SFB 680

Molecular Basis of
Evolutionary Innovations

Molekulare Grundlagen evolutionärer Innovatione

Susanne Still

University of Hawaii, Manoa

Information theoretic clustering

Clustering lies at the heart of most cognitive activities. For example, we group visual data such that they represent the same object. Different acoustic signals stand for the same word and, on a higher level of abstraction, different words can convey the same idea or belong to the same class of concepts. Unfortunately, most clustering algorithms have conceptual shortcomings, because they make a number of ad hoc assumptions. Shannon's Rate-distortion theory offers an information-theoretic approach to clustering whereby the data is compressed under the constraint that the average distortion is fixed. The Information Bottleneck method (introduced in 1999) then obviates the need of an a priori distortion measure, in the clustering application thatwould be a (dis-) similarity) measure. altogether, this method allows for a principled approach to clustering. I will discuss two important results of the Information Bottleneck approach applied to clustering. First, we have shown that the finite size of a data set determines the optimum number of clusters, a fundamental problem not addressed in any general way by standard clustering methods. Second, I will discuss how the frequently used K-means and deterministic annealing algorithms can be derived from the Information Bottleneck principle. In addition to conceptual unification, this approach leads to a new algorithm that is less likelyto get trapped in local minima than K-means.

December 17, 2009

10.00 a. m.

Institute for Genetics, Seminar Room, 4th floor

Host: Michael Lässig

www.sfb680.uni-koeln.de