Christian Felt and Jadon Wagstaff

Data Mining Spring 2017

Instructor: Jeff Phillips

February 15, 2017

Data Collection Report

1. I downloaded the data from the IUCN Red List website at <http://www.iucnredlist.org/technical-documents/spatial-data>.

2. The data is 74,526 19-dimensional points, representing 38,186 mammals, 18,693 amphibians, and 17,647 reptiles. I have left out ocean-dwelling species and birds because their ranges are less discrete.

3. I am thinking of the data as a matrix with 74,526 rows and 19 columns. However, many of these columns/dimensions will be pruned away when I discover which features of the data are most interesting for the purposes of clustering. The file format is .csv, with each data point occupying a different line, and each value within each data point separated by a comma. I will read this data using a simple Python program.

4. The data came originally as three large Esri shapefiles (about 300 MB each), containing polygons representing the geographical range of each species and 15 other fields with information such as scientific name, whether the species is currently present in the area, whether it is native or invasive, etc. Using the ArcMap software, I computed centroids for each polygon in a way that makes sure the centroid is located inside the polygon, added these centroids as XY (latitude/longitude) points to the map, and exported the map data to text files, whose total size is only 15 MB.

5. Similar data could be simulated by generating random latitude and longitude coordinates, limiting their possible values to areas where there is land rather than ocean, and making certain locations more likely than others so that clusters appear. Values for the other fields, such as whether the species is native or invasive, present or extinct, could also be generated randomly, most easily if the value for every field is represented using an integer (for instance, by hashing the strings for a species’ scientific name.)