Annotation on the Cheap

Sanjoy Dasgupta
UCSD

We consider the task of labeling an entire data set, given the ability to query the labels of specific points within it.

Specifically, we define the ”finite population annotation” (FPA) problem as follows. The input consists of: a set of n data points, each of which has an associated label that is initially missing; and parameters delta, epsilon. The algorithm is given the ability to ask for any of the missing labels, and has to produce a set of n labels such that with probability at least 1-delta, at most an epsilon fraction of them are incorrect. The goal is to meet this statistical requirement without asking for very many labels.

This problem is motivated by scenarios in e-discovery and active learning. We describe three algorithms that solve it by exploiting different types of structure in the data. In each case, we characterize the types of data for which the algorithm requires few labels, and we give experimental illustrations of its behavior.

This is joint work with David Lewis and Matus Telgarsky.