

Data Mining
Using the EM Clustering Algorithm
on
Places Rated Almanac Data

Jim Rogers
April 23, 2001
INFT

979

TABLE OF CONTENTS

	<u>Pages</u>	
Background	3	
Purpose	3	
Expectation Maximization (EM) Algorithm	3-5	
Data	5-6	
Data Cleaning and Preparation	6	
Clustering Trials	6-8	
Clustering Results	9-22	
Summary	22-23	
References	24	
Appendices:		
EM output using all attributes	25-30	
EM output using one attribute (Climature & Terrain)	31-33	
EM output using two attributes (Climature & Terrain with Housing)	34-36	

BACKGROUND

This dataset is from the Places Rated Almanac, by Richard Boyer and David Savageau, copyrighted and published by Rand Mc Nally. The dataset analyzed for this project is from 1985. The Places Rated Almanac updates this data regularly at approximately three year intervals.

The Places Rated Almanac is a guide to finding the best places to live in the United States. The Places Rated Almanac assigns scores to characteristics/attributes of cities. They assign these scores to all U.S. metropolitan areas with a population greater than 50,000 people. In 1985, there were 329 metropolitan areas with a population greater than 50,000 people. 80% of the entire U.S. population lived in these 329 metropolitan areas.

PURPOSE

The purpose of this project is to determine if clustering software can be effectively utilized to analyze the places rated data. The goal is to perform exploratory data analysis on the data, and see if the clusters that are generated are meaningful.

In regards to the places rated data, (1) determine if the geographic regions within the U.S. exhibit clusters/patterns in regards to the data, (2) determine if there is a relationship between the different ratings/attributes, (3) identify how many locations fall into each cluster, (4) explore how the different ratings/attributes of the cities will cluster, and (5) attempt to divide the data into natural groups.

EXPECTATION MAXIMIZATION (EM) ALGORITHM

The EM algorithm, implemented in the Waikato Environment for Knowledge Analysis (Weka), was utilized to perform the clustering on the data. Weka is a system developed at the University of Waikato in New Zealand. Weka is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and Weka has been tested under Linux, Windows, and Macintosh operating systems. Java provides a uniform interface to many different learning algorithms, along with methods for

pre- and postprocessing and for evaluating the result of learning schemes on any given dataset.

EM is a mixture based algorithm that attempts to maximize the likelihood of the model. EM models the distribution of instances probabilistically, so that an instance belongs to a group with a certain probability. EM does not actually calculate probabilities, instead it calculates densities. The assumption is made by EM that the attributes are independent random variables. EM can handle both numeric and nominal attributes. The first step, calculation of the cluster probabilities (which are the "expected" class values) is expectation; the second step is calculation of the distribution parameters, is "maximization" of the likelihood of the distributions given the data.

The output of EM is the number of clusters printed first, followed by a description of each cluster, the cluster's prior probability, and a probability distribution for all attributes in the dataset. For a nominal attribute, the distribution is represented by the count associated with each value (plus one); for a numeric attribute it is a standard normal distribution. EM also outputs the number of training instances in each cluster, and the log-likelihood of the training data with respect to the clustering that it generates.

By default, EM selects the number of clusters automatically by maximizing the logarithm of the likelihood of future data, estimated using cross-validation. Beginning with one cluster, it continues to add clusters until the estimated log-likelihood decreases.

The foundation for statistical clustering is a statistical model called finite mixtures. A mixture is a set of k probability distributions, representing k clusters, that govern the attribute values for members of that cluster. Each distribution gives the probability that a particular instance would have a certain set of attribute values if it were known to be a member of that cluster. Each cluster has a different distribution. Any particular instance belongs to one and only one of the clusters, but it is not known which one. The clusters are not equally likely. There is some probability distribution that reflects their relative populations. The mixture model combines several normal distributions, and its probability density function

looks like a mountain range with a peak for each of the components.

DATA

The data analyzed by EM are rating criteria used by the Places Rated Almanac to rate characteristics/attributes of U.S. metropolitan areas. The rating criteria are Climate & Terrain, Health Care & Environment, Crime, Transportation, Education, The Arts, Recreation, and Economics. For all but two of the above criteria, the higher score is better. For Housing and Crime, a lower score is better. The scores are computed using component statistics for each criterion.

Climate & Terrain: very hot and very cold months, seasonal temperature variation, heating and cooling degree days, freezing days, zero degree days, and ninety degree days.

Housing: utility bills, property taxes, and mortgage payments.

Health Care & Environment: per capita physicians, teaching hospitals, medical schools, cardiac rehabilitation centers, comprehensive cancer treatment centers, hospices, insurance/hospitalization costs index, fluoridation of drinking water, and air pollution.

Crime: violent crime rate and property crime rate.

Transportation: daily commute, public transportation, interstate highways, air service, and passenger rail service.

Education: pupil/teacher ratio in the public K-12 system, effort index in K-12, and academic options in higher education.

The Arts: museums, fine arts and public radio stations, public television stations, universities offering a degree or degrees in the arts, symphony orchestras, theatres, opera companies, dance companies, and public libraries.

Recreation: good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, aquariums, family theme parks, sanctioned automobile race tracks, pari-mutuel betting attractions, major and minor

league professional sports teams, NCAA Division I football and basketball teams, miles of ocean or Great Lakes coastline, inland water, national forests, national parks, or national wildlife refuges, and Consolidated Metropolitan Statistical Area access.

Economics: average household income adjusted for taxes and living costs, income growth, and job growth.

The data are recorded in an ASCII file. The file contains 329 observations plus an index column. The index is for the location. All data are numeric except for the index which is nominal.

DATA CLEANING AND PREPARATION

The actual dataset was obtain from `lib.stat.cmu.edu/datasets/places.data`

The fields in the database were tab separated. There was no missing data, since all the fields were filled.

To be processed by the EM algorithm in Weka, data must be in the .arff format. The .arff format uses commas to separate fields and places the @ symbol in front of the relation name and each attribute. The line preceding the data is @DATA.

The database was opened in Excel, and the values in the columns were adjusted so that the first value in each column lined up perfectly. The data was then saved in a .csv format which is a format where commas are placed between values in adjacent columns. The database was then opened in Word, header information added, and the file extension changed to .arff prior to saving.

CLUSTERING TRIALS

To try and identify geographic regions within a cluster, I modified the nominal place attribute/index. Each city was indexed in alphabetical order in the original database from 1 to 329. I separated the mainland U.S. into 16 geographical regions, and added one region each for Hawaii and Alaska. I then went through the database index and changed their index to my geographical index. I looked at

each city in the database and assigned to a geographical index. So, my index went from 1 to 18.

Place index 1 includes all cities in Maine, New Hampshire, and Vermont.

Place index 2 includes all cities in Connecticut, Massachusetts, and Rhode Island.

Place index 3 includes all cities in New York, Pennsylvania, and New Jersey.

Place index 4 includes all cities in Maryland, West Virginia, Delaware, and Washington, DC.

Place index 5 includes all cities in North Carolina, Virginia, and South Carolina.

Place index 6 includes all cities in Florida, Georgia, and Alabama.

Place index 7 includes all cities in Louisiana, Mississippi, and Tennessee.

Place index 8 includes all cities in Kentucky, Ohio, and Indiana.

Place index 9 includes all cities in Michigan, Wisconsin, and Minnesota.

Place index 10 includes all cities in Illinois, Iowa, and Missouri.

Place index 11 includes all cities in Kansas, Nebraska, and Oklahoma.

Place index 12 includes all cities in Arkansas, New Mexico, and Texas.

Place index 13 includes all cities in North Dakota, South Dakota, and Montana.

