Advanced Pattern Recognition -- Distance Functions and Clustering

Xiaojun Qi

1

Outline

- Distance Measures
- Minimum-distance pattern classification (single prototypes and multiple prototypes)
- Cluster Seeking
  - K-Means Algorithm
  - Hierarchical Clustering

2

Introduction

- The motivation for using distance functions as a classification tool follows naturally from the fact that the most obvious way of establishing a measure of similarity between pattern vectors, which we also consider as in Euclidean space, is by determining their proximity.
- The method of pattern classification by distance functions can be expected to yield practical and satisfactory results only when the pattern classes tend to have clustering properties.

3

- The term *minimum-distance pattern classification* will be used to characterize this particular approach since the proximity of an unknown pattern to the patterns of a class will serve as a measure for its classification.
- Several cluster-seeking algorithm will be introduced in this class since clustering properties play an important role in the performance of classifiers based on a distance concept.

4

Distance Measures

(a) Point -to-Point Distance

In Euclidean *n*-dimensional space, the distance between two points **a** and **b** is given by

$$D(\mathbf{a},\mathbf{b}) = \|\mathbf{a}-\mathbf{b}\| = \sqrt{(\mathbf{a}-\mathbf{b})'(\mathbf{a}-\mathbf{b})} = \sqrt{\sum_{k=1}^{n}(a_k - b_k)^2}$$

Where **a** and **b** are *n*-dimensional vectors with the *k*th components equal to a_k and b_k, respectively.

5

(b) Point-to-Set Distance

The distance between a pattern point **x** and a set of pattern points $\{\mathbf{a}^i\}$ which represent a class of ***K*** patterns is defined as the mean-square distance between the point **x** and the ***K*** members of the set $\{\mathbf{a}^i\}$. The squared distance between **x** and $\mathbf{a}^i$ is

$$D^2(\mathbf{x},\mathbf{a}^i) = (\mathbf{x}-\mathbf{a}^i)'(\mathbf{x}-\mathbf{a}^i) = \sum_{k=1}^{n}(x_k - a_k^i)^2$$

The mean-square distance is then given by

$$\overline{D^2(\mathbf{x},\{\mathbf{a}^i\})} = \frac{1}{K}\sum_{i=1}^{K} D^2(\mathbf{x},\mathbf{a}^i) = \frac{1}{K}\sum_{i=1}^{K}\sum_{k=1}^{n}(x_k - a_k^i)^2$$

6

(c) Intraset Distance

The intraset distance for a set of pattern points $\{a^i, i=1,2, \ldots, K\}$ is given by:

$$\overline{D^2(\{\mathbf{a}^j\},\{\mathbf{a}^i\})}, \qquad i, j = 1,2,\ldots,K-1; \qquad i \neq j$$

$$D^2(\mathbf{a}^j,\mathbf{a}^i) = \sum_{k=1}^{n} (a_k{}^j - a_k{}^i)^2$$

For fixed a^j and with a^i ranging over all of the K-1 other points in the set $\{a^i\}$, the partial average is:

$$\overline{D^2(\mathbf{a}^j,\{\mathbf{a}^i\})} = \frac{1}{K-1}\sum_{i=1}^{K}\sum_{k=1}^{n}(a_k{}^j - a_k{}^i)^2$$

7

Intraset Distance is calculated as:

$$\overline{D^2(\{\mathbf{a}^j\},\{\mathbf{a}^i\})} = \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{K-1}\sum_{i=1}^{K}\sum_{k=1}^{n}(a_k{}^j - a_k{}^i)^2\right]$$

$$= \frac{1}{K(K-1)}\sum_{j=1}^{K}\sum_{i=1}^{K}\sum_{k=1}^{n}(a_k{}^j - a_k{}^i)^2$$

It can also be expressed in terms of the variances associated with the components of the pattern points.

$$\overline{D^2} = \frac{K}{K-1}\sum_{k=1}^{n}\left[\frac{1}{K^2}\sum_{j=1}^{K}\sum_{i=1}^{K}(a_k{}^j - a_k{}^i)^2\right]$$

$$= \frac{K}{K-1}\sum_{k=1}^{n}\left[\frac{1}{K^2}\sum_{j=1}^{K}\sum_{i=1}^{K}(a_k{}^j)^2 - \frac{2}{K^2}\sum_{j=1}^{K}\sum_{i=1}^{K}a_k{}^j a_k{}^i + \frac{1}{K^2}\sum_{j=1}^{K}\sum_{i=1}^{K}(a_k{}^i)^2\right]$$

$$= \frac{K}{K-1}\sum_{k=1}^{n}\left[\frac{1}{K}\sum_{i=1}^{K}\overline{(a_k{}^j)^2} - 2\overline{(a_k{}^j)}\,\overline{(a_k{}^i)} + \frac{1}{K}\sum_{j=1}^{K}\overline{(a_k{}^i)^2}\right]$$

$$= \boxed{\frac{2K}{K-1}\sum_{k=1}^{n}\left[\overline{(a_k{}^i)^2} - \overline{(a_k{}^i)}^2\right]}$$

