Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

A Practical Example

Staal A. Vinterbo

Harvard-MIT Division of Health Science and Technology

Decision Systems Group, BWH

Harvard Medical School

Dec 2005: HST 951/MIT 6.873 Class

Overview

- Obtain data, prepare it
- Create, validate and compare classifiers
- Determine predictors if possible: hypotheses
- Write report

Data

The data we plan on using:

> *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression, T.R. Golub et. al, Science 286:531-537. (1999).*

Google: “golub all aml data”

Data format

The data comes as:

- Two files: training set and test set
- Each gene on a row
- class in separate file

Need to transform.

Transform

- Use Excel to strip away first column.
- Load into R using read.delim
- Filter columns, transpose and attach class labels

Repeat Original Experiment

- Repeat Classification task of paper
- 4 errors on test

Validate Method

- 8 fold CV

CV comparison with NN

- Compare to ANN using
  - 8 fold CV – T-Test
  - 5×2 CV (Alpaydin, E. Combined 5x2CV *F* Test for Comparing Supervised Classification Learning Algorithms Neural Computation, 1999, 11, 1885-1982)

5×2 CV

The 5x2CV F-test can be used to quantitatively compare the performance of two classifiers. As its name implies, the test is based on performing five replications of 2-fold CV.
Let Δ_{ij} denote the *difference* between the performance measures of the two classifiers on fold $j \in \{1, 2\}$ of replication $i \in \{1, \ldots, 5\}$. The average difference in performance on replication i is $\bar{\Delta}_i$ and the estimated variance is s_i^2.

$$\bar{\Delta}_i = \frac{(\Delta_{i1} + \Delta_{i2})}{2}$$

$$s_i^2 = (\Delta_{i1} - \bar{\Delta}_i)^2 + (\Delta_{i2} - \bar{\Delta}_i)^2$$

5×2 CV

Let H_0 denote the null hypothesis that the two classifiers perform equally well. Under H_0, Δ_{ij} can be treated as being $N(0, \sigma^2)$ distributed, and we have:

$$A = \sum_{i=1}^{5} \sum_{j=1}^{2} \frac{\Delta_{ij}^2}{\sigma^2} \sim \chi^2_{10}$$

$$B = \sum_{i=1}^{5} \frac{s_i^2}{\sigma^2} \sim \chi^2_5$$

$$f = \frac{A/10}{B/5} = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} \Delta_{ij}^2}{2 \sum_{i=1}^{5} s_i^2} \sim F_{10,5}$$

We then reject H_0 if the statistic f is sufficiently large. For 95% confidence, $f = 4.74$.

Markers?

Bioinformatics: Can we suggest markers that discerns between ALL and AML?

The report

What we want to tell:

- Fuzzy Classification Trees are worth while
- They are interpretable
- We found a good marker for discerning ALL from AML

Support

- Comparison to other classifiers
- Repeat of paper classification task
- Show stability of marker

The paper parts:

- Introduction:
  - background – why is this question important
  - what we did – results and significance
- Methods
  - Mathematical preliminaries and definitions
  - Fuzzy Classification Trees
  - Validation methods
- Experiments
  - Data description, and preparation
  - Experimental protocol: hypotheses and what results are needed to confirm
- Results: state the results
- Discussion:
  - Link hypotheses and results and draw conclusion
  - Discuss weaknesses/strengths and items needed to reproduce
  - Hint at further research