Introduction to Machine Learning

Martin Morgan
Bioconductor / Fred Hutchinson Cancer Research Center
Seattle, WA, USA

24 November, 2009

Machine Learning

- *Unsupervised* machine learning involves discovery of classes without *a priori* knowledge or use of sample classification. Unsupervised machine learning is sometimes called *cluster analysis*.
- *Supervised* machine learning uses prior knowledge of sample classification to develop algorithms for class membership prediction.
- *Dimensional reduction* is often relevant for gene expression data.

Data and Algorithms for Machine Learning

- Machine learning requires selection of *samples* and *features*, choice of distance (*similarity*) metric, and choice of algorithm.
- Features are usually pre-processed, log-transformed gene expression values. Dimensional reduction will often be used to identify a subset of features, or mathematical combinations of features, that greatly reduces the size of the machine learning problem.
- A distance metric represents how far samples are separated from one another in 'feature space'.
- There are many machine learning algorithms implemented in R; choice requires prior motivation or careful assessment of algorithm performance.

Data Input and Reduction

Samples

- We use the `ALL` (acute lymphocytic leukemia) data set of Chiaretti *et al.*, subset to just those samples from B-cells with BCR/ABL or NEG molecular biology.
- The subset is partly pragmatic (a readily usable example), but also reflects valid research directions.

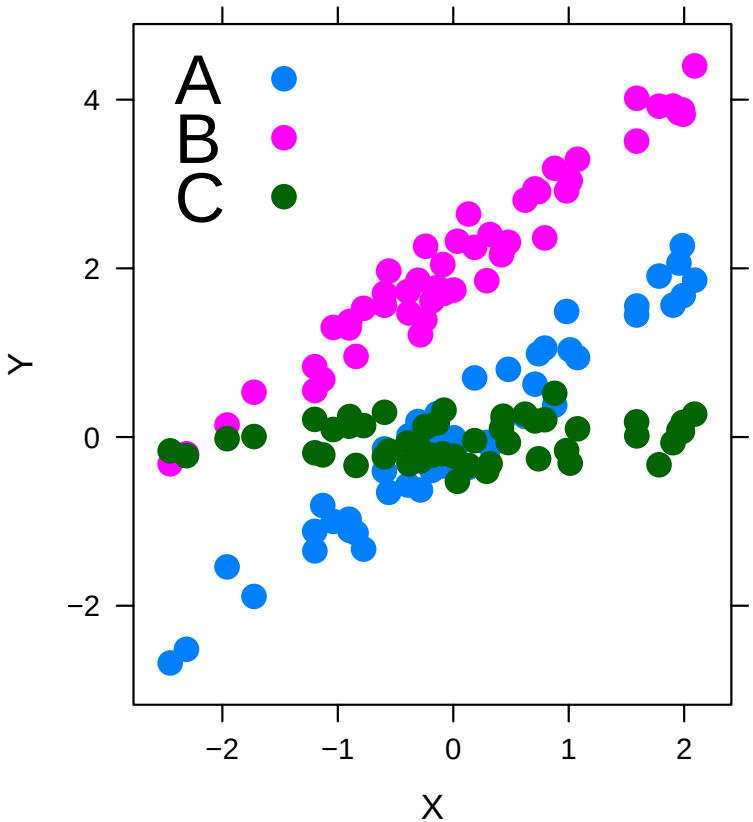
Features

- Variability: use a non-specific (with respect to sample classification) variance filter to remove non-variable probes.
- Biologically relevant: e.g., reliably annotated with transcription-related Gene Ontology (GO) ids.

Distance Metrics

- Choice of metric has important consequences for outcome.
- Several metrics available in the bioDist and cluster packages. Examples: Euclidean, Manhattan, 1-correlation, Mahalanobis.
- Selecting a distance metric is a necessary step; sometimes implicit in machine learning algorithm, e.g., Support Vector Machines use Euclidean distance.
- All features *a priori* equally informative? Center and scale before calculating distance.

Distance Metrics – Examples



	Corrl'n	Euclid.
A vs. B	Low	High
A vs. C	High	Low
B vs. C	High	Med

Distance Metrics – Choice

What is desired?

