Title: Annotation on the cheap

Sanjoy Dasgupta

University of California – San Diego

4:30 pm, 09/05/2018, 1195 Bordeaux Dr, Sunnyvale, CA 94089

ABSTRACT

We consider algorithms that take an unlabeled data set and label it in its entirety, given the ability to interact with a human expert. The goal is to minimize the amount of interaction while producing a labeling that satisfies an (epsilon, delta) guarantee: with probability at least 1-delta over the randomness in the algorithm, at most an epsilon fraction of the labels are incorrect.

Scenario 1: The algorithm asks the expert for labels of specific points. This is the standard problem of active learning, except that the final product is a labeled data set rather than a classifier.

Scenario 2: The expert also provides "weak rules" or helpful features.

We will summarize the state of the art on these problems, in terms of promising algorithms and statistical guarantees, and identify key challenges and open problems.

Bio: Sanjoy Dasgupta is a Professor of Computer Science and Engineering at UC San Diego, where he has been since 2002. He works on algorithmic statistics, with a particular focus on unsupervised and minimally supervised learning. He is author of a textbook, "Algorithms" (with Christos Papadimitriou and Umesh Vazirani). He was program co-chair of the COLT conference in 2009 and of ICML in 2013.