Abstract:
Many data sets have missing values in practical application contexts, but the majority of commonly studied machine learning or statistical methods cannot be applied directly when there are incomplete samples. However, most such methods only depend on the relative differences between samples instead of their particular values, and thus one useful approach is to directly estimate the pairwise distances between all samples in the data set. This can be accomplished by fitting a Gaussian mixture model to the data, and using it to derive estimates for the distances. In the first part of the talk, a variant of such model for high-dimensional data with missing values will be presented. Experimental simulations confirm that the proposed method provides accurate estimates compared to alternative methods for estimating distances. In particular, using the mixture model for estimating distances is on average more accurate than using the same model to impute any missing values and then calculating distances. The experimental evaluation additionally shows that more accurately estimating distances lead to improved prediction performance for classification and regression tasks when used as inputs for a neural network. In the second part of the talk, visualization of datasets with missing values will Big Data will be introduced.