

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, December 4, 2001
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Robert Tibshirani
Department of Statistics
Stanford University
Stanford, CA 94305

Cluster Validation by Prediction Strength

We propose a new quantity for assessing the number of groups or clusters in a dataset. The key idea is to view clustering as a supervised classification problem, in which we must also estimate the “true” class labels. The resulting “prediction strength” measure assesses how many groups can be predicted from the data, and how well. In the process, we develop novel notions of bias and variance for unlabelled data. Prediction strength performs well in simulation studies, and we apply it to clusters of breast cancer samples from a DNA microarray study. Finally, some consistency properties of the method are established.

(This is joint work with Guenther Walther, Pat Brown and David Botstein)