Place index 14 includes all cities in Wyoming, Utah, and Colorado.

Place index 15 includes all cities in Washington, Oregon, and Idaho.

Place index 16 includes all cities in California, Nevada, and Arizona.

Place index 17 includes all cities in Hawaii.

Place index 18 includes all cities in Alaska.

The first clustering trial was run on all the attributes including the nominal place attribute with geographical index.

The second series of clustering trials was run on individual attributes with the place attribute (e.g., crime and place, The Arts and place). So, the second series of trials resulted in a total of nine trials, with one for each characteristic/attribute.

The third series of clustering trials was run on pairs of attributes and the place attribute (e.g., crime, recreation, and place, Economics, Education, and place). This third series of trials resulted in a total of thirty-six trials, with one for each unique pairing of characteristics/attributes.

In some practice trials, I removed the place attribute from the input, and this did not impact the number of clusters, the size of the clusters, or the mean of the clusters. Therefore, utilization of the place attribute, which was the only nominal attribute, did not effect the clustering results.

The output of the running the EM algorithm within Weka on the dataset includes: (1) the type of data mining performed (e.g. clusters), (2) the name of the relation, (3) the number of instances, (4) the number of attributes, (5) the names of the attributes, (6) the number of clusters selected by cross validation, (7) the mean and standard deviation of the normal distribution of each cluster that is generated, (8) the counts for each possible value of the nominal attribute, (9) the number of clustered instances in each cluster, (10) the percentage of instances in each cluster, (11) and the log likelihood.

CLUSTERING RESULTS

ALL ATTRIBUTES

When the EM clustering algorithm was run on all the attributes of the Places Rated database, six clusters were generated. There was good separation between the mean values of each of the six clusters.

One of the six clusters had the best (the highest mean value for the attribute in any of the clusters) Health Care and Environment, Transportation, Education, Arts, and Recreation. This was the smallest of the six clusters with only 23 of the 329 cities (7%) in this cluster. The highest percentage of cities in this cluster were from the eastern mid-atlantic region from New York to Washington, DC.

One of the other six clusters had the worst (the lowest mean value for the attribute in any of the clusters) Climate and Terrain, Health Care and Environment, Transportation, Education, and Arts. 54 of the 329 cities (16%) were in this cluster. The highest percentage of cities in this cluster were from the southern U.S. ranging from Florida to New Mexico.

The largest of the six clusters had 87 of the 329 cities (26%) in the cluster. The highest percentage of cities in this cluster were primarily from the northern states of Ohio, Indiana, Michigan, Wisconsin, Minnesota, Illinois, and Iowa. This cluster had the best crime and the worst recreation and the worst economics.

In one of the six clusters, all its attributes were close to a medium value. This cluster had very few cities from the western states.

INDIVIDUAL ATTRIBUTES

Climate and Terrain

When the EM clustering algorithm was run on only the Climate and Terrain attribute with the place index, six clusters were generated with good separation between the mean value of each cluster. The best cluster, the one with the highest mean value for Climate and Terrain, had 19 of 329 cities (6%) in the cluster. The highest percentage of

cities in this cluster were from the western states of Washington, Oregon, Idaho, California, Nevada and Arizona.

The worst cluster, the one with the lowest mean value for Climate and Terrain, had 16 of 329 cities (5%) in the cluster. The cities in this cluster were primarily from the southeast states (Florida, Georgia, and Alabama), the northern states (Michigan, Wisconsin, Minnesota, North Dakota, South Dakota, and Montana), and Alaska. The cities in the southeast states were probably in this cluster because their temperatures are too hot, and the other cities in this cluster were probably in this cluster because their temperatures are too cold.

The second best cluster had only 4 of 329 cities (1%) in the cluster. Those four cities were Honolulu, Hawaii, Galveson, Texas, Olympia, Washington, and Salem, Oregon.

Most of the cities, 290 of 329 (88%) fell in the clusters with medium mean values.

Housing

When the EM algorithm was run on the Housing attribute with the Place index, four clusters were formed with good separation between the mean values of those four clusters.

The best cluster had 147 of the 329 cities (45%) in the cluster. The cities in the best housing cluster were primarily from the eastern and central time zones excluding cities from the New England states.

The worst cluster had 33 of the 329 cities (10%) in the cluster. The cities in the worst housing cluster were primarily from the southwestern states of California, Nevada, and Arizona, Hawaii, and Alaska. The number of cities in the worst housing cluster were much less than the number of cities in the best housing cluster.

Health Care and Environment

When the EM algorithm was run on the Health Care and Environment attribute with the Place index, four clusters were formed with very good separation between the mean values of the clusters.

The best cluster had 16 of the 329 cities (5%) in the cluster. The cities in this cluster were Atlanta, Georgia; Baltimore, Maryland; Boston, Massachusetts; Chicago, Illinois; Cleveland, Ohio; Detroit, Michigan; Los Angeles, California; Minneapolis-St. Paul, Minnesota; Nassua-Suffolk, New York; New York, New York; Newark, New Jersey; Philadelphia, Pennsylvania; Pittsburgh, Pennsylvania; Raleigh-Durham, North Carolina; San Francisco, California; and Washington, DC. These cities are very large metropolitan areas with large hospitals, specialized treatment facilities, and many physicians. All of these cities are eastern half of the country except for Los Angeles and San Francisco.

The two clusters with the lowest mean Health Care and Environment scores contained 240 of the 329 cities (73%), and the worst cluster of these two had 112 of the 329 cities (34%) in its cluster. The cities in the worst cluster were primarily from the southern states of Florida, Alabama, Louisiana, Mississippi, Tennessee, Arkansas, Texas, and New Mexico; the northwestern states of Wyoming, Utah, Colorado, Washington, Oregon, and Idaho; and Alaska.

Crime

When the EM algorithm was run on the Crime attribute with the Place index, three clusters were generated with good separation between the mean of each cluster. There were an equal number of cities in the best cluster and the worst cluster. For the three clusters, 25% of the cities were in the best cluster, 25% of the cities were in the worst cluster, and 50% of the cities were in the middle cluster.

The cities in the best cluster were primarily from the northern states of Maine, New Hampshire, Vermont, New York, Pennsylvania, New Jersey, Michigan, Wisconsin, Minnesota, North Dakota, South Dakota, and Montana. The cities in the worst cluster were primarily from the southern states of Florida and Georgia, the southwestern states of Texas, New Mexico and Arizona, and Alaska.

Transportation

When the EM algorithm was run on the Transportation attribute with the Place index, only two clusters were generated. There was a large separation between the mean values of these two clusters.

There were an approximately equal number of cities in the best cluster and the worst cluster. There were 158/329 cities (48%) in the best cluster and 171/329 cities (52%) in the worst cluster. Cities in the best cluster were spread rather evenly across the U.S., but included very few cities from the northeastern states. The cities with the very best transportation were from Michigan, Wisconsin, and Minnesota. The cities in the worst cluster were primarily from the southern states of Florida, Alabama, Arkansas, Texas and New Mexico.

Education

When the EM algorithm was run on the Education attribute with the Place index, three clusters were formed with good separation between the mean values of the three clusters.

There was almost an equal number of cities in the best cluster and the worst cluster. There were 69/329 cities (21%) in the best cluster, 57/329 cities (17%) in the worst cluster, and 203/329 cities (62%) in the middle cluster. Cities in the best cluster were primarily from the eastern U.S., and cities in the worst cluster were primarily from the western U.S., Hawaii, and Alaska.

The Arts

When the EM algorithm was run on the Arts attribute with the Place index, three clusters were formed with a very large separation between the mean values of the three clusters.