- Expression in a time course experiment?
- Find all genes affected by the same transcription factor?
- Samples with known phenotype and related expression profile?

Example: time course.

- Seeking (a) correlated; (b) anti-correlated; (c) lagged genes?
- 1-correlation only relevant for (a); outliers can be disastrous.

Example: transcription factors.

- Unknown what pattern of correlation to expect.
- One strategy: use known targets to guide choice of distance.

Statistical Data Reduction

Issues

- Data is very high-dimensional, and so data difficult to visualize and interpret.
- Many features add little new (uncorrelated) information, so do not help discriminate between groups.

One approach

- Data reduction via multidimensional scaling (or other techniques).
- Transformation reflects feature selection, e.g., features selected on basis of t-test to distinguish groups will likely result in data reduction that effectively discriminates two groups.
- Examples: `stats::cmdscale`, `MASS::sammon`

Unsupervised Machine Learning

Goal: divide data sets so that there is larger within- than between-group similarity.

- How many groups g in the data?
- Which features define groups?
- Both features and distance metric already selected.

Major types of algorithms.

- *Hierarchical clustering*: a hierarchy of clusters from 1 to g. *Agglomerative* starts with g groups, successively coalescing most similar. *Divisive* splits 1 group into 2, 3, . . . , g.
- *Partitioning*: divide data into g groups using a (re)allocation algorithm.

Hierarchical Clustering

- Rely on between-group distance calculations: *single*, *average*, and *complete* linkage, corresponding to the minimum, average, and maximum distance between an element of each group.
- Hierarchical clustering is not deterministic, i.e,. not all possible splits explored.
- Examples: `stats::hclust` (agglomerative), `cluster::diana` (divisive).

Hierarchical Clustering – Dendrograms

- Tree-like structure, with samples (leaf nodes) at the bottom.
- Height of the join indicated distance between the left and right branches.
- *Impose* structure – induced distances that may differ from distanced used to compute the dendrogram. *Cophenetic correlation* (`stats::cophenetic`) provides guidance on whether dendrogram and sample distances agree.
- `stats::cutree` returns trees cut at specified heights.
- Example: handout page 156.

Partitioning

- User selects desired number of groups g.
- Group or cluster centers determined, objects assigned to groups (e.g., randomly).
- Group membership and cluster centers updated through various algorithms. Update to increase goodness of fit.
- Output: assignment of samples to groups.
- Examples: k-means, PAM (partition around mediods), self-organizing maps

Partitioning – Examples

- k-means: partition samples into groups to minimize the sum of squared distances from the samples to the group centers.
- PAM: choose k representative samples (medoids). Assign samples to nearest medoid. Repeat, attempting to minimize dissimilarities of samples to the closest representative object.

Supervised Machine Learning

Goal.

- Classify new samples into groups identified *a priori* or in a *training set*.

Key attributes.

- Groupings are defined *a priori* (e.g., `BCR/ABL` versus `NEG`).
- Some samples are used to train the algorithm; other samples assess the trained algorithm performance.
- *Cross-validation* repeats sample assignment, training, and testing many times, assigning different samples to each set.
- Both supervised and unsupervised machine learning filter samples (to match biological questions) and features (to focus on informative probesets). Both require distance metrics.

Supervised Machine Learning – Algorithms

Examples

- k nearest neighbors: assign each test sample to the same class as the k nearest neighbors in the training set.
- Linear discriminant analysis: identify a linear combination of features that best discriminate between (minimize within-group vs. between-group sums of squares) classes.
- The MLInterfaces package provides an interface to diverse machine learning algorithms.

Cross-validation

- Divide the data into a test and a training set. E.g., *leave-one-out* cross validation assigns all samples but one to the training set. The remaining sample is assigned to the test set.
- Fit the model to the training set, and estimate the error rate on the test set. The *confusion matrix* summarizes how often the trained algorithm mis-classifies test individuals.
- Repeat with data divided into a different training set, e.g., leaving a different sample out.
- Summarize over training sets.

Summary

- Unsupervised machine learning groups samples using a specified distance metrix. It can be a useful tool in quality assessment and exploratory analysis.
- Supervised machine learning emphasizes class prediction and feature selection.