8

The last step follows from the fact that $\overline{(a_k{}^j)^2} = \overline{(a_k{}^i)^2}$ since we are referring to the same sample set

$$\frac{1}{K}\sum_{i=1}^{K}\overline{(a_k{}^j)^2} = \overline{(a_k{}^j)^2} \quad \text{and} \quad \frac{1}{K}\sum_{j=1}^{K}\overline{(a_k{}^i)^2} = \overline{(a_k{}^i)^2}$$

Using the biased sample variance, the intraset distance is:

$$\boxed{\overline{D^2} = \frac{2K}{K-1}\sum_{k=1}^{n}(\sigma_k^*)^2}$$

$$\overline{(a_k{}^i)^2} - \overline{(a_k{}^i)}^2 = (\sigma_k^*)^2$$

The biased variance is: $(\sigma_k^*)^2 = \frac{1}{K}\sum_{i=1}^{K}(a_k{}^i - \overline{a_k{}^i})^2$

The unbiased variance is: $(\sigma_k)^2 = \frac{1}{K-1}\sum_{i=1}^{K}(a_k{}^i - \overline{a_k{}^i})^2$

Using the unbiased sample variance, the intraset distance is:

$$(\sigma_k)^2 = \frac{K}{K-1}(\sigma_k^*)^2$$

$$\boxed{\overline{D^2} = 2\sum_{k=1}^{n}(\sigma_k)^2}$$

9

(d) Interset Distance

The interset distance between sets $\{a^i\}$ and $\{b^j\}$ containing K_a and K_b samples, respectively, is given by:

$$\overline{D^2(\{\mathbf{a}^i\},\{b^j\})}, \qquad i = 1,2,\ldots,Ka\,;\; j = 1,2,\ldots,Kb$$

- However, this expression is not easily reduced to a simple closed form in terms of statistical properties.
- An alternative way to measure interset distances is to use the distance between the centriods of the two sets under consideration or to use the Mahalanobis distance.

10

Minimum-Distance Pattern Classification

- Pattern classification by distance functions is one of the earliest concepts in automatic pattern recognition.
- This simple classification technique is an effective tool for the solution of problems in which the pattern classes exhibit a reasonable limited degree of variability.

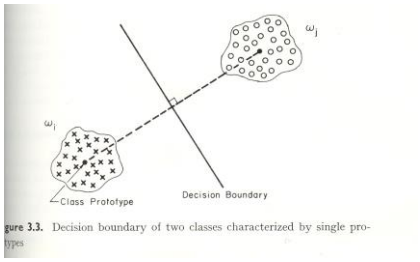
11

Minimum-Distance Pattern Classification -- Single Prototype

- In some situations, the patterns of each class tend to cluster tightly about a typical or representative pattern for that class. This occurs in cases where pattern variability and other corruptive influences are well behaved.
  - Example: The characters in the checks are highly stylized and are usually printed in magnetic ink to facilitate the measurement process. Therefore, the resulting measurement vectors (patterns) of each character class will be almost identical since the same characters in different checks are identical for all practical purposes.

12

- The linear decision surface separating every pair of prototype points z_i and z_j is the hyperplane which is the perpendicular bisector of the line segment joining the two points.



gure 3.3. Decision boundary of two classes characterized by single pro-types

13

- The minimum-distance classifiers are a special case of linear classifiers, in which the decision boundaries are constrained to have this property.
- Since a minimum-distance classifier categorizes a pattern on the basis of the closest match between the pattern and the respective class prototypes, this approach is also known as *correlation* and *cluster matching*.

