2 Pattern recognition

Humans are particularly good at recognizing many patterns such as faces and voices of other individuals. A possibly harmful behaviour of another person or the appearance of a possibly dangerous animal may also be quickly identified. Obviously such pattern recognition abilities have implied a survival advantