Derek Gloudemans

Advanced Machine Learning

CS 6362-01

Clustering Written Portion

11/18/2019

1. Prove analytically that the minimum of the sum of intra-cluster distances from cluster mean is non-increasing with increasing K.

Let equal the number of clusters, and equal the minimum sum of intra-cluster distances for a clustering of data into clusters. Consider any cluster in the clustering. Let be the sum of intra-cluster distances for cluster . If are the examples that are grouped into cluster , then can be written as:

Let be the example grouped into cluster with the largest distance from the cluster mean. Note that this distance must be greater than or equal to 0. Consider the clusters resulting from partitioning cluster into and . The intra-cluster distance for the cluster with only one element is 0, and the intra-cluster distance for the cluster with – 1 elements is at most equal to Furthermore, it is strictly less than unless every element in the cluster is exactly equal to the centroid.

Since the and all other intra-cluster distances remain unchanged, the minimum sum of intra-cluster distances for this clustering cannot be larger than the sum of intra-cluster distances before making this partition. **Thus, since there exists at least one possible value for and the minimum of all possible values for is selected, cannot be greater than so is non-increasing.**

1. Find the update equations for Kernel K-means clustering.

**Assign examples to clusters:**

In normal K-means, the cluster assignment of example xi to cluster Sk is accomplished by the following formula:

In kernel K-means clustering, this corresponds to:

We know:

Essentially what it says is that each example is assigned to the cluster for which the sum of the distances between the features for that example and the features for all other examples in the cluster (in the transformed feature-space) is minimized. Theoretically, using the kernel trick precludes having to convert each example into the transformed feature space explicitly.

**Update Cluster Centroids:**

using the updated cluster memberships for each