14

Mean Vector of the Pattern Class

$$m_j = \frac{1}{N_j} \sum_{x \in w_j} x_j \qquad j = 1,2,..., W$$

Distance Measure

$$D_j(x) = \| x - m_j \| \qquad j = 1,2,..., W$$

Decision Functions

$$d_j(x) = x^T m_j - \frac{1}{2} m_j^T m_j \qquad j = 1,2,..., W$$

Decision Boundary

$$d_{ij} = d_i(x) - d_j(x)$$
$$= x^T (m_i - m_j) - \frac{1}{2}(m_i - m_j)^T (m_i - m_j) = 0$$

15

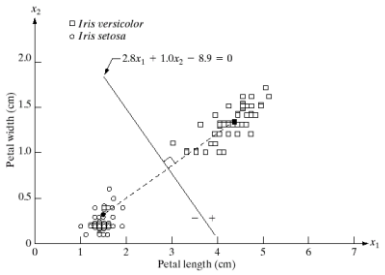


FIGURE 12.6 Decision boundary of minimum distance classifier for the classes of *Iris versicolor* and *Iris setosa*. The dark dot and square are the means.

M1 = (4.3, 1.3)

M2 = (1.5, 0.3)

$$d_1(x) = x^T m_1 - \frac{1}{2} m_1^T m_1 = 4.3x_1 + 1.3x_2 - 10.1$$

$$d_2(x) = x^T m_2 - \frac{1}{2} m_2^T m_2 = 1.5x_1 + 0.3x_2 - 1.17$$

$$d_{12}(x) = d_1(x) - d_2(x) = 2.8x_1 + 1.0x_2 - 8.9 = 0$$

16

Minimum-Distance Pattern Classification -- Multi-Prototypes

- Each pattern of class w_i tends to cluster around one of the prototypes $z_i^1, z_i^2, ..., z_i^{N_i}$ where *Ni* is the number of prototypes in the *i*th pattern class.

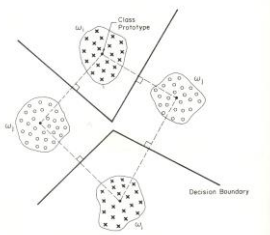


Figure 3.4. Piecewise-linear decision boundaries for two classes, each of which is characterized by two prototypes.

17

Minimum-Distance Pattern Classification -- Multi-Prototypes

- For the case that the decision boundaries for a two-class case in which each class contains two prototypes, the boundaries between the two classes are piecewise linear.
- Since we could have defined this as a single-prototype, four-class problem, the sections of the boundaries are the perpendicular bisectors of the lines joining the prototypes of different classes. This is in agreement with the decision boundaries of single-prototype classifiers.

18

Cluster Seeking

- Cluster seeking is very much an experiment-oriented "art" in the sense that the performance of a given algorithm depends on:
  - The type of data being analyzed;
  - The chosen measure of pattern similarity;
  - The method used for identifying clusters in the data.

19

Unsupervised Pattern Recognition

- Cluster seeking belongs to unsupervised pattern recognition.
- The unsupervised learning problem may be stated as that of identifying the classes in the given set of patterns.
- If cluster centers are used as a method of representation, one way of characterizing a given set of data is by cluster identification.
- Once the cluster centers are determined
  1. The classes can be used to determine decision functions by means of one or more of the training algorithms.
  2. The identified cluster centers may be used as the basis for a minimum-distance classifier.

20

Cluster Seeking
-- Measures of Similarity

- To define a data cluster, it is necessary to first define a measure of similarity which will establish a rule for assigning patterns to the domain of a particular cluster center.
- Distance (Similarity) Measures
  - Euclidean Distance: D = ||x-z||.
  - Mahalanobis Distance:

$$D = (x - m)' C^{-1} (x - m)$$

where C is the covariance matrix of a pattern population, m is the mean vector, and x represents a variable pattern.

21

Cluster Seeking
-- Measures of Similarity

- Nonmetric similarity function

$$s(x, z) = \frac{x' z}{\| x \| \| z \|}$$

Which is the cosine of the angle between the vectors x and z. It is maximum when x and z are oriented in the same direction with respect to the origin.

- This measure of similarity is useful when cluster regions tend to develop along principal axes.
- The use of this similarity measure is governed by certain qualifications, such as sufficient separation of cluster regions with respect to each other as well as with respect to the coordinate system origin.

22

Cluster Seeking
-- K-Means Algorithm

- The K-means algorithm is based on the minimization of a performance index which is defined as the sum of the squared distances from all points in a cluster domain to the cluster center.

23

Cluster Seeking
-- K-Means Algorithm (Cont.)

