An introduction to variable and feature selection (Part 1)

Isabelle Guyon, Andre Elisseeff

Berkeley,Max Planck Institute for Biological Cybernetics

JMLR,2004

Outline

Outline

Introduction

Why do variable & feature selection

- Facilitate data visualization & data understanding.
- Reduce the measurement & storage requirements
- Reduce training & utilization times
- Defying the curse of dimensionality and improve the prediction performance.

Why do feature construction

- Improve the prediction performance.

This paper focuses on **constructing & selecting subsets of features that are useful to build a good predictor.**

Several methods

Three main parts in this slides

- Variable Ranking
- Small but Revealing Examples
- Variable subset selection

Outline

Motivation

About variable ranking

- **Nice properties**: simplicity, computational and statistical scalability, and good empirical success!
- **Variable ranking is a filter method**: it is a preprocessing step, independent of the choice of the predictor.
- **Under some certain assumptions (independence or orthogonality), it may be optimal with respect to a given predictor.**
  - Only requires computating n scores and sorting the scores.
  - Robust against overfitting because it introduces bias but it may have considerably less variance.

Since variable ranking considers variables one by one, it **ignores relationships between variables**, but it is quite simple and practical.(*The following techniques can be seen frequently in kaggle notebooks with high popularity*).

Correlation Criteria

Pearson correlation coefficient

$$R(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$$

- $X_i \in R^m$: the ith feature of all examples, m is the number of the examples; Y: the label

In linear regression

$R(i)^2$: represents the fraction of the total variance around the mean value $\bar{y}$ that is explained by the **linear relation** between x_i and y.

- $R(i)$: only detect linear dependencies between variable and target.
- One way is to make a non-linear fit of the target with single variables and rank according to the goodness of fit.

Single Variable Classifiers

Single Variable Classifiers

- Select variables according to their individual predictive power, using as criterion the performance(error rate,fpr,fnr) of a classifier built with a single variable.
- When there is lots of variables that separate the data perfectly, ranking criteria based on classification success rate cannot distinguish between the top ranking variables.
  - Use correlation coefficient or another statistic like the margin (the distance between the examples of opposite classes that are closest to one another for a given variable).

Information Theoretic Ranking Criteria

Information Theoretic Ranking Criteria

$$I(i) = \int_{x_i}\int_{y} p(x_i, y) log \frac{p(x_i, y)}{p(x_i)p(y)} dxdy$$

- $p(x_i), p(y)$: the probability densities of x_i and y.
- $p(x_i, y)$: the joint probability.
- The probabilities are usually **estimated from frequency counts.**

The case of continuous variables (and possibly continuous targets) is the hardest. **One can consider discretizing the variables or approximating their densities with a non-parametric method such as Parzen windows.**

Outline

Motivation

Motivation

This part use some examples to outline the usefulness and the limitations of variable ranking techniques and present several situations in which **the variable dependencies cannot be ignored**.

Can Presumably Redundant Variables Help Each Other?

Drawbacks of variable ranking

- Leads to the selection of a redundant subset.(Same performance could possibly be achieved with a smaller subset of complementary variables).

Noise reduction and better class separation may be obtained by adding variables that are presumably redundant. Variables that are independently and identically distributed are not truly redundant.

How Does Correlation Impact Variable Redundancy?

Two important conclusions

- Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them.
- Very high variable correlation (or anti-correlation) does not mean absence of variable complementarity

Can a Variable that is Useless by Itself be Useful with Others?

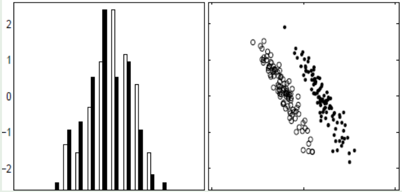
Tempt and worry

Multivariate methods are prone to overfitting especially when the number of variables to select from is large compared to the number of examples.

- It is tempting to use a variable ranking method to filter out the least promising variables before using a multivariate method.
- But whether we will lose some valuable variables through filtering process? **(YES!)**

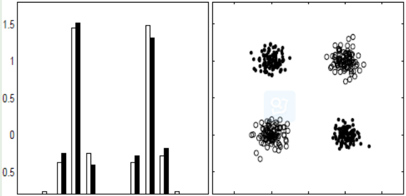
Can a Variable that is Useless by Itself be Useful with Others?

Example



One variable is useless while two dimensional separation is better than the separation using the useful variable alone.

Example



Two variables useless by themselves can be useful together.

Outline

Motivation

Motivation

Since the usefulness of selecting subsets of variables that together have good predictive power, as opposed to ranking variables according to their individual predictive power. This part outlines **main directions** that have been taken to tackle it.

- **Wrappers**: utilize the learning machine of interest as a black box to score subsets of variable according to their predictive power.
- **Filters**: select subsets of variables as a pre-processing step, independently of the chosen predictor.
- **Embedded methods**: perform variable selection in the process of training and are usually specific to given learning machines.

Wrappers and Embedded Methods

Wrapper methodology

- Wrapper methodology consists in using the **prediction performance** of a given learning machine to assess the relative usefulness of subsets of variables.
  - How to search the space of all possible variable subsets;
  - How to assess the prediction performance of a learning machine to guide the search and halt it;
  - Which predictor to use.
- Use exhaustive search if number of variables is not large.
- When search becomes computationally intractable(try best-first, branch-and-bound, genetic algorithms).
- Often criticized for they seem to be a brute force method.
  - Coarse search strategies may alleviate overfitting.
  - Greedy search strategies are computationally efficient and robust against overfitting **(two favors:forward selection and backward elimination)**.

Wrappers and Embedded Methods

Embedded methods(such as CART)

- Incorporate variable selection(Decision Tree like gini index) as part of the training process which may be efficient in several respects.
  - Better use of the available data by not needing to split the training data into a training and validation set;
  - Reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated.

Wrappers and Embedded Methods

Embedded methods:predict the change in objective function

- s: number of variables selected at a given algorithm step.
- $J(s)$: the value of the objective function of the trained learning machine using such a variable subset. Predicting the change in the objective function is obtained by:
  - **Finite difference calculation**: The difference between $J(s)$ and $J(s+1)$ or $J(s1)$(**linear least-square model**: The Gram-Schmidt orthogonolization procedure permits the performance of forward variable selection by **adding at each step the variable that most decreases the mean-squared-error**).
  - **Quadratic approximation of the cost function**:
  - **Sensitivity of the objective function calculation**: (One variant: replace the objective function by the leave-one-out cross-validation error).

Direct Objective Optimization

Components of objective functions

- The goodness-of-fit (to be maximized)
- The number of variables (to be minimized),usually referred as **regularization term**.
  - l_0 norm is the original intuition. But for its hardness in optimization, so l_1, l_2 norm are preferred.
  - In practice, l_1-norm minimization suffices to drive enough weights to zero.

To my knowledge, no algorithm has been proposed to directly minimize the number of variables for non-linear predictors.

Filters for Subset Selection

Benifits of filters compared with wrapper methods

- Some filters (e.g. those based on mutual information criteria) provide a generic selection of variables, not tuned for/by a given learning machine.
- Filtering can be used as a preprocessing step to reduce space dimensionality and overcome overfitting.

It seems reasonable to use a wrapper (or embedded method) with a linear predictor as a filter and then train a more complex non-linear predictor on the resulting variables.