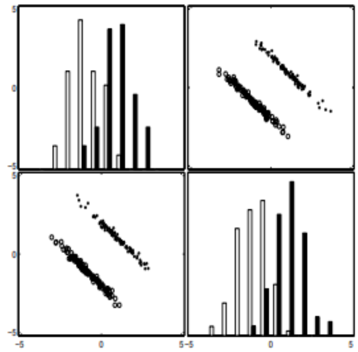
14

Filter Method (cont.)

- Example: Why high correlated variables are redundant?



No information is gained when two variables are highly correlated.



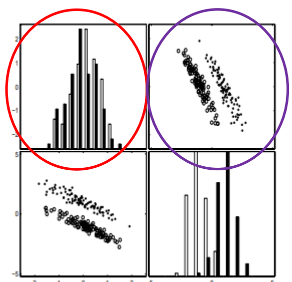
More information is gained when two variables are used

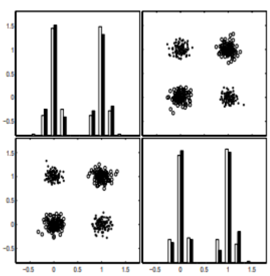
Image source; [1]

15

Filter Method (cont.)

- Example 2: “Useless” variable may be meaningful when used with the other





Looking from histogram, two variables seem to have same distribution. -> One variable seems to be “useless”

Look at scatter plot & correlation, more information is gained if two variables are used.

Image source; [1]

16

Filter Method (cont.)

- Advantage
  - Simple
  - Not computationally intensive
  - Independent from learning model

- Disadvantage
  - In general, feature subset provides lower performance than the other methods
  - Do not consider collinearity among features

17

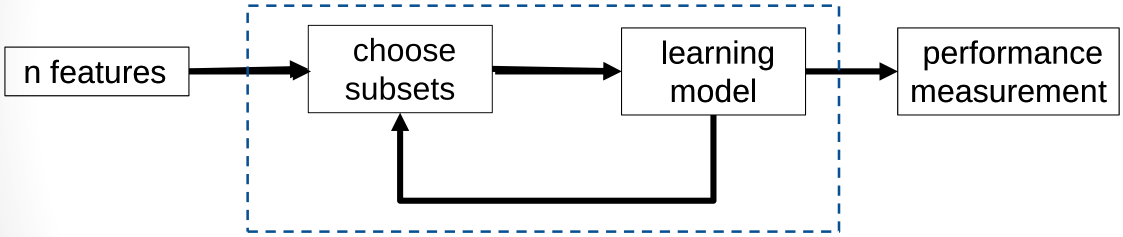
Outline

- Introduction
- Feature Selection Methods
  - Filter Methods
  - Wrapper Methods
  - Embedded Methods

18

Wrapper Method

- Repetitively select subset of features and test with learning model
- Adjust current subset based on performance from the last subset, until obtaining the best subset



- Equivalent to search algorithm
- Computationally intensive

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Wrapper Method (cont.)

- Common methods:
  - Forward feature selection
  - Backward feature elimination
  - Recursive feature elimination
  - Others: Exhaustive feature selection, Bidirectional search

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Forward Feature Selection

- Process of adding features
- Start with subset = no variable
- From each original variable, do the followings
  1. Derive model and compute performance
  2. Choose one variable that results to highest improve of performance
  3. Add variable in step 2 to subset and use with each of remaining variable to derive model
  4. Go back to step 1 unless no addition of variable improves performance

Forward Feature Selection (cont.)

- Example: Features = {1,2,3,4,5}

  Subset = {}

  Use one variable -> Train model

1	2	3	4	5

Assume variable 1
results to highest performance

Subset ={ 1 + ? }

Add one more variable to subset -> Train model

1,2	1,3	1,4	1,5

Assume 1,3
results to highest improve of performance

Forward Feature Selection (cont.)

- Example (cont.)

Subset = {1,3 + ?}

Add one variable to subset -> Train model

| 1,3,2 | 1,3,4 | 1,3,5 |

Assume 1,3, 5
results to highest improve of performance

...

Continue until performance no longer (or hardly) improves

Backward Feature Elimination

- Process of removing variables
- Start with subset = n original variables
- Do the followings:
  1. Derive model and compute performance
  2. Drop one variable that results to least improve of performance
  3. Go back to step 1 unless no performance is improved

Backward Feature Elimination (cont.)

- Example: Features = {1,2,3,4,5}

Subset = {1,2,3,4,5} -> Train model

Drop one variable -> Train model

| 2,3,4,5 | 1,3,4,5 | 1,2,4,5 | 1,2,3,5 | 1,2,3,4 |

Assume dropping 1
results to least improve of performance

Subset = { 2,3,4,5}

Drop one variable -> Train model

| 3,4,5 | 2,4,5 | 2,3,5 | 2,3,4 |

Assume dropping 3
results to least improve of performance

Supaporn Erjongmanee
fengspe@ku.ac.th
Department of Computer Engineering
Kasetsart University

Backward Feature Elimination (cont.)

- Example (cont.)

