SUMMARY OF BLACK TREE PRO SOFTWARE

LIST OF ALGORITHMS

Data Classification

1. Nearest Neighbor Classification

Uses the Nearest Neighbor method to predict classifications. The Nearest Neighbor of row i is row j, if the norm of the difference between rows i and j (as Euclidean vectors), is minimum for row i.

2. Delta Classification

Generates spherical clusters for each row, with a radius of delta. The value of delta is solved for by the algorithm on an unsupervised basis. Then the Nearest Neighbor algorithm is applied to each row, and if the norm of the difference between the Nearest Neighbor for row i is greater than delta, that prediction is rejected. Delta is therefore used to filter predictions, and if the distance between an input and the prediction exceeds delta, that prediction is rejected.

3. Supervised Delta Classification

Generates spherical clusters for each row, with a unique radius of delta for each row. The values of delta are solved for by the algorithm on a supervised basis. Then the Nearest Neighbor algorithm is applied to each row, and if the norm of the difference between the Nearest Neighbor for row i is greater than the applicable delta, that prediction is rejected. Delta is therefore used to filter predictions, and if the distance between an input and the prediction exceeds the applicable delta, that prediction is rejected.

4. Delta Modal Classification

Generates spherical clusters for each row, with a radius of delta. The value of delta is solved for by the algorithm on an unsupervised basis. The predicted class for row i is the modal class (i.e., the most frequent class) in the cluster for row i.

5. Delta Tree Classification

Repeatedly clusters the dataset, generating sub-clusters, and therefore a hierarchy of clusters that in turn defines a tree. Each sub-cluster is a spherical cluster with a unique radius of delta, generating multiple values of delta. The values of delta are solved for on an unsupervised basis. Then the Nearest Neighbor algorithm is applied to each row, and if the norm of the difference between the Nearest Neighbor for row i is greater than applicable value of delta, that prediction is rejected. Delta is therefore used to filter

predictions, and if the distance between an input and the prediction exceeds the applicable value of delta, that prediction is rejected.

6. Std. Dev. Classification

The standard deviation of each dimension is calculated, and collected into a row vector of dimension 1 x N, where N is the number of columns of the dataset. The norm of that vector is then calculated, and used as the value of delta to generate spherical clusters for each row. The predicted class for row i is the modal class of the cluster for row i.

7. Permutation-Based Classification

Generates randomly permuted (by row) copies of the dataset. Then the Nearest Neighbor algorithm is applied to increasingly large subsets of those permutations, beginning with one row, and increasing up to the total number of rows in the dataset. This causes unique Nearest Neighbors to be returned each iteration. The final size of the cluster is optimized by another function that repeatedly calls this underlying function. This process is unsupervised, and clusters are guaranteed to be non-empty for every row. The predicted class for row i is the modal class of the cluster for row i.

8. Discrete Classification

Assumes each row of the dataset contains discrete labels (i.e., not Euclidean vectors), and each row is therefore treated as a set. A variant of the Nearest Neighbor algorithm is applied, where the Nearest Neighbor of row i is row j, if the set intersection between rows i and j is maximum, for row i. Assumes the dataset is a CSV, plain text file, with a dimension of M x N + 1, where M is the number of rows in the dataset, N is the number of labels, and all entries are either 0 or 1, where entry (i,j) is 1 only if row i contains label j.

Data Clustering

1. Nearest Neighbor Clustering

Uses the Nearest Neighbor method to generate clusters. Row j is included in the cluster for row i, if row i is the Nearest Neighbor of row j.

2. Delta Clustering

Generates spherical clusters for each row, with a radius of delta. The value of delta is solved for by the algorithm on an unsupervised basis.

3. Supervised Delta Clustering

Generates spherical clusters for each row, with a unique radius of delta for each row. The values of delta are solved for by the algorithm on a supervised basis.

4. Std. Dev. Clustering

The standard deviation of each dimension is calculated, and collected into a row vector of dimension 1 x N, where N is the number of columns in the dataset. The norm of that vector is then calculated, and used as the value of delta to generate spherical clusters for each row.

5. Delta Tree Clustering

Repeatedly clusters the dataset, generating sub-clusters, and therefore a hierarchy of clusters that in turn defines a tree. Each sub-cluster is a spherical cluster with a unique radius of delta, generating multiple values of delta. The values of delta are solved for on an unsupervised basis.

6. Permutation-Based Clustering

Generates randomly permuted (by row) copies of the dataset. Then the Nearest Neighbor algorithm is applied to increasingly large subsets of those permutations, beginning with one row, and increasing up to the total number of rows in the dataset. This causes unique Nearest Neighbors to be returned each iteration. The final size of the cluster is optimized by another function that repeatedly calls this underlying function. This process is unsupervised, and clusters are guaranteed to be non-empty for every row.