The best cluster was extremely small with only 8 of 329 cities (2%) in the cluster. These 8 cities were Baltimore, Maryland; Boston, Massachusetts; Chicago, Illinois; Los Angeles, California; New York, New York; Philadelphia, Pennsylvania; San Francisco, California; and Washington, DC. These are all very large urban metropolitan areas with several museums, television stations, and symphony orchestras.

The worst cluster was the largest of the three clusters with 216 of the 329 cities (66%) in the cluster. Cities from all over the U.S. were included in this cluster.

The middle cluster had 105 of the 329 cities (32%) in the cluster. Cities in the midwestern states of Kansas, Nebraska, and Oklahoma; the far western states of California, Nevada, and Arizona; and Hawaii had more cities in the middle cluster than the worst cluster.

Recreation

When the EM algorithm was run on the Recreation attribute with the Place index, three clusters were formed with good separation between the mean values of the clusters. Each of the best and worst clusters had 78 of the 329 cities (24%) in its cluster. There were 173 of the 329 cities (52%) in the middle cluster.

Cities in the best cluster were primarily from the western U.S., Hawaii, Alaska, and Florida. Cities in Florida had the highest recreation scores.

Cities in the worst cluster were primarily from the eastern mid-Atlantic states and southwestern states. Cities in West Virginia had the lowest recreation scores.

Economics

When the EM algorithm was run on the Economics attribute with the Place index, three clusters were formed with a good separation between the mean values of the clusters. There were almost an equal number of cities in the best cluster and the worst cluster. There were 83 of 329 cities (25%) in the best cluster, 91 of 329 cities (28%) in the worst cluster, and 155 of 329 cities (47%) in the middle cluster.

Cities in the best cluster were primarily from the New England states, southeast states, and Midwest states. Cities in the worst cluster were mostly from the northern states of New York, Pennsylvania, New Jersey, Ohio, Indiana, Michigan, Wisconsin, Minnesota, Washington, and Idaho.

TWO ATTRIBUTES PLUS THE PLACE INDEX

When the EM algorithm was run on the Climate & Terrain attribute and the Housing attribute, 60% of the cities were clustered with a good mean housing rating and medium mean climate & terrain rating. 7% of the cities were clustered with the best mean climate & terrain rating and the worst mean housing rating. These 7% of the cities were primarily from the western states and Hawaii.

When the EM algorithm was run on the Climate & Terrain attribute and the Health Care & Environment attribute, 22 of the 329 cities (7%) had the best mean climate & terrain rating and the best mean health care & environment rating. These 7% of the cities were primarily from New York, New Jersey, and Pennsylvania. 129 of the 329 cities (39%) had a poor mean climate & terrain rating and the worst mean health care & environment rating. These 39% of the cities were primarily from the northern New England states of Maine, New Hampshire, and Vermont; the southern states of Florida, Georgia, Alabama, Louisiana, Mississippi, Tennessee, Arkansas, New Mexico and Texas; and states in the mountain time zone states of Wyoming, Utah, and Colorado.

When the EM algorithm was run on the Climate & Terrain attribute and the Crime attribute, 18 of the 329 cities (5%) had the worst mean climate & terrain rating and the best mean crime rating. These 5% of the cities were from the traditionally colder areas of the U.S. which are the northern states (Michigan, Wisconsin, Minnesota, North Dakota, South Dakota, and Montana) and Alaska. 24 of the 329 cities (7%) had the best climate & terrain rating and a poor mean crime rating. These 7% of the cities were from some of the traditionally warmer areas of the U.S. which are the western states (California, Nevada, and Arizona) and Hawaii. 232 of the 329 cities (71%) fell in the middle cluster and had a medium mean climate & terrain rating and a medium mean crime rating.

When the EM algorithm was run on the Climate & Terrain attribute and the Transportation attribute, only 33 of the 329 cities (12%) were in a cluster that had the best mean climate & terrain rating with the best mean transportation rating. These 12% of the cities were primarily from the western states and Hawaii. 138 of the 329 cities (42%)

were in a cluster that had the worst mean climate and terrain rating with the worst mean transportation rating. These 42% of the cities were primarily from the northeastern section of New England, the southeastern states, and the midwestern states.

When the EM algorithm was run on the Climate & Terrain attribute and the Education attribute, only 9 of the 329 cities (3%) were in a cluster that had the worst mean climate & terrain rating and the worst mean education rating. These 3% of the cities were from the cold weather states in the northern U.S. and Alaska. 36 of the 329 cities (11%) were in a cluster that had the best mean climate & terrain rating with a poor mean education rating. These 11% of the cities were primarily from the western states.

When the EM algorithm was run on the Climate & Terrain attribute and the Arts attribute, eight clusters were formed. Only 8 of the 329 cities (2%) were in a cluster that had the best mean arts rating and a medium mean climate and terrain rating. These 8 cities were all major metropolitan area. 63 of the 329 cities (19%) were in a cluster with the worst mean arts rating and a medium mean climate and terrain rating. Only 7 of the 329 cities (2%) were in a cluster with the worst mean climate and terrain rating and a low mean arts rating. 21 of the 329 cities (6%) were in a cluster with the best mean climate and terrain rating and a medium mean arts rating.

When the EM algorithm was run on the Climate & Terrain attribute and the Recreation attribute, many of the cities (239 of 329 cities = 73%) were in clusters with a poor mean recreation rating and a medium mean climate & terrain rating.

When the EM algorithm was run on the Climate & Terrain attribute and the Economics attribute, the worst mean climate & terrain rating clustered with the worst mean economics rating. Several cities were in this cluster, and these cities in this cluster were primarily from northeast New England states of Maine, New Hampshire, and Vermont; the southern states of Florida, Georgia, Alabama, Louisiana, Mississippi, Tennessee, Texas, New Mexico, and Arkansas; the midwestern states of Kansas, Nebraska, and Oklahoma; the Rocky Mountain states of Wyoming, Utah, and Colorado; and Alaska.

When the EM algorithm was run on the Housing attribute and the Health Care & Environment attribute, the worst mean housing rating clustered with the best mean health care & environment rating. Cities in this cluster were primarily from the states of California, Nevada, Arizona, Hawaii, and Alaska.

When the EM algorithm was run on the Housing attribute and the Crime attribute, 48% of the cities were clustered with the best mean housing rating and a medium mean crime rating. These cities were primarily from the southeastern states.

When the EM algorithm was run on the Housing attribute and the Transportation attribute, six clusters were formed. The largest of these six clusters, with 25% of the cities, had the best mean housing rating and the worst mean transportation rating. Cities in this largest cluster were mostly from the southeastern states. A cluster with the worst mean housing rating and a very good mean transportation rating was formed. Cities in this cluster were from New York, Pennsylvania, New Jersey, California, Nevada, Arizona, Hawaii, and Alaska.

When the EM algorithm was run on the Housing attribute and the Education attribute, the worst mean housing rating clustered with the best mean education rating. 10% of the cities were in this cluster, and they were mostly from California, Nevada, Arizona, Hawaii, and Alaska. 59% of the cities were in a cluster with the best mean housing attribute clustered with the worst mean education rating. Cities in this cluster were spread out across the U.S., except from the northeastern U.S., western U.S., Hawaii, and Alaska.

When the EM algorithm was run on the Housing attribute and the Arts attribute, 7 of the 329 cities (2%) were in a cluster with the worst mean housing cluster and the best mean Arts rating. These 7 cities were Boston, Massachusetts; New York, New York; Philadelphia, Pennsylvania; Washington, DC; Chicago, Illinois; Los Angeles, California; and San Francisco, California. The largest cluster, with 103 of the 329 cities (31%), had the best mean housing rating clustered with the worst Arts rating. Cities in this cluster were mainly from the southeastern states.