Subset = {2,4,5}

Drop one variable -> Train model

| 4,5 | 2,5 | 2,4 |

...

Continue until no performance (or hardly) improves

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Recursive Feature Elimination

- Apply greedy algorithm
- Start with subset = n original variables -> Train model
- Obtain rank of performance for each variable (e.g., coefficient in regression)
- Do the followings:
  1. Drop variable with lowest rank from subset
  2. Use subset with remaining variables to train model
  3. Obtain rank of performance for each remaining variable
  4. Go back to step 1 unless all variables are used.

Wrapper Method (cont.)

- Advantage
  - Generally, provide best performance set of features
- Disadvantage
  - Computationally intensive
  - Some methods do not consider collinearity among features

Outline

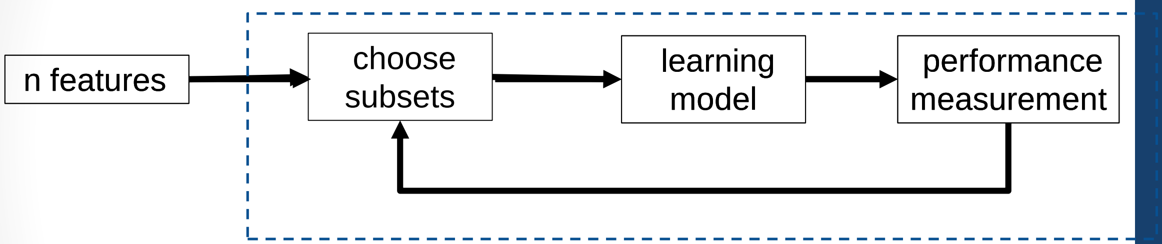
- Introduction
- Feature Selection Methods
  - Filter Methods
  - Wrapper Methods
  - Embedded Methods

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Embedded Method

- Combine filter and wrapper together
- Built in learning model



Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Embedded Method (cont.)

- Regularization: Adding penalty term during training model to remove insignificant variables
  - Regression
    - Lasso, Ridge, Elastic net
- Tree-based Feature Importance

Regularization

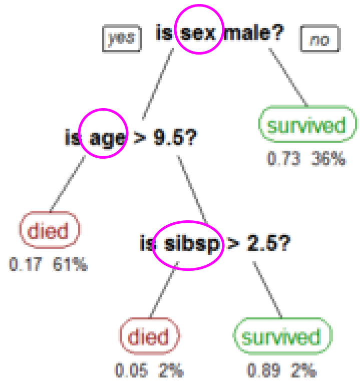
Model = Original_Regression + Penalty

- Add penalty during training model to remove insignificant variables
  - Lasso Regulation (L1)
    - During training, add penalty term (using *absolute distance*) to decrease some coefficients of variables to zero.

      These variables are not significant. -> They can be removed
  - Ridge Regulation (L2)
    - During training, add penalty term (using *square distance*) to decrease some coefficients of insignificant variables to zero.
  - Elastic Net (L1/L2)
    - Combination of using both absolute distance and square distance in penalty

Tree-Based Feature Importance

- Decision tree / Random forest concept
  - Classification model
  - Important variables are put on the top of the tree
  - Variables can be both categorical and numerical data
- Choose subset of *significant variables* from top of the tree



Sex > Age > Sibsp

Image source: https://upload.wikimedia.org/wikipedia/commons/f/f3/CART_tree_titanic_survivors.png

Tree-Based Feature Importance (cont.)

- Feature importance is computed based on
  - Mean decrease accuracy
    - When a variable is left out from model, how much accuracy decreases
  - Mean decrease impurity
    - Impurity: probability of incorrect classification
    - When a variable is left out from model, how much impurity decreases
- The more decrease in accuracy and impurity, the more important features

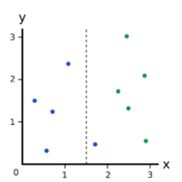


Image source: https://victorzhou.com/blog/gini-impurity/

Embedded Method (cont.)

- Advantage
  - Faster than wrapper methods
  - Better performance than filter methods
- Disadvantage
  - Bound to specific learning model

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Conclusion

- Feature selection is to choose subset of original features for learning model
  - Aim to choose more relevant features to output
  - Save computation time and data storage
- Mainly there are 3 methods:
  - Filters: Correlation
  - Wrapper: Forward Feature Selection, Backward Feature Elimination
  - Embedded: Bounded to learning model
    - Example: Regression, Forest tree

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

References

1. I. Guon and A. Elisse, "An introduction to variable and feature selection," Journal of Machine Learning Research, No. 3, pp.1157-1182, 2003, Available at http://www.ai.mit.edu/projects/jmlr/papers/volume3/guyon03a/source/old/guyon03a.pdf
2. B. Ghojogh, et.al., "Feature selection and feature extraction in pattern analysis: a literature review",
3. A. Jovic, et. al., "A review of feature selection methods in applications"
4. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/
5. https://heartbeat.fritz.ai/hands-on-with-feature-selection-techniques-embedded-methods-84747e814dab