INFO-F-528 Machine learning methods for bioinformatics
Exercise session 5: dimensionality reduction in gene expression data

Gianluca Bontempi
Machine Learning Group, Computer Science Department,
Université Libre de Bruxelles

1 Gene expression dataset

Why analysing microarray datasets?

- In a study of breast cancer in women, suppose that one of the phenotype variables be the lymph node status. We could be interested in studying the correlation between gene expression levels and this phenotype.
- Suppose to have observed the dynamics of the expression level of $G = 6000$ yeast genes for $T = 10$ instant times under $C = 10$ nitrogen conditions. We could be interested in studying the correlation of the time pattern of expression to the nitrogen condition.

An example of microarray dataset is the one used by Golub et al. in [**?**]. This dataset contains the genome expressions of $n = 7129$ genes of $N = 72$ patients, for which $V = 11$ phenotype variables are measured. In the following we will focus on the correlation between the expression of the genes and one specific phenotype variable, the ALL.AML factor indicating the leukemia type: lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML).

The expression matrix X and the phenotype vector Y are contained in the file `golub.Rdata`.

Classification of microarray is characterized by a very large dimensionality (e.g. large number of gene probes) with respect to the number of samples (e.g. number of patients). We will consider several ways of reducing dimensionality

2 Gene filtering

Not all genes are expressed in all tissue types. In order to ease the computational burden (and reduce the chance of spurious results) it is useful to remove those

genes that have little or no variation in the samples being analyzed. In Golub the following transformations were applied to data

1. thresholding: between 100 and 16,000
2. exclusion of genes with $\frac{\max}{\min} \leq 5$ or $(\max - \min) \leq 500$, where max and min refer to all the values taken by the gene in the different cases.
3. take a base 10 logarithmic transformation.

A technique of filtering based on the t-test has been discussed in the course of Statistics. The Golub filter and two t-test filters (having different p-values) are implemented in the file `filter.R`.

Exercises

Use a subset of the genes to classify the cancer type. Make a script `fs_fil.R` (by changing the `KNNvalid.R` script implemented in the session 4) which:

1. performs the 1-NN classification with the subset of genes obtained by the 3-steps Golub procedure : the index of genes is contained in the variable `ind.filter1` in the file `golub.filter.Rout`.
2. performs the 1-NN classification with the subset of genes obtained by the t-test filtering ($p = 1e-6$) procedure : the index of genes is contained in the variable `ind.filter2` in the file `golub.filter.Rout`).
3. performs the 1-NN classification with the subset of genes obtained by the t-test filtering ($p = 1e-5$) procedure: the index of genes is contained in the variable `ind.filter3` in the file `golub.filter.Rout`).
4. Did the accuracy improve? Discuss the results.

3 Feature selection

Two are the main approaches to feature selection:

Filter methods: they are preprocessing methods. They attempt to assess the merits of features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by ranking them through compression techniques (like PCA) or by computing correlation with the output.

Wrapper methods: these methods assess subsets of variables according to their usefulness to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. The problem boils down to a problem of stochastic state space search. Example are the stepwise methods proposed in linear regression analysis.

Exercises

1. Write a script `fs_pca.R` which
   - makes the PCA transformation of the training dataset restricted to the genes listed in the variable `ind.filter3`
   - assesses the classification accuracy of different numbers of principal components ($n = 2, 5, 20$) by using the Golub dataset, a 1NN learner and a leave-one-out procedure.
2. Write a script `fs_rank.R` which
   - ranks the variables according to the bivariate correlation with the target
   - assesses the classification accuracy of different sizes of the ranked variables by using the Golub dataset, a 1NN learner and a leave-one-out procedure.

3.1 Wrapping search

The wrapper search can be seen as a search in a space $W = \{0, 1\}^n$ where a generic vector $w \in W$ is such that

$$w[j] = \begin{cases} 0 & \text{if the input } j \text{ does NOT belong to the set of features} \\ 1 & \text{if the input } j \text{ belongs to the set of features} \end{cases}$$

We look for the optimal vector $w^* \in \{0, 1\}^n$ such that

$$w^* = \arg\min_{w \in W} \text{MSE}_w$$

where MSE_w is the generalization error of the model based on the set of variables described by w. Since the number of vectors in W is equal to 2^n, for moderately large n, the exhaustive search is no more possible. Various methods have been developed for evaluating only a small number of variables by either adding or deleting one variable at a time. Three are the main categories:

Forward selection: the procedure starts with no variables. The first input selected is the one which allows the lowest generalization error. The second input selected is the one that, together with the first, has the lowest error, and so on, till when no improvement is made.

Backward selection: it works in the opposite direction of the forward approach. We begin with a model that contains all the n variables. The first input to be removed is the one that allows the lowest generalization error.

Stepwise selection: it combines the previous two techniques, by testing for each set of variables, first the removal of features beloning to the set, then the addition of variables not in the set.

Exercise

Implement a R script `fs_wrap.R` which

- implements a forward feature selection function which assesses the subsets' accuracy by leave-one-out error of a 1NN classifier.
- assesses the classification accuracy of different sizes of the selected variables by using the Golub dataset, a 1NN learner and a training-and-test procedure.