- Step 1: Choose K initial cluster centers **z**1(1), **z**2(1), …, **z**K(1). These are arbitrary and are usually selected as the first K samples of the given sample set.
- Step 2: At the *k*th iterative step, distribute the samples {**x**} among the K cluster domains, using the relation,

$$\mathbf{x} \in S_j(k) \quad \text{if} \quad \|\mathbf{x} - \mathbf{z}_j(k)\| < \|\mathbf{x} - \mathbf{z}_i(k)\| \qquad (1)$$

For all $i = 1, 2, \ldots, K$, $i \neq j$, where $S_j(k)$ denotes the set of samples whose cluster center is $\mathbf{z}_j(k)$. Ties in expression (1) are resolved arbitrarily.

24

- Step 3: From the results of Step 2, compute the new cluster centers $\mathbf{z}j(k+1)$, $j = 1, 2, \ldots, K$, such that the sum of the squared distances from all points in $Sj(k)$ to the new cluster center is minimized. In other words, the new cluster center $\mathbf{z}j(k+1)$ is computed so that the performance index

$$J_j = \sum_{\mathbf{x} \in S_j(k)} \|\mathbf{x} - \mathbf{z}_j(k+1)\|^2, \quad j = 1, 2, \ldots, K$$

is minimized. The $\mathbf{z}j(k+1)$, which minimizes this performance index is simply the sample means of $Sj(k)$. Therefore, the new cluster center is given by

$$\mathbf{z}_j(k+1) = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j(k)} \mathbf{x}, \quad j = 1, 2, \ldots, K$$

Where Nj is the number of samples in $Sj(k)$. The name "K-means" is obviously derived from the manner in which cluster centers are sequentially updated.

25

- Step 4: If $\mathbf{z}j(k+1) = \mathbf{z}j(k)$ for $j = 1, 2, \ldots, K$, the algorithm has converged and the procedure is terminated. Otherwise go to Step 2.
- ➔ The behavior of the K-means algorithm is influenced by:
  - The number of cluster centers specified;
  - The choice of initial cluster centers;
  - The order in which the samples are taken;
  - The geometrical properties of the data.
- ➔In most practical cases the application of this algorithm will require experimenting with various values of K as well as different choices of starting configurations.

26

Cluster Seeking
-- K-Means Algorithm Example

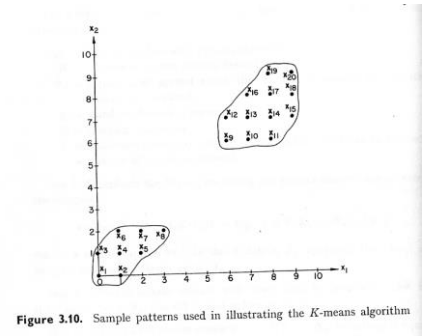


Figure 3.10. Sample patterns used in illustrating the *K*-means algorithm

27

- Step 1. Let K = 2 and choose $\mathbf{z}1(1) = \mathbf{x}1 = (0, 0)'$, $\mathbf{z}2(1) = \mathbf{x}2 = (1, 0)'$.
- Step 2. Since $\|\mathbf{x}_1 - \mathbf{z}_1(1)\| < \|\mathbf{x}_1 - \mathbf{z}_i(1)\|$ and

$$\|\mathbf{x}_3 - \mathbf{z}_1(1)\| < \|\mathbf{x}_3 - \mathbf{z}_i(1)\|, \quad i = 2$$

we have that $S_1(1) = \{\mathbf{x}_1, \mathbf{x}_3\}$. Similarly, the remaining patterns are closer to $\mathbf{z}2(1)$, so

$$S_2(1) = \{\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5, \ldots, \mathbf{x}_{20}\}$$

28

- Step 3. Update the cluster centers:

$$\mathbf{z}_1(2) = \frac{1}{N_1} \sum_{\mathbf{x} \in S_1(1)} \mathbf{x}$$
$$= \frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_3)$$
$$= \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix}$$
$$\mathbf{z}_2(2) = \frac{1}{N_2} \sum_{\mathbf{x} \in S_2(1)} \mathbf{x}$$
$$= \frac{1}{18}(\mathbf{x}_2 + \mathbf{x}_4 + \cdots + \mathbf{x}_{20})$$
$$= \begin{pmatrix} 5.67 \\ 5.33 \end{pmatrix}$$

29

- Step 4. Since $\mathbf{z}j(2) \neq \mathbf{z}j(1)$, for $j = 1, 2$, we return to Step 2.
- Step 2. With the new cluster centers we obtain

$$\|\mathbf{x}_l - \mathbf{z}_1(2)\| < \|\mathbf{x}_l - \mathbf{z}_2(2)\|$$
for $l = 1, 2, \ldots, 8$
and
$$\|\mathbf{x}_l - \mathbf{z}_2(2)\| < \|\mathbf{x}_l - \mathbf{z}_1(2)\|$$

