2006 Autumn semester Pattern Information Processing
Topic 4. Pattern recognition

Session 10. (1) Discriminant analysis

Pattern recognition is one of the most important topics related to image processing. The first session of Topic 4 explains discriminant analysis, which is a statistical framework of pattern recognition. The concept of pattern recognition is formalized as classification of vectors representing patterns. Since the vectors corresponding to each category are affected by random effects, the classification should be formalized statistically. This session presents the concept of discriminant analysis based on Mahalanobis distance.

What is pattern recognition?

Pattern recognition is classification of target patterns into several categories. The number of different patterns is usually much larger than the number of categories. Consider the case of hand-written character recognition. Although hand-written patterns of the character “A” have a lot of variants, all of them have to be categorized into the character “A.”

If we consider the simplest case, a recognition system where possible input characters are “A” and “B,” the system has to classify each pattern into either the category “A” or the category “B.” This simple classification is often difficult since the patterns, hand-written characters in this case, are usually deformed, and such a pattern often appears that is difficult to determine which category to be classified into.

In the pattern recognition procedure, a feature vector is extracted from each pattern. The feature vector is a summarization of a pattern, and contains, in case of image pattern for example, average brightness, edge orientations, spatial frequency spectra or size density, etc. The objective of generating a feature vector is the reduction of dimensionality. A pattern itself is often a very high dimensional vector; For example an image of 256×256 pixels is a 65536-dimensional vector. If we classify the patterns in 65536-dimensional space, the recognition system must become too sensitive and often misclassifies degraded patterns. Thus the feature vector is introduced to reduce the dimensionality.

The methods of extracting various features are discussed in the area of image analysis for visual patterns, speech analysis for audio patterns, etc. Several methods of image analysis have been discussed in this course. The aim of this topic is explaining methods of categorizing feature vectors. The statistical discriminant analysis based on Mahalanobis distance is one of the fundamental idea of the optimal classification.

Discriminant analysis – in one dimensional case

The discriminant analysis is a method classifying an unclassified vector using training vectors that have been already correctly classified. Let us consider at first the case that the vectors are one-dimensional, i. e. equivalent to scalers. We assume that the training vectors have been classified into category A or category B. Let us consider how to determine which category a new vector to be classified into.

It is natural to get an idea at first that the new vector should be classified to the “nearer” category, by measuring the distances between the new vector and the average of vectors in each category. However, this method is not sufficient, as shown in Fig. 1. Although the new vector is classified into the category A by the above method, it should be classified into the category B, since the new vector is in the area where the vectors in category B are distributed.

The reason of this misclassification is that the variance of category B is much larger than that of category A. The distance should be defined considering the difference of variances. The difference between the new vector and the mean of the category of large (small) variance should be evaluated smaller (larger) than the simple distance.

It is assumed that the preclassified vectors of each

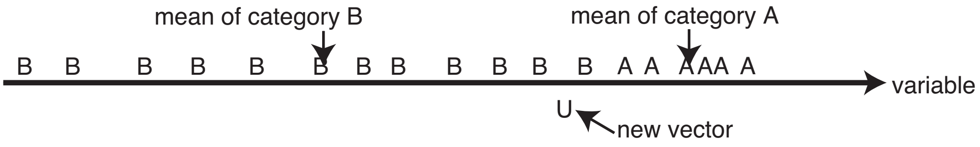

Fig. 1: Which category to be classified into?

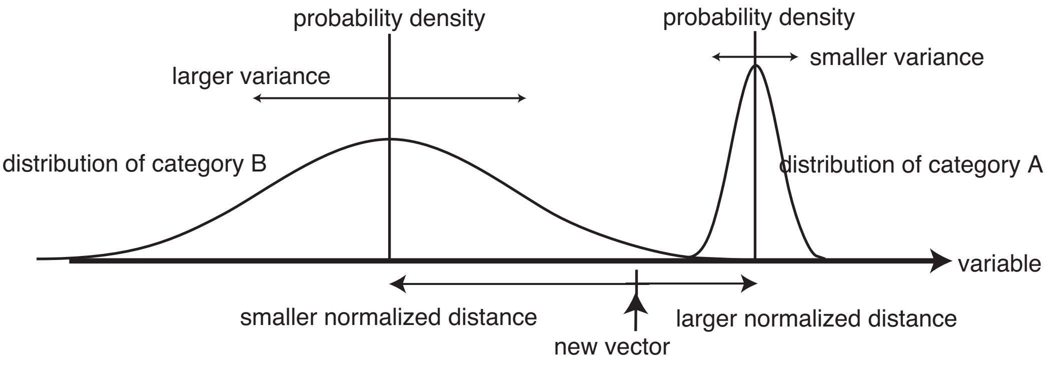
probability density
probability density
larger variance
smaller variance
distribution of category B
distribution of category A
variable
smaller normalized distance
larger normalized distance
new vector