When the EM algorithm was run on the Housing attribute and the Recreation attribute, 24% of the cities were in a cluster with the best mean housing rating and the worst mean recreation rating. Cities in this cluster were primarily from the southeastern states.

When the EM algorithm was run on the Housing attribute and the Economics attribute, 7 clusters were formed. 66 of the 329 cities (20%) were in a cluster with the worst mean economics rating and a good mean housing rating. 34 of the 329 cities (10%) were in a cluster with the best mean economics rating and a medium mean housing rating. 31 of the 329 cities (9%) were in a cluster with the worst mean housing rating and a medium mean economics rating. The largest cluster, with 83 of the 329 cities (25%) in the cluster, was formed with the best mean housing rating and a medium mean economics rating.

When the EM algorithm was run on the Health Care & Environment attribute and the Crime attribute, 15 of the 329 cities (5%) were in a cluster with the best mean health care and environment rating and the worst mean crime rating. Cities in this cluster were the very large metropolitan areas.

When the EM algorithm was run on the Health Care & Environment attribute and the Transportation attribute, 115 of the 329 cities (35%) were in a cluster with the worst mean health care & environment rating and the worst mean transportation rating. Cities in this cluster were mainly from the southeastern states and the western states. 19 of the 329 cities (6%) were in a cluster with the best mean health care & environment rating and the best mean transportation rating. Cities in this cluster were the large U.S. metropolitan urban areas that have large medical facilities, many physicians, many interstate highways, and much public transportation.

When the EM algorithm was run on the Health Care & Environment attribute and the Education attribute, 102 of the 329 cities (32%) were in a cluster with the worst mean health care & environment rating and the worst mean education rating. Cities in this cluster were mainly from the southern states, western states, and Alaska. 33 of the 329 cities (10%) were in a cluster with the best mean health care & environment rating clustered with the best

mean education rating. Cities in this small cluster were mainly from the states of New York, New Jersey, and Pennsylvania.

When the EM algorithm was run on the Health Care & Environment attribute and the Arts attribute, 99 of the 329 cities (30%) were in a cluster with the worst mean health care & environment rating and the worst mean Arts rating. Cities in this cluster, which was the largest of the five clusters involving the health care & environment attribute and the arts attribute, were mainly from the southern states and the western states. A very small cluster was formed with only 11 of the 329 cities (3%) in the cluster. This very small cluster had the best mean health care & environment rating and the best mean Arts rating. Cities in this very small cluster were from the large metropolitan areas of Boston, Massachusetts; New York, New York; Philadelphia, Pennsylvania; Newark, New Jersey; Washington, DC; Baltimore, Maryland; Detroit, Michigan; Minneapolis-St. Paul, Minnesota; Chicago, Illinois; Los Angeles, California; and San Francisco, California.

When the EM algorithm was run on the Health Care & Environment attribute and the Recreation attribute, 93 of the 329 cities (28%) were in a cluster with the worst mean health care & environment rating and the worst mean recreation rating. Cities in this cluster were mainly from the southern states excluding Florida, and the northwest states. A small cluster was formed with only 22 of the 329 cities (7%) in the cluster. This small cluster had the best mean health care & environment rating and the best mean recreation rating. Cities in this small cluster were from the large metropolitan areas that have large and specialized medical facilities and several professional sports teams.

When the EM algorithm was run on the Health Care & Environment attribute and the Economics attribute, five clusters were formed. 85 of the 329 cities (26%) were in a cluster with the worst mean economics rating and a low mean health care & environment rating. The largest cluster, with 103 of the 329 cities (31%) in the cluster, was formed with the worst mean health care & environment rating and a medium mean economics rating. 36 of the 329 cities were in a cluster with the best mean economics rating and a medium mean health care & environment rating. The smallest cluster, with only 21 of the 329 cities (6%) in the

cluster, was formed with the best mean health care & environment rating and a medium mean economics rating.

When the EM algorithm was run on the Crime attribute and the Transportation attribute, a very small cluster with only 9 of the 329 cities (3%) was formed. This very small cluster had the worst mean crime rating and the best mean transportation rating. These cities had high crime rates, many interstate highways, and availability to public transportation. These 9 cities were Boston, Massachusetts; New York, New York; Philadelphia, Pennsylvania; Baltimore, Maryland; Washington, DC; Atlanta, Georgia; Chicago, Illinois; Los Angeles, California; and San Francisco, California.

When the EM algorithm was run on the Crime attribute and the Education attribute, a cluster was formed with the worst mean crime rating and the best mean education rating. 93 of the 329 cities (28%) were in this cluster that had the worst crime clustered with the best education. Cities in this cluster were from the southern U.S., California, and Nevada.

When the EM algorithm was run on the Crime attribute and the Arts attribute, a very small cluster was formed with worst mean crime rating and the best mean arts rating. This very small cluster had 15 of the 329 cities in the cluster. Cities in this cluster were major metropolitan areas with high crime rates, several museums, television stations, and symphonies. Cities in this cluster were Boston, Massachusetts; New York, New York; Newark, New Jersey; Philadelphia, Pennsylvania; Washington, DC; Baltimore, Maryland; Atlanta, Georgia; Detroit, Michigan; Minneapolis-St. Paul, Minnesota; Chicago, Illinois; St. Louis, Missouri; Dallas, Texas; Los Angeles, California; San Francisco, California; and Seattle, Washington.

When the EM algorithm was run on the Crime attribute and the Recreation attribute, a cluster was formed with the worst mean crime rating and the best mean recreation rating. 90 of the 329 cities (27%) were in this cluster. Cities in this cluster were mainly from Florida, Georgia, Utah, Colorado, California, Nevada, Hawaii, and Alaska.

When the EM algorithm was run on the Crime attribute and the Economics attribute, four clusters were formed. The largest cluster, with 131 of 329 cities (40%) in the

cluster, was formed with the worst mean economics rating and a medium mean crime rating. The smallest cluster, with 51 of the 329 cities (16%) in the cluster, was formed with the worst mean crime rating and a medium mean economics rating.

When the EM algorithm was run on the Transportation attribute and the Education attribute, only two clusters were formed. The worst mean transportation rating clustered with the worst mean education rating, and almost all of the cities (281 of the 329 cities = 85%) were in this cluster. In the other cluster, the best mean transportation rating clustered with the best mean education rating. There were 48 of the 329 cities (15%) in this cluster, and these cities were primarily from the New England states.

When the EM algorithm was run on the Transportation attribute and the Arts attribute, one of the clusters was formed with the worst mean transportation rating and the worst mean Arts rating. 67 of the 329 cities (20%) of the cities were in this cluster. Another cluster was formed with the best mean transportation rating clustered with the best mean arts rating. This was a very small cluster with only 9 of the 329 cities (3%) in this cluster. Cities in this cluster were large urban metropolitan areas with good public transportation, many interstates, several museums, and several television stations. These 9 cities were Boston, Massachusetts; Philadelphia, Pennsylvania; New York, New York; Newark, New Jersey; Baltimore, Maryland; Washington, DC; Chicago, Illinois; Los Angeles, California; and San Francisco, California.

When the EM algorithm was run on the Transportation attribute and the Recreation attribute, one of the clusters was formed with the worst mean transportation rating and the worst mean recreation rating. 72 of the 329 cities (22%) were in this cluster. Almost all the cities in Arkansas and New Mexico were in this cluster. Another cluster was formed with the best mean transportation rating and the best mean recreation rating. 75 of the 329 cities (23%) were in this cluster. Cities in this cluster were primarily from Florida, Georgia, Utah, Colorado, California, and Hawaii.

