EVALUATING THE LEARNABILITY OF K-MEANS CLUSTERING

Yuxuan Li '17
Faculty Sponsor: Takunari Miyazaki

Recent years have witnessed increased interest in research and the application of machine learning, where models update themselves without being explicitly instructed. Machine learning models are widely used in tasks such as email spam detection, recommender systems, and speech translation. Promising results were shown even for complicated tasks such as image recognition and video recognition. However, the credit is largely due to supervised learning, where models are trained with known labels. What happens when models are trained without any prior knowledge? This is known as unsupervised learning, which does not allow for the tuning of the representation by consulting known labels. The current project investigated this issue of learnability in a pure unsupervised learning setting. In particular, we focused on *k*-means clustering as a representative algorithm. *k*-means groups data instances into k clusters so that each data instance falls into the cluster with the nearest centroid based on the Euclidean distances across the attributes. Using datasets from the UCI machine learning repository and Python's machine learning library Scikit-learn, *k*-means' performances on a variety of datasets (n=40) were evaluated with Adjusted Rand Index (ARI). ARI scores range from -1 to 1, with 1 indicating identical clustering assignments and 0 indicating randomly matched clustering assignments. We found that nearly half of the datasets resulted in an ARI score around 0. For datasets whose scores were above 0, *k*-means did not generate the optimal performance at the optimal k, i.e., the ground-truth number of clusters in each dataset. Rather, the optimal performance scattered in k values deviating from the optimal k. These results suggested that the reasonable partition from *k*-means clustering does not adhere to the desirable outcome, revealing the inherent difficulty of pure unsupervised learning.