- Therefore: for $l = 9, 10, \ldots, 20$

$$S_1(2) = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_8\}$$
and
$$S_2(2) = \{\mathbf{x}_9, \mathbf{x}_{10}, \ldots, \mathbf{x}_{20}\}$$

30

- Step 3. Update the cluster centers:

$$\mathbf{z}_1(3) = \frac{1}{N_1}\sum_{\mathbf{x}\in S_1(2)}\mathbf{x}$$

$$= \frac{1}{8}(\mathbf{x}_1 + \mathbf{x}_2 + \cdots \mathbf{x}_8)$$

$$= \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}$$

$$\mathbf{z}_2(3) = \frac{1}{N_2}\sum_{\mathbf{x}\in S_2(2)}\mathbf{x}$$

$$= \frac{1}{12}(\mathbf{x}_9 + \mathbf{x}_{10} + \cdots + \mathbf{x}_{20})$$

$$= \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}$$

31

- Step 4. Since $\mathbf{z}j(3) \neq \mathbf{z}j(2)$, for j = 1, 2, we return to Step 2.
- Step 2 yields the same results as in the previous iteration:

$$S_1(4) = S_1(3) \text{ and } S_2(4) = S_2(3)$$

- Step 3 also yields the same results.
- Step 4. Since $\mathbf{z}j(4) = \mathbf{z}j(3)$ for j = 1, 2, the algorithm has converged, yielding these cluster centers:

$$\mathbf{z}_1 = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}, \quad \mathbf{z}_2 = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}$$

These results agree with what we would expect from inspection of the given data.

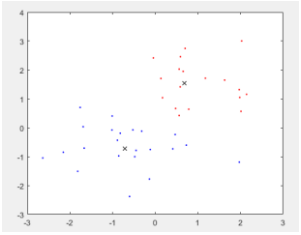
32

kmeans Matlab Function

```
X = [randn(20,2)+ones(20,2); ...
     randn(20,2)-ones(20,2)];
opts = statset('Display','final');
[cidx, ctrs] = kmeans(X, 2, 'Distance','city', ...
          'Replicates',5, 'Options',opts);
plot(X(cidx==1,1),X(cidx==1,2),'r.', ...
   X(cidx==2,1),X(cidx==2,2),'b.', ctrs(:,1),ctrs(:,2),'kx');
```

33

- Replicate 1, 2 iterations, total sum of distances = 51.1223.
- Replicate 2, 4 iterations, total sum of distances = 51.1223.
- Replicate 3, 2 iterations, total sum of distances = 51.1223.
- Replicate 4, 2 iterations, total sum of distances = 51.1223.
- Replicate 5, 2 iterations, total sum of distances = 51.1223.
- Best total sum of distances = 51.1223



34

Cluster Seeking
-- Hierarchical Clustering

- Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of Johnson's (1967) hierarchical clustering is this:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters. In single link clustering, the rule is that the distance from the compound object to another object is equal to the shortest distance from any member of the cluster to the outside object.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

35

clusterdata Matlab Function

```
% Compute four clusters of the Fisher iris data using Ward %
linkage and ignoring species information, and see how % the
cluster assignments correspond to the three species.
    load fisheriris
    c = clusterdata(meas,'linkage','ward','maxclust',4);
    crosstab(c,species)
      0   25    1
      0   24   14
      0    1   35
     50    0    0
when using c = clusterdata(meas,'linkage','ward','maxclust',3);
    crosstab(c,species)
0  1  35
0  49 15
50  0  0
```

36

Cluster Seeking
-- Evaluation of Clustering Results

- The principal difficulty in evaluating the results of clustering algorithms is inability to visualize the geometrical properties of a high-dimensional space.
  - A very useful interpretation tool is the distance between cluster centers. This information is best presented in a table.
  - Cluster membership information can also be used for merging purposes. If two cluster centers are relatively close, and one of the centers is associated with a much larger number of samples, it is often possible to merge the two cluster domains.

37

  - The variances of a cluster domain about its mean can be used to infer the relative distribution of the samples in the domain.
  - It is also useful to know the closest and farthest points from the cluster center in each domain.
  - The average distance between cluster centers can be used to augment the information present in a distance table.
  - The covariance matrix of each sample set can also be of value, although it is difficult to interpret in high-dimensionality problems and can add computational difficulties to an iterative algorithm.

38