When the EM algorithm was run on the Transportation attribute and the Economics attribute, one of the clusters

was formed with the best mean transportation rating and the best mean economics rating. Only 7 of the 329 cities (2%) were in this very small cluster, and these cities were in the New England area.

When the EM algorithm was run on the Education attribute and the Arts attribute, 48% of the cities were in a cluster that had a medium mean education rating clustered with the worst mean arts rating. Another cluster was very small. This very small cluster, with only 7 of the 329 cities (2%) in the cluster, had the best mean education rating clustered with the best mean arts rating. These cities were all major urban metropolitan areas. Cities in this very small cluster were Boston, Massachusetts; New York, New York; Philadelphia, Pennsylvania; Washington, DC; Chicago, Illinois; Los Angeles, California; and San Francisco, California.

When the EM algorithm was run on the Education attribute and the Recreation attribute, 69 of the 329 cities (21%) were in a cluster that was formed with the worst mean education rating and the best mean recreation rating. Cities in this cluster were primarily from Florida, Georgia, Alabama, Wyoming, Utah, Colorado, Washington, Oregon, Idaho, California, Nevada, Arizona, Hawaii, and Alaska.

When the EM algorithm was run on the Education attribute and the Economics attribute, four clusters were formed. There were an equal number of cities, 94 of the 329 cities (29%), in the clusters with the worst mean education rating clustered with a medium mean economics rating, and the worst mean economics rating clustered with a medium mean education rating.

When the EM algorithm was run on the Arts attribute and the Recreation attribute, 50 of the 329 cities (15%) were in a cluster that was formed with the worst mean arts rating and the worst mean recreation meaning. Most of the cities in this cluster were small in size and population. Another cluster was formed with the best mean arts rating clustering with the best mean recreation rating. This was a very small cluster with only 9 of the 329 cities in this cluster. Cities in this cluster were all major metropolitan areas. The cities in this cluster were Boston, Massachusetts; New York, New York; Philadelphia, Pennsylvania; Newark, New Jersey; Washington, DC; Chicago,

Illinois; Los Angeles, California; San Francisco, California; and Seattle, Washington.

When the EM algorithm was run on the Arts attribute and the Economics attribute, 145 of the 329 cities (44%) were in a cluster that was formed with a poor mean arts rating and the worst mean economy rating. Another cluster was formed with the worst mean arts rating clustered with the best mean economy rating. This cluster had 76 of the 329 cities (23%) in the cluster. Cities in this cluster were primarily from Maine, New Hampshire, Vermont, Florida, Georgia, Alabama, Arkansas, New Mexico, and Texas.

When the EM algorithm was run on the Recreation attribute and the Economics attribute, 64 of the 329 cities (19%) were in a cluster with the worst mean recreation rating clustered with the worst mean economics rating. Cities in this cluster were from Maryland, Delaware, West Virginia, Virginia, North Carolina, and South Carolina.

SUMMARY

Using the EM algorithm for exploratory analysis on the Places Rated Almanac data, produced clusters with good separation between the mean values of the clusters. For the 329 cities in the database, the number of clusters generated varied between two and six regardless of the number of attributes included in the clustering. The option with the EM algorithm, to set the maximum number of clusters, did not have to be utilized. The elapsed time to generate the clusters was between one and four minutes. The elapsed time appeared to be based on the number of clusters generated, because it took longer to generate the larger clusters than it took to generate the smaller clusters.

The largest metropolitan urban areas in the U.S. were in the clusters that had the best mean Health Care & Environment, Transportation, Education, Arts, and Recreation. These urban areas were approximately only 2% of all the cities. The cities with the worst Climate & Terrain, Health Care & Environment, Transportation, Education, and Arts always clustered together and had many cities in those clusters. The cities with the best mean crime rating, the worst mean recreation rating, and the worst mean economics rating often clustered together.

These cities were primarily in the northern states. The housing attribute did not seem to directly relate to or directly affect any of the other attributes when clusters were formed.

REFERENCES

1. Places Rated Almanac, by Richard Boyer and David Savageau, copyrighted and published by Rand McNally. SBN number is 0-528-88008-X.
2. <http://lib.stat.cmu.edu/datasets/places.data>
3. Data Mining, by Ian H. Witten and Eibe Frank, published by Morgan Kaufmann Publishers, copyrighted in 2000 by Academic Press.

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: placesdata
Instances: 329
Attributes: 10
ClimateandTerrain
Housing
HealthCareandEnvironment
Crime
Transportation
Education
TheArts
Recreation
Economics
place
Test mode: evaluate on training data

=== Clustering model (full training set) ===

EM
==

Number of clusters selected by cross validation: 6

Cluster: 0 Prior probability: 0.1882

Attribute: ClimateandTerrain
Normal Distribution. Mean = 532.0545 StdDev = 70.6808
Attribute: Housing
Normal Distribution. Mean = 7548.768 StdDev = 675.0553
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 1411.5821 StdDev = 514.1417
Attribute: Crime
Normal Distribution. Mean = 1154.2126 StdDev = 257.1689
Attribute: Transportation
Normal Distribution. Mean = 4871.4604 StdDev = 1104.4103
Attribute: Education
Normal Distribution. Mean = 2876.1647 StdDev = 229.3147
Attribute: TheArts
Normal Distribution. Mean = 3642.4122 StdDev = 1504.4715
Attribute: Recreation
Normal Distribution. Mean = 1792.1064 StdDev = 522.7199
Attribute: Economics
Normal Distribution. Mean = 5533.1689 StdDev = 824.3638
Attribute: place
Discrete Estimator. Counts = 1.07 2.48 4.07 2.02 7.96 8.31 6.74 8.46
5.25 7.06 8.36 10.85 1.03 1.01 1.16 2.1 1 1 (Total = 79.91)

Cluster: 1 Prior probability: 0.1809

Attribute: ClimateandTerrain
Normal Distribution. Mean = 510.7558 StdDev = 163.9554
Attribute: Housing
Normal Distribution. Mean = 9038.3189 StdDev = 1296.7464

Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 664.8186 StdDev = 303.9268
Attribute: Crime
Normal Distribution. Mean = 940.5176 StdDev = 382.5712
Attribute: Transportation
Normal Distribution. Mean = 4459.629 StdDev = 1155.7252
Attribute: Education
Normal Distribution. Mean = 2694.5195 StdDev = 300.1912
Attribute: TheArts
Normal Distribution. Mean = 1514.1912 StdDev = 963.9469
Attribute: Recreation
Normal Distribution. Mean = 2199.6111 StdDev = 945.368
Attribute: Economics
Normal Distribution. Mean = 6107.8811 StdDev = 1405.4033
Attribute: place
Discrete Estimator. Counts = 5.23 10.12 3.06 1.01 1.11 10.31 2.19 1.01
5.66 1.44 1.02 2.7 5.99 5.91 10.18 7.59 1 2 (Total = 77.52)