Fig. 2: Normalized distances.

category are sampled from the population corresponding to each category. For example, the vectors in a category of character "A" are assumed to be sampled from the population of all of considered hand-written patterns of character "A." The probability density of the appearance of a preclassified vector is equivalent to its density in the population. Let the mean and the variance of the population corresponding to category A be μ_A and σ_A^2, respectively, and those of category B be μ_B and σ_B^2. The "normalized" squared distances D_A^2 and D_B^2 from the new vector x to μ_A and μ_B, respectively, are defined as follows:

$$D_A^2 = \frac{(x-\mu_A)^2}{\sigma_A^2},\ D_B^2 = \frac{(x-\mu_B)^2}{\sigma_B^2}. \qquad (1)$$

We get from the above equation that the larger (smaller) the variance of the category is, the smaller (larger) this distance is. The new vector can be classified into the "nearer" category based on this distance. Figure 2 illustrates the normalized distance.

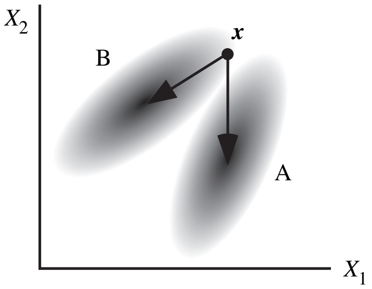

Fig. 3: Which is nearer from x, A or B?

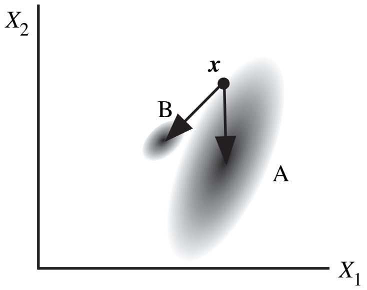

Fig. 4: x is "nearer" to A.

Mahalanobis distance

Let us consider the two dimensional case, i. e. the case where each vector has two elements. A set of two dimensional vectors is illustrated on a scattergram. We also consider that there are a number of preclassified vectors on the scattergram, and these are samples from each population. A two-dimensional probability density is defined for each population.

Figure 3 shows example density functions of categories A and B. Let us consider to determine which category the new vector $x = (x_1, x_2)$ should be classified into. It is still insufficient that it is classified into "nearer" category based on the Euclidian distance. Figure 4 shows a counterexample. The Euclidian distance from x to the centroid of category A is equal to that from x to B. However, the probability that x belongs to category B is almost zero, since the probability density of B concentrates to the small neighborhood of the centroid. On the contrary, the probability that x belongs to category A is a little larger than zero, since category A has a broader density function. Thus x should be classified into category A, and the distance should be normalized similarly to the one-dimensional case.

The two-dimensional case has one more issue to be considered. The distribution in Fig. 5 has a large variance in b–d direction, and a small variance in a–c direction. Thus the distance from the centroid to the vectors on a, b, c, and d should be defined identically. It indicates that the definition of the normalized distance in the twodimensional case should consider the shape of distribution, or the covariance matrix, of the distribution.

Let us consider the distance between a vector and the centroid of a distribution, satisfying the above requirement. We assume a two-dimensional distribution as shown in Fig. 6. Let the means of the variables x_1, x_2 be μ_1, μ_2, respectively, and the variances be σ_1^2, σ_2^2, respectively, and the correlation coefficient be ρ. Normalizing x_1, x_2 as

$$u_1 = \frac{x_1 - \mu_1}{\sigma_1},\ u_2 = \frac{x_2 - \mu_2}{\sigma_2}, \tag{2}$$

we get that both of the means of u_1, u_2 are 0, the variances are 1, and the correlation coefficient is still ρ.

We further convert u_1, u_2 into $z_{(1)}, z_{(2)}$, which are the variables not correlated to each other. These variables are known as the principal components, which were explained in Topic 2. In case that the number of variables is two, the principal components on normalized bases u_1, u_2 are independent on the correlation coefficient, as follows:

$$z_{(1)} = \frac{u_1 + u_2}{\sqrt{2}},\ z_{(1)} = \frac{u_1 - u_2}{\sqrt{2}}, \tag{3}$$

as shown in Fig. 7 [1].

