An introduction to variable and feature selection (Part 2)

Isabelle Guyon, Andre Elisseeff

Berkeley,Max Planck Institute for Biological Cybernetics

JMLR,2004

Outline

Outline

Motivation

Motivation

Reducing dimensionality of data by selecting a subset of the original variables may be advantageous including the expense of making, storing and processing measurements.

- If these are not of concern, other means of space dimensionality reduction should also be considered.

Goal of feature construction

- Achieve best reconstruction of the data.
  - Unsupervised learning problem which is closely related to data compression and lots of algorithms are used across both fields.
  - Less prone to overfitting.
- Be most efficient for making predictions.
  - Supervised: more easily to overfit.

Feature construction: usual methods

Generic feature construction methods

- Clustering
- Linear transforms of input variables (PCA,LDA).
- Sophisticated linear transforms like spectral,wavelet transforms (Fourier, Hadamard) or convolutions of kernels.
- Simple functions to subsets of variables, like products to create monomials.

Feature construction: Clustering

Clustering

Replace a group of similar variables by a cluster centroid, which ***becomes a feature***.

- Unsupervised methods (e.g. K-means, works well in text processing, research shows that few words end up belonging to several clusters, hinting that hard cluster assignment may be sufficient, hierarchical clustering).
  - Rarely overfit.
- Supervised methods(e.g. distributional clustering):can obtain more discriminant features.
  - Easy to overfit(utilize the label of the training data).

Feature construction: Matrix Factorization

Matrix Factorization

- Singular value decomposition(SVD).
  - Goal of SVD is to form a set of features that are linear combinations of the original variables.
- Sufficient dimensionality reduction(SDR).
  - Most informative features are extracted by solving an optimization problem that monitors the tradeoff between data reconstruction and data compression.
- Etc.

Outline

Motivation

Motivation

Distinguish out-of-sample performance prediction (generalization prediction) and model selection.

Model selection & Final performance of the predictor evaluation

Model selection & Final performance of the predictor evaluation.

- **Evaluate the final performance of the predictor**: important to **set aside an independent test set.The remaining data is used both for training and performing model selection**.
- **For model selection (including variable/feature selection and hyperparameter optimization)**: the data not used for testing may be further split between fixed training and validation sets, various methods or cross-validation.
  - **For a fixed validation set**, statistical tests can be used, but their validity is doubtful for cross-validation because independence assumptions are violated.
  - **If there are sufficiently many examples**, it may not be necessary to split the training data.

Model selection & Final performance of the predictor evaluation

Hints

- Leave-one-out formulas can be viewed as corrected values of the training error.
  - High variance estimator of generalization error & give overly optimistic results, particularly when data are not properly independently & identically sampled from the ”true” distribution.
  - Need lots of resources to train models,so usually N fold cross validation is enough.
- Other suggestions, please refer to newest papers.

Outline

Outline

Outline

This part focuses on explaining many problems relying on methods described in this paper and give some introduction.

Variance of Variable Subset Selection

Variance of Variable Subset Selection

- **If the data has redundant variables**, different subsets of variables with identical predictive power may be obtained according to initial conditions of the algorithm, removal or addition of a few variables or training examples, or addition of noise. **High variance!**.
  - Variance often generates ”bad” model that does not generalize well;
  - Results are not reproducible
  - One subset fails to capture the ”whole picture”.
- One method to ”stabilize” variable selection explored in this issue is to use several **”bootstraps” (very practical!)**.
- Related ideas can be found in the context of Bayesian variable selection.

Variable Ranking in the Context of Others

Variance of Variable Subset Selection

- This paper limits presenting variable ranking methods using only a criterion computed from single variables, ignoring the context of others.
- Introduce nested subset methods that provide a useful ranking of subsets, not of individual variables: some variables may have a low rank because they are redundant and yet be highly relevant.
- Bootstrap and Bayesian methods may be instrumental in producing a good variable ranking incorporating the context of others.

Unsupervised Variable Selection

Sometimes, **no target y is provided**, but we still want to select a set of most significant variables with respect to a defined criterion.

Unsupervised Variable Selection

- A number of variable ranking criteria are useful across applications, including *saliency, entropy, smoothness, density and reliability*. A variable is salient if it has a high variance or a large range, compared to others.
- A variable is reliable if measurement error bars computed by repeating measurements are small compared to the variability of the variable values.
- Perform variable or feature selection for clustering applications. [Xing and Karp, 2001, Ben-Hur and Guyon, 2003, and references therein]

Forward vs. Backward Selection

Comparisons between forward and backward selection

- Forward selection is computationally more efficient to generate nested subsets of variables.
- Weaker subsets are found by forward selection because the importance of variables is not assessed in the context of other variables not included yet.
- A variable is reliable if measurement error bars computed by repeating measurements are small compared to the variability of the variable values.
- Perform variable or feature selection for clustering applications. [Xing and Karp, 2001, Ben-Hur and Guyon, 2003, and references therein]

The Multi-class Problem

The Multi-class Problem

- All the methods based on *mutual information criteria* extend naturally to the multi-class case
- Multi-class variable ranking criteria include Fishers criterion (the ratio of the between class variance to the within-class variance) which is closely related to F statistic used in the ANOVA test.
- Wrappers or embedded methods **depend upon the capability of the classifier** used to handle the multi-class case(e.g. SVM, LDA).
- **Multi-class setting is in some sense easier for variable selection than two-class case**.

Selection of Examples & Inverse Problems

Selection of Examples

- Dual problems of feature selection/construction are those of pattern selection/construction.
- Selecting wrong variables associated with a confounding factor may be avoided by focusing on informative patterns that are close to the decision boundary.

Inverse Problems

- In some applications(particularly in bioinformatics), use a subset of variables to improve the predictor is not the only goal of variable selection.
  - In diagnosis problems, for instance, it is important to identify the factors that triggered a particular disease or unravel the chain of events from the causes to the symptoms.

Outline

Conclusions

Motivation

- Sophisticated wrapper or embedded methods improve predictor performance compared to simpler variable ranking methods, but improvements *are not always significant*: domains with large numbers of input variables suffer from the curse of dimensionality and multivariate methods may overfit the data.
- For some domains, applying first a method of automatic feature construction yields improved performance and a more compact set of features.
- We recommend using a linear predictor of choice (e.g. a linear SVM) and select variables in two alternate ways
  - With a variable ranking method using a correlation coefficient or mutual information.
  - With a nested subset selectio method performing forward or backward selection or with multiplicative updates.