Cluster: 2 Prior probability: 0.0705

Attribute: ClimateandTerrain
Normal Distribution. Mean = 600.8237 StdDev = 137.6176
Attribute: Housing
Normal Distribution. Mean = 10602.7497 StdDev = 2348.6821
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 3754.148 StdDev = 1363.4347
Attribute: Crime
Normal Distribution. Mean = 1376.3008 StdDev = 482.1663
Attribute: Transportation
Normal Distribution. Mean = 6258.3294 StdDev = 1687.1317
Attribute: Education
Normal Distribution. Mean = 3253.2196 StdDev = 260.0774
Attribute: TheArts
Normal Distribution. Mean = 14136.6059 StdDev = 10725.2928
Attribute: Recreation
Normal Distribution. Mean = 2838.8535 StdDev = 958.8307
Attribute: Economics
Normal Distribution. Mean = 5962.91 StdDev = 888.2103
Attribute: place
Discrete Estimator. Counts = 1.71 2.16 5.89 3 1.38 2.99 1.88 2.03 3.12
2.76 1 3 1 1.98 2.26 3.01 1 1 (Total = 41.19)

Cluster: 3 Prior probability: 0.1346

Attribute: ClimateandTerrain
Normal Distribution. Mean = 627.5839 StdDev = 144.5412
Attribute: Housing
Normal Distribution. Mean = 11841.5326 StdDev = 3321.9603
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 1895.5676 StdDev = 714.1009
Attribute: Crime
Normal Distribution. Mean = 953.8666 StdDev = 251.9297
Attribute: Transportation
Normal Distribution. Mean = 4619.7238 StdDev = 1307.8657
Attribute: Education
Normal Distribution. Mean = 2926.074 StdDev = 407.424
Attribute: TheArts

Normal Distribution. Mean = 5548.37 StdDev = 2448.0803
Attribute: Recreation
Normal Distribution. Mean = 2284.3763 StdDev = 696.604
Attribute: Economics
Normal Distribution. Mean = 5756.3644 StdDev = 641.3096
Attribute: place
Discrete Estimator. Counts = 1.25 9.87 9.1 1.01 3.65 1.04 1.16 2.04 4.9
3.93 1.01 1.15 1 2.18 1.81 14.16 2 1 (Total = 62.27)

Cluster: 4 Prior probability: 0.1648

Attribute: ClimateandTerrain
Normal Distribution. Mean = 486.9409 StdDev = 83.3461
Attribute: Housing
Normal Distribution. Mean = 6535.9656 StdDev = 817.2448
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 471.434 StdDev = 162.2386
Attribute: Crime
Normal Distribution. Mean = 993.169 StdDev = 267.4051
Attribute: Transportation
Normal Distribution. Mean = 2767.6433 StdDev = 663.5388
Attribute: Education
Normal Distribution. Mean = 2635.7936 StdDev = 285.6398
Attribute: TheArts
Normal Distribution. Mean = 729.6243 StdDev = 613.7206
Attribute: Recreation
Normal Distribution. Mean = 1448.8958 StdDev = 636.2039
Attribute: Economics
Normal Distribution. Mean = 5787.8387 StdDev = 1121.9664
Attribute: place
Discrete Estimator. Counts = 1.06 1.26 1.39 1.01 2.12 13.35 7.95 3.75
2.51 3.14 3 20.42 1.23 2.13 1.92 3.99 1 1 (Total = 72.22)

Cluster: 5 Prior probability: 0.2611

Attribute: ClimateandTerrain
Normal Distribution. Mean = 533.0649 StdDev = 76.3398
Attribute: Housing
Normal Distribution. Mean = 7174.5118 StdDev = 854.9404
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 775.392 StdDev = 222.2988
Attribute: Crime
Normal Distribution. Mean = 707.3485 StdDev = 242.0842
Attribute: Transportation
Normal Distribution. Mean = 3706.6555 StdDev = 1079.1211
Attribute: Education
Normal Distribution. Mean = 2791.5066 StdDev = 213.347
Attribute: TheArts
Normal Distribution. Mean = 1256.8288 StdDev = 872.3675
Attribute: Recreation
Normal Distribution. Mean = 1396.2272 StdDev = 408.0389
Attribute: Economics
Normal Distribution. Mean = 4713.1403 StdDev = 628.8
Attribute: place
Discrete Estimator. Counts = 2.68 4.11 18.49 6.95 9.78 5.01 3.08 16.71
10.55 13.67 1.61 1.88 1.75 1.8 2.67 1.15 1 1 (Total = 103.89)

=== Evaluation on training set ===

EM

==

Number of clusters selected by cross validation: 6

Cluster: 0 Prior probability: 0.1882

Attribute: ClimateandTerrain

Normal Distribution. Mean = 532.0545 StdDev = 70.6808

Attribute: Housing

Normal Distribution. Mean = 7548.768 StdDev = 675.0553

Attribute: HealthCareandEnvironment

Normal Distribution. Mean = 1411.5821 StdDev = 514.1417

Attribute: Crime

Normal Distribution. Mean = 1154.2126 StdDev = 257.1689

Attribute: Transportation

Normal Distribution. Mean = 4871.4604 StdDev = 1104.4103

Attribute: Education

Normal Distribution. Mean = 2876.1647 StdDev = 229.3147

Attribute: TheArts

Normal Distribution. Mean = 3642.4122 StdDev = 1504.4715

Attribute: Recreation

Normal Distribution. Mean = 1792.1064 StdDev = 522.7199

Attribute: Economics

Normal Distribution. Mean = 5533.1689 StdDev = 824.3638

Attribute: place

Discrete Estimator. Counts = 1.07 2.48 4.07 2.02 7.96 8.31 6.74 8.46
5.25 7.06 8.36 10.85 1.03 1.01 1.16 2.1 1 1 (Total = 79.91)

Cluster: 1 Prior probability: 0.1809

Attribute: ClimateandTerrain

Normal Distribution. Mean = 510.7558 StdDev = 163.9554

Attribute: Housing

Normal Distribution. Mean = 9038.3189 StdDev = 1296.7464

Attribute: HealthCareandEnvironment

Normal Distribution. Mean = 664.8186 StdDev = 303.9268

Attribute: Crime

Normal Distribution. Mean = 940.5176 StdDev = 382.5712

Attribute: Transportation

Normal Distribution. Mean = 4459.629 StdDev = 1155.7252

Attribute: Education

Normal Distribution. Mean = 2694.5195 StdDev = 300.1912

Attribute: TheArts

Normal Distribution. Mean = 1514.1912 StdDev = 963.9469

Attribute: Recreation

Normal Distribution. Mean = 2199.6111 StdDev = 945.368

Attribute: Economics

Normal Distribution. Mean = 6107.8811 StdDev = 1405.4033

Attribute: place

Discrete Estimator. Counts = 5.23 10.12 3.06 1.01 1.11 10.31 2.19 1.01
5.66 1.44 1.02 2.7 5.99 5.91 10.18 7.59 1 2 (Total = 77.52)

Cluster: 2 Prior probability: 0.0705

Attribute: ClimateandTerrain
Normal Distribution. Mean = 600.8237 StdDev = 137.6176
Attribute: Housing
Normal Distribution. Mean = 10602.7497 StdDev = 2348.6821
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 3754.148 StdDev = 1363.4347
Attribute: Crime
Normal Distribution. Mean = 1376.3008 StdDev = 482.1663
Attribute: Transportation
Normal Distribution. Mean = 6258.3294 StdDev = 1687.1317
Attribute: Education
Normal Distribution. Mean = 3253.2196 StdDev = 260.0774
Attribute: TheArts
Normal Distribution. Mean = 14136.6059 StdDev = 10725.2928
Attribute: Recreation
Normal Distribution. Mean = 2838.8535 StdDev = 958.8307
Attribute: Economics
Normal Distribution. Mean = 5962.91 StdDev = 888.2103
Attribute: place
Discrete Estimator. Counts = 1.71 2.16 5.89 3 1.38 2.99 1.88 2.03 3.12
2.76 1 3 1 1.98 2.26 3.01 1 1 (Total = 41.19)