By these conversions, the correlation between $z_{(1)}$ and $z_{(2)}$ need not to be considered. The desired distance between a vector and the centroid is defined by the Euclidian distance simply normalized by the variance on each of $z_{(1)}, z_{(2)}$ – bases, denoted $V(z_{(1)})$ and $V(z_{(2)})$, respectively, as follows:

$$D^2 = \frac{Z_{(1)}^2}{V(z_{(1)})} + \frac{Z_{(2)}^2}{V(z_{(2)})}. \tag{4}$$

This squared distance is called *Mahalanobis distance,* which is an admissible distance between a vector and the centroid of a distribution.

The variances $V(z_{(1)})$ and $V(z_{(2)})$ are equal to the eigenvalues corresponding to $z_{(1)}$ and $z_{(2)}$, respectively, and denoted as follows by a simple calculation:

$$V(z_{(1)}) = 1 + \rho,\ V(z_{(2)}) = 1 - \rho. \tag{5}$$

We get from Eqs. (2) and (5) that

$$\begin{aligned} D^2 &= \frac{Z_{(1)}^2}{V(z_{(1)})} + \frac{Z_{(2)}^2}{V(z_{(2)})} = \frac{\left(\frac{u_1+u_2}{\sqrt{2}}\right)^2}{1+\rho} + \frac{\left(\frac{u_1-u_2}{\sqrt{2}}\right)^2}{1-\rho} \\ &= \frac{(1-\rho)(u_1+u_2)^2 + (1+\rho)(u_1-u_2)^2}{2(1+\rho)(1-\rho)} \\ &= \frac{2(u_1^2+u_2^2) - 2\rho \cdot 2u_1u_2}{2(1-\rho^2)} \\ &= \frac{u_1^2+u_2^2-2\rho u_1u_2}{1-\rho^2}. \end{aligned} \tag{6}$$

[1] Note that the principal components depend on the correlation coefficient for the case of three or more variables, not as in Eq. (3)

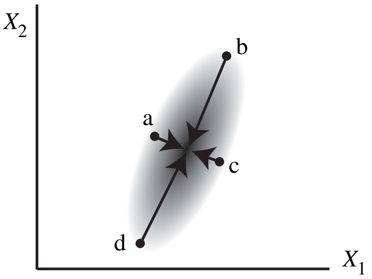

Fig. 5: “Same” distances from $a - d$.

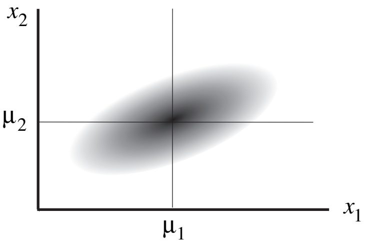

Fig. 6: Two-dimensional distribution.

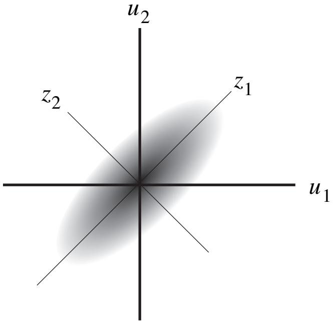

Fig. 7: Conversion to principal components.

Curse of dimensionality

The Mahalanobis distance in the form of Eq. (6) depends on the correlation coefficient ρ and the variances σ_1^2 and σ_2^2, in other words, the covariance matrix. These parameters are defined on the distribution of population. However, the distribution is unknown and only the samples, i. e. preclassified vectors, are given. Thus the distribution should be estimated from the preclassified vectors.

In the case of pattern recognition problems, the number of features to be incorporated into the feature vectors is often large, since a lot of features need to be considered to characterize the patterns. This means that the dimensionality of the feature vectors is high. On the other hand, it is difficult to prepare a large number of preclassified vectors, or training examples. A small number of preclassified vectors in high-dimensional space causes low-quality estimation of the distribution, since the arrangement of the preclassified vectors is too sparse in the space. This is called *curse of dimensionality.* A lot of methods to avoid the curse of dimensionality have been developed. The support vector machine with the kernel method, which will be explained in the third session of this topic, is one of these efforts.

References

R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification,* second edition, John Wiley & Sons, ISBN0-471-05669-3 (2001).（邦訳　パターン識別（尾上守夫監訳），新技術コミュニケーションズ，ISBN4-915851-24-9 (2001).）