Dataset Analysis

1. Imitate Dataset

Generates spherical clusters for each row, with a radius of delta. The value of delta is solved for by the algorithm on an unsupervised basis. Then random noise is added to the original rows of the dataset, and all resultant vectors that are within delta of some row of the original dataset are accepted as new rows in a new dataset, thereby imitating the original dataset.

2. Dimension Correlation

Implements Kendall Tau Correlation on the specified dimensions.

3. Dimension Compression

Tests accuracy before and after removing individual dimensions of the dataset using the Nearest Neighbor algorithm, thereby identifying dimensions that reduce accuracy.

Function Prediction / Analysis

1. Periodic Detection

Tests to what extent each row of the the dataset contains periodic data (e.g., [1.0 3.0 3.0 2.0], [1.001 3.0 3.001 2.0001])

2. Random Periodic Detection

Tests to what extent each row of the dataset contains periodic data with a constant average (e.g., [1 2 3], [1 0 5]).

3. Linear interpolation (Monte Carlo)

Assumes the dataset is a time-series, and generates a linear function using random weights.

4. Polynomial Interpolation (Boot Strap)

Assumes the dataset is a time-series, and differentiates the function a fixed number of times, treating the derivative closest to zero as the true zero derivative of a polynomial function. It then integrates back upwards, producing a polynomial function.

5. Nearest Neighbor Prediction

Returns the Nearest Neighbor of an input vector.

6. Delta Prediction

Clusters the dataset using Delta Clustering. Then returns all vectors within delta of the input vector.

7. Delta Modal Prediction

Clusters the dataset using Delta Clustering. Then returns all vectors within delta of the input vector, and runs clustering on that set of vectors (producing sub-clusters). The largest sub-cluster is returned as the prediction.

8. Supervised Delta Prediction

Clusters the dataset using Supervised Delta Clustering. Then returns all vectors within the applicable delta of the input vector.

9. Delta Tree Prediction

Clusters the dataset using Delta Tree Clustering. Then returns all vectors within the applicable delta of the input vector.

10. Std. Dev. Prediction

Clusters the dataset using Standard Deviation Clustering. Then returns the entire cluster associated with each input vector.

11. Permutation-Based Prediction

Clusters the dataset using Permutation-Based Clustering. Then returns the entire cluster associated with each input vector.

12. Numerical Differentiation

Assumes the dataset is a time-series, and takes the difference between adjacent rows to calculate the derivative.

13. Numerical Integration

Assumes the dataset is a time-series, and integrates over adjacent rows to calculate the definite integral.

LIST OF CHECKBOX OPTIONS

1. Normalize Dataset

Normalizes the dataset by iterating through various weights for each dimension, testing accuracy in each case, using the Nearest Neighbor method.

2. Testing File Missing Data

Indicates that the dataset contains missing data. Each such missing entry must be identified in a separate, plain text, CSV filed named “missing_data_matrix.txt”, with dimensions of M x N (the number of rows and columns (ex. classifiers) in the dataset, respectively), where all entries are either 0 or 1, and entry (i,j) is 1 only if entry (i,j) of the dataset is a missing value. In that case, the actual entry (i,j) of the dataset itself, can be any numerical value, as it will be ignored.

3. No Training Classifiers

Indicates that the training dataset does not contain classifier labels. If so, then supervised algorithms may not be used.

4. No Testing Classifiers

Indicates that the testing dataset does not contain classifier labels. If so, then accuracy cannot be calculated.

5. No Testing File

Indicates that there is no testing file, and that the selected algorithm is to be applied to the training file alone.

6. Random Testing Dataset

Causes the testing dataset to be generated using random rows of the training dataset (i.e., a random subset of the training dataset).

7. Datasets Contain Ordinal Data

Indicates that the dataset contains ordinal data (i.e., ordinal integer labels with no meaningful notion of distance between them). This occurs when you have rankings with arbitrary ordered labels (e.g., levels of enumerated education). Each such dimension must be identified in a separate, plain text, CSV filed named “ordinal_dimensions.txt”, with dimensions of 1 x K, where K is the number of ordinal dimensions. So for example, if dimensions 3 and 5 are the only ordinal dimensions, then the file will simply contain the string “3,5”.

8. Apply Confidence Metrics

Generates a measure of confidence for every prediction, and allows for predictions to be filtered using that measure. This is applicable only to predictions that generate clusters (e.g., this will not work for Nearest Neighbor).

9. Apply Arithmetic Averaging

This causes the algorithms to use an arithmetic average of the standard deviation to calculate the maximum radius size of the cluster. Unselecting this option will instead use geometric averaging. If clusters are small, or mostly empty using arithmetic averaging, try geometric averaging instead.

10. Mutually Exclusive Clusters

Applies the selected clustering algorithm, but causes the clusters to be mutually exclusive (i.e., the intersection between the clusters for rows i and j will always be empty).

LIST OF CMND OPTIONS

1. Number of Iterations (if applicable)

2. Black Tree Geometry (Easter Egg)