Cluster: 3 Prior probability: 0.1346

Attribute: ClimateandTerrain
Normal Distribution. Mean = 627.5839 StdDev = 144.5412
Attribute: Housing
Normal Distribution. Mean = 11841.5326 StdDev = 3321.9603
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 1895.5676 StdDev = 714.1009
Attribute: Crime
Normal Distribution. Mean = 953.8666 StdDev = 251.9297
Attribute: Transportation
Normal Distribution. Mean = 4619.7238 StdDev = 1307.8657
Attribute: Education
Normal Distribution. Mean = 2926.074 StdDev = 407.424
Attribute: TheArts
Normal Distribution. Mean = 5548.37 StdDev = 2448.0803
Attribute: Recreation
Normal Distribution. Mean = 2284.3763 StdDev = 696.604
Attribute: Economics
Normal Distribution. Mean = 5756.3644 StdDev = 641.3096
Attribute: place
Discrete Estimator. Counts = 1.25 9.87 9.1 1.01 3.65 1.04 1.16 2.04 4.9
3.93 1.01 1.15 1 2.18 1.81 14.16 2 1 (Total = 62.27)

Cluster: 4 Prior probability: 0.1648

Attribute: ClimateandTerrain
Normal Distribution. Mean = 486.9409 StdDev = 83.3461
Attribute: Housing
Normal Distribution. Mean = 6535.9656 StdDev = 817.2448
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 471.434 StdDev = 162.2386
Attribute: Crime
Normal Distribution. Mean = 993.169 StdDev = 267.4051

Attribute: Transportation
Normal Distribution. Mean = 2767.6433 StdDev = 663.5388
Attribute: Education
Normal Distribution. Mean = 2635.7936 StdDev = 285.6398
Attribute: TheArts
Normal Distribution. Mean = 729.6243 StdDev = 613.7206
Attribute: Recreation
Normal Distribution. Mean = 1448.8958 StdDev = 636.2039
Attribute: Economics
Normal Distribution. Mean = 5787.8387 StdDev = 1121.9664
Attribute: place
Discrete Estimator. Counts = 1.06 1.26 1.39 1.01 2.12 13.35 7.95 3.75
2.51 3.14 3 20.42 1.23 2.13 1.92 3.99 1 1 (Total = 72.22)

Cluster: 5 Prior probability: 0.2611

Attribute: ClimateandTerrain
Normal Distribution. Mean = 533.0649 StdDev = 76.3398
Attribute: Housing
Normal Distribution. Mean = 7174.5118 StdDev = 854.9404
Attribute: HealthCareandEnvironment
Normal Distribution. Mean = 775.392 StdDev = 222.2988
Attribute: Crime
Normal Distribution. Mean = 707.3485 StdDev = 242.0842
Attribute: Transportation
Normal Distribution. Mean = 3706.6555 StdDev = 1079.1211
Attribute: Education
Normal Distribution. Mean = 2791.5066 StdDev = 213.347
Attribute: TheArts
Normal Distribution. Mean = 1256.8288 StdDev = 872.3675
Attribute: Recreation
Normal Distribution. Mean = 1396.2272 StdDev = 408.0389
Attribute: Economics
Normal Distribution. Mean = 4713.1403 StdDev = 628.8
Attribute: place
Discrete Estimator. Counts = 2.68 4.11 18.49 6.95 9.78 5.01 3.08 16.71
10.55 13.67 1.61 1.88 1.75 1.8 2.67 1.15 1 1 (Total = 103.89)
Clustered Instances

0	62 (19%)
1	59 (18%)
2	23 (7%)
3	44 (13%)
4	54 (16%)
5	87 (26%)

Log likelihood: -72.59355

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: placesdata-weka.filters.AttributeFilter-R2,3,4,5,6,7,8,9
Instances: 329
Attributes: 2
ClimateandTerrain
place
Test mode: evaluate on training data

=== Clustering model (full training set) ===

EM
==

Number of clusters selected by cross validation: 6

Cluster: 0 Prior probability: 0.2583

Attribute: ClimateandTerrain
Normal Distribution. Mean = 543.6388 StdDev = 22.495
Attribute: place
Discrete Estimator. Counts = 1.6 7.47 14.67 2.31 4.26 8.76 3.18 22.04
8.29 7.98 3.95 4.27 1 3.18 2.13 5.83 1 1 (Total = 102.92)

Cluster: 1 Prior probability: 0.2821

Attribute: ClimateandTerrain
Normal Distribution. Mean = 606.9175 StdDev = 36.3556
Attribute: place
Discrete Estimator. Counts = 1.04 15.23 18.9 8.6 17.57 8.64 6.4 4.99
1.36 1.29 1.12 7.18 1 1.06 4.75 10 1.01 1 (Total = 111.13)

Cluster: 2 Prior probability: 0.3356

Attribute: ClimateandTerrain
Normal Distribution. Mean = 463.6783 StdDev = 51.8841
Attribute: place
Discrete Estimator. Counts = 7.25 4.29 5.41 1.08 1.14 17.98 10.35 3.96
11.85 19.52 7.88 22.73 2.77 7.72 1.1 1.14 1 1 (Total = 128.12)

Cluster: 3 Prior probability: 0.0594

Attribute: ClimateandTerrain
Normal Distribution. Mean = 827.838 StdDev = 59.3705
Attribute: place
Discrete Estimator. Counts = 1 1 1 1 1.03 1.08 1.01 1 1 1 1 1.05 1 1
8.49 12.84 1.04 1 (Total = 37.55)

Cluster: 4 Prior probability: 0.0536

Attribute: ClimateandTerrain
Normal Distribution. Mean = 266.5395 StdDev = 87.3278
Attribute: place

Discrete Estimator. Counts = 1.11 1.02 1.02 1 1 3.55 1.05 1.01 8.5 1.21
1.05 2.84 5.3 1.04 1 1 1 2 (Total = 35.68)

Cluster: 5 Prior probability: 0.0109

Attribute: ClimateandTerrain

Normal Distribution. Mean = 722.1376 StdDev = 5.4709

Attribute: place

Discrete Estimator. Counts = 1 1 1 1 1 1 1 1 1 1 1 1.93 1 1 2.53 1.19
1.95 1 (Total = 21.6)

=== Evaluation on training set ===

EM

==

Number of clusters selected by cross validation: 6

Cluster: 0 Prior probability: 0.2583

Attribute: ClimateandTerrain

Normal Distribution. Mean = 543.6388 StdDev = 22.495

Attribute: place

Discrete Estimator. Counts = 1.6 7.47 14.67 2.31 4.26 8.76 3.18 22.04
8.29 7.98 3.95 4.27 1 3.18 2.13 5.83 1 1 (Total = 102.92)

Cluster: 1 Prior probability: 0.2821

Attribute: ClimateandTerrain

Normal Distribution. Mean = 606.9175 StdDev = 36.3556

Attribute: place

Discrete Estimator. Counts = 1.04 15.23 18.9 8.6 17.57 8.64 6.4 4.99
1.36 1.29 1.12 7.18 1 1.06 4.75 10 1.01 1 (Total = 111.13)

Cluster: 2 Prior probability: 0.3356

Attribute: ClimateandTerrain

Normal Distribution. Mean = 463.6783 StdDev = 51.8841

Attribute: place

Discrete Estimator. Counts = 7.25 4.29 5.41 1.08 1.14 17.98 10.35 3.96
11.85 19.52 7.88 22.73 2.77 7.72 1.1 1.14 1 1 (Total = 128.12)

