Dataset: "leukemia"

The source

T. R. Golub *et al*., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, vol. 286, 15 October 1999. ***www.sciencemag.org***

The issue

Classifying cancer on the basis of DNA microarray analysis rather than morphological appearance analysis.

Application to acute leukemias.

The aims

1. Class prediction: discriminate between known types of tumor
2. Class discovery: reveal previously unknown subtypes of tumor

The classes

Two main class labels:

AML - ***acute mieloid leukemia***

ALL - acute lymphoblastic leukemia

+ other informations annotated for each sample.

The files

The data available at

http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html

are split into a training set and a test set (the independent set**).**

We have organized them in two ASCII files:

leukemia.train and **leukemia.test**

Feel free to experiment with different arrangements

(e.g. shuffling patterns, different normalizations and pre-processing...)

The data

Data are 7129-dimensional patterns.

leukemia.train: 38 patterns ; **leukemia.test:** 34 patterns

Each feature is the expression level of a particular gene.

Each pattern is a sample from a unique patient.

Training set is from bone marrow samples. Test set is from other samples

The format

```
<number-of-patterns>
<dimension-of-patterns>                          (= 7129)
<gene1-expression-value>  <gene2-expression-value> ...
  (7129 values for pattern 1)
<class label>                  (= 1 for ALL, -1 for AML)
<gene1-expression-value>  <gene2-expression-value> ...
  (7129 values for pattern 1)
<class label> ...
```

Values are column-wise normalized into [0,1]

Refer to the paper and to its companion website

See the ENSEMBLE LAB webpage!