JOINING ALL PIECES TOGETHER

The purpose of this chapter is to provide several examples of MATLAB® code that would demonstrate how to combine all pieces (feature selection, classification, ensembles, and performance evaluation) together. The first three examples show feature selection followed by classification and performance evaluation, while the last example is intended to demonstrate the superiority of ensembles over individual classifiers.

All examples utilized the colon cancer data set (Alon, et al., 1999). The original data underwent preprocessing customary to this data set. This data set contains expression of 2,000 genes for 62 cases (22 normal and 40 colon tumor). Preprocessing includes the logarithmic transformation to base 10, followed by normalization to zero mean and unit variance. The code implementing normalization is given in functions preprocess and normalize. If you have already preprocess colon data, these functions are unnecessary to run.

The only thing to remember is the fact that all functions expect data matrices to be \( D \times N \), where \( D \) is the number of features (genes) and \( N \) is the number of data instances, while vectors of class labels need to be \( 1 \times N \). If your matrices/vectors do not meet these requirements, then transposition will do the job for you. The code
of `preprocess` converts class labels to either -1 (colon tumor instance) or +1 (normal instance). A different labeling (e.g., 0/1) can be easily obtained if desired.

As you remember, there are two types of feature selection models: filters and wrappers. In all examples, I chose filter-like models that do not use a classifier to access feature importance. I recommend you to follow this choice as a safeguard against model overfitting. As for classifiers, I prefer those algorithms that do not require training (e.g., $k$-nearest neighbors). Given that we have only a few instances of high-dimensional data, classifier training will put pressure on our scarce data resources, which can easily result in overfitting. Of course, as you saw in one of the chapters, describing the bolstered error estimation, extra data may be artificially and cheaply generated but I would not abuse this option too often, too much, since artificial data are only a temporal solution to substitute real data until the latter become abundant. This way of reasoning does not make wrappers and classifiers requiring training completely useless: given enough data, they might yield better results than filters and classifiers without the need for training.

In all examples that will follow, the original (training) data are used for feature selection. After that, the classification occurs, using only selected genes. The bolstered resubstitution error is an estimate of the classification performance. It is estimated based on artificially generated data that play the role of test data. Since feature selection and classification are carried out on different data sets, there is no danger of severe overfitting, which could result in optimistically biased performance characteristics.

```matlab
function [patterns, targets] = preprocess(filename1, filename2)

% Preprocess colon data for further analysis
% Colon data can be found at
% http://microarray.princeton.edu/oncology/affydata/index.html
% They were used by U. Alon, N. Barkai, D. A. Notterman, K. Gish,
% S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene
% expression revealed by clustering analysis of tumor and normal colon
% tissues probed by oligonucleotide arrays", Proceedings of National
% Academy of Sciences, Vol. 96, pp. 6745-6750, June 1999
%
% Inputs:
%   filename1    - Name of a file with data
%   filename2    - Name of a file with labels
%
% Outputs:
%   patterns     - Data matrix (DxN), where D is the number of
```

Copyright © 2011, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Related Content

Human Biobanks: Selected Examples from and beyond Europe
[www.igi-global.com/chapter/human-biobanks-selected-examples-beyond/47301?camid=4v1a](www.igi-global.com/chapter/human-biobanks-selected-examples-beyond/47301?camid=4v1a)

Prediction of Protein-Protein Interactions Between Human Host and Two Mycobacterial Organisms
[www.igi-global.com/article/prediction-protein-protein-interactions-between/40968?camid=4v1a](www.igi-global.com/article/prediction-protein-protein-interactions-between/40968?camid=4v1a)

Privacy Preserving Principal Component Analysis Clustering for Distributed Heterogeneous Gene Expression Datasets
Xin Li (2013). *Methods, Models, and Computation for Medical Informatics* (pp. 238-271).
[www.igi-global.com/chapter/privacy-preserving-principal-component-analysis/73081?camid=4v1a](www.igi-global.com/chapter/privacy-preserving-principal-component-analysis/73081?camid=4v1a)
Data Mining for Biologists

[www.igi-global.com/article/data-mining-for-biologists/101239?camid=4v1a](www.igi-global.com/article/data-mining-for-biologists/101239?camid=4v1a)