2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo

INFORMATION GEOMETRY AND DOCUMENT CLASSIFICATION

GUY LEBANON

Department of Statistics and School of Electrical and Computer Engineering,
Purdue University

The task of classifying documents according to topic is traditionally based on extracting features, and treating the features as points in a Euclidean space, equipped with Euclidean geometry. We argue that this may be improved upon by examining a more appropriate geometry for text documents, and adapting classification models to this geometry. By embedding documents in the multinomial simplex, we identify a canonical geometry for them - the Fisher geometry on the multinomial simplex. Adapting popular classification models such as radial basis support vector machines and logistic regression to the Fisher geometry yields impressive results in text classification. The application of information geometry to text classification results in an improvement over the-state-of-the-art in this field.

If time remains, I will discuss an extension of Cencov's theorem for spaces of conditional models and a novel geometric representation for documents that moves beyond the standard bag of words assumption.

Collaborator: John Lafferty, Carnegie Mellon University.