Machine, Data and Learning (CS7.301)

Spring 2022, IIIT Hyderabad
14 Mar, Monday (Lecture 14)

Taught by Prof. Vikram Pudi

Overview of Data Analytics (contd.)

Measurement of Similarity

There are different measures of similarity depending on the type of variable we are considering.

Categorical or nominal variables use the metric

$$d(x, y) = 1 - \frac{m}{n},$$

where m is the sum of weights of matching attributes, and n is the sum of weights of all attributes.

Numeric variables have several possible functions – euclidean distance, Manhattan distance, and so on.

Clustering (contd.)

A partition can be evaluated based on square error. That is, if there are N clusters, we can find the square error as

$$d = \sum_{i=1}^{N} \sum_{x \in C_i} d(x, m_i).$$

There are also hierarchical methods (agglomerative and divisive) of clustering.