

Department of Applied Mathematics and Statistics
The Johns Hopkins University

SEMINAR

Damianos Karakos
Center for Language and Speech Processing
The Johns Hopkins University

September 16, 2004
304 Whitehead Hall
Refreshments: 3:30 p.m.
Seminar: 4:00 p.m.

**WEIGHTED SUMS OF K–L DIVERGENCES FOR UNSUPERVISED
CLASSIFICATION VIA SENSING AND PROCESSING DECISION TREES**

ABSTRACT

Sensing and Processing Decision Trees (SPDTs) for unsupervised clustering are grown recursively, by (i) splitting high-dimensional data and (ii) transforming the data in each resulting cluster independently of the other clusters, guided by some goodness criterion. It will be shown that maximizing a weighted sum of partition-dependent Kullback–Leibler divergences is an appropriate criterion for unsupervised clustering. In particular, if each data-point is a finite-length sample from a collection of stationary and ergodic distributions, where each distribution corresponds to one class-label, then this criterion is equivalent to sequentially maximizing the mutual information between the unseen class-label of the data-point and the path from the root of the SPDT to the leaf-cluster where the data-point is placed. Empirical results for text classification will be presented to demonstrate the effectiveness of this criterion.