Cluster: 3 Prior probability: 0.0594

Attribute: ClimateandTerrain

Normal Distribution. Mean = 827.838 StdDev = 59.3705

Attribute: place

Discrete Estimator. Counts = 1 1 1 1 1.03 1.08 1.01 1 1 1 1 1.05 1 1
8.49 12.84 1.04 1 (Total = 37.55)

Cluster: 4 Prior probability: 0.0536

Attribute: ClimateandTerrain

Normal Distribution. Mean = 266.5395 StdDev = 87.3278

Attribute: place

Discrete Estimator. Counts = 1.11 1.02 1.02 1 1 3.55 1.05 1.01 8.5 1.21
1.05 2.84 5.3 1.04 1 1 1 2 (Total = 35.68)

Cluster: 5 Prior probability: 0.0109

Attribute: ClimateandTerrain

Normal Distribution. Mean = 722.1376 StdDev = 5.4709

Attribute: place

Discrete Estimator. Counts = 1 1 1 1 1 1 1 1 1 1 1 1.93 1 1 2.53 1.19
1.95 1 (Total = 21.6)

Clustered Instances

0	97	(29%)
1	89	(27%)
2	104	(32%)
3	19	(6%)
4	16	(5%)
5	4	(1%)

Log likelihood: -8.38441

=== Run information ===

Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: placesdata-weka.filters.AttributeFilter-R3,4,5,6,7,8,9
Instances: 329
Attributes: 3
ClimateandTerrain
Housing
place
Test mode: evaluate on training data

=== Clustering model (full training set) ===

EM
==

Number of clusters selected by cross validation: 6

Cluster: 0 Prior probability: 0.1559

Attribute: ClimateandTerrain
Normal Distribution. Mean = 567.1551 StdDev = 49.7187
Attribute: Housing
Normal Distribution. Mean = 10614.7553 StdDev = 1731.8766
Attribute: place
Discrete Estimator. Counts = 2.13 18.09 13.03 2.78 2.12 4.28 1.08 1.48
2.3 3.46 1.1 1.33 1.01 5.15 1.18 6.76 1 1 (Total = 69.29)

Cluster: 1 Prior probability: 0.3039

Attribute: ClimateandTerrain
Normal Distribution. Mean = 554.8504 StdDev = 50.3935
Attribute: Housing
Normal Distribution. Mean = 6661.832 StdDev = 586.4624
Attribute: place
Discrete Estimator. Counts = 1.28 1.05 10.11 5.5 16.45 14.52 11.16
20.98 5.56 6.14 4.91 11.78 1.02 1.44 2.94 1.06 1 1 (Total = 117.89)

Cluster: 2 Prior probability: 0.2859

Attribute: ClimateandTerrain
Normal Distribution. Mean = 450.5618 StdDev = 60.4011
Attribute: Housing
Normal Distribution. Mean = 7774.2508 StdDev = 1185.6562
Attribute: place
Discrete Estimator. Counts = 6.52 2.87 2.46 1.1 1.09 16.49 7.2 1.9
15.54 18.95 6.32 19.51 2.97 4.93 1.06 1.14 1 1 (Total = 112.04)

Cluster: 3 Prior probability: 0.0342

Attribute: ClimateandTerrain
Normal Distribution. Mean = 220.7674 StdDev = 68.6723
Attribute: Housing
Normal Distribution. Mean = 7989.8055 StdDev = 1720.1835
Attribute: place

Discrete Estimator. Counts = 1.02 1 1 1 1 1.36 1 1 5.98 1.02 1 2.06
4.98 1.01 1 1 1 2 (Total = 29.43)

Cluster: 4 Prior probability: 0.07

Attribute: ClimateandTerrain

Normal Distribution. Mean = 803.7862 StdDev = 80.0236

Attribute: Housing

Normal Distribution. Mean = 13278.4591 StdDev = 3973.36

Attribute: place

Discrete Estimator. Counts = 1 3.11 1.05 1.09 1.03 1.03 1.01 1 1 1 1
1.15 1 1 8.49 13.08 2 1 (Total = 41.05)

Cluster: 5 Prior probability: 0.15

Attribute: ClimateandTerrain

Normal Distribution. Mean = 594.4763 StdDev = 52.3758

Attribute: Housing

Normal Distribution. Mean = 8268.7978 StdDev = 388.1457

Attribute: place

Discrete Estimator. Counts = 1.05 3.87 14.35 3.54 4.3 3.32 1.55 7.64
1.62 1.44 1.67 4.17 1.03 1.48 5.33 8.96 1 1 (Total = 67.31)

=== Evaluation on training set ===

EM

==

Number of clusters selected by cross validation: 6

Cluster: 0 Prior probability: 0.1559

Attribute: ClimateandTerrain

Normal Distribution. Mean = 567.1551 StdDev = 49.7187

Attribute: Housing

Normal Distribution. Mean = 10614.7553 StdDev = 1731.8766

Attribute: place

Discrete Estimator. Counts = 2.13 18.09 13.03 2.78 2.12 4.28 1.08 1.48
2.3 3.46 1.1 1.33 1.01 5.15 1.18 6.76 1 1 (Total = 69.29)

Cluster: 1 Prior probability: 0.3039

Attribute: ClimateandTerrain

Normal Distribution. Mean = 554.8504 StdDev = 50.3935

Attribute: Housing

Normal Distribution. Mean = 6661.832 StdDev = 586.4624

Attribute: place

Discrete Estimator. Counts = 1.28 1.05 10.11 5.5 16.45 14.52 11.16
20.98 5.56 6.14 4.91 11.78 1.02 1.44 2.94 1.06 1 1 (Total = 117.89)

Cluster: 2 Prior probability: 0.2859

Attribute: ClimateandTerrain

Normal Distribution. Mean = 450.5618 StdDev = 60.4011

Attribute: Housing

Normal Distribution. Mean = 7774.2508 StdDev = 1185.6562

Attribute: place
Discrete Estimator. Counts = 6.52 2.87 2.46 1.1 1.09 16.49 7.2 1.9
15.54 18.95 6.32 19.51 2.97 4.93 1.06 1.14 1 1 (Total = 112.04)

Cluster: 3 Prior probability: 0.0342

Attribute: ClimateandTerrain
Normal Distribution. Mean = 220.7674 StdDev = 68.6723
Attribute: Housing
Normal Distribution. Mean = 7989.8055 StdDev = 1720.1835
Attribute: place
Discrete Estimator. Counts = 1.02 1 1 1 1 1.36 1 1 5.98 1.02 1 2.06
4.98 1.01 1 1 1 2 (Total = 29.43)

Cluster: 4 Prior probability: 0.07

Attribute: ClimateandTerrain
Normal Distribution. Mean = 803.7862 StdDev = 80.0236
Attribute: Housing
Normal Distribution. Mean = 13278.4591 StdDev = 3973.36
Attribute: place
Discrete Estimator. Counts = 1 3.11 1.05 1.09 1.03 1.03 1.01 1 1 1 1
1.15 1 1 8.49 13.08 2 1 (Total = 41.05)

Cluster: 5 Prior probability: 0.15

Attribute: ClimateandTerrain
Normal Distribution. Mean = 594.4763 StdDev = 52.3758
Attribute: Housing
Normal Distribution. Mean = 8268.7978 StdDev = 388.1457
Attribute: place
Discrete Estimator. Counts = 1.05 3.87 14.35 3.54 4.3 3.32 1.55 7.64
1.62 1.44 1.67 4.17 1.03 1.48 5.33 8.96 1 1 (Total = 67.31)
Clustered Instances

0	46 (14%)
1	101 (31%)
2	96 (29%)
3	11 (3%)
4	22 (7%)
5	53 (16%)

Log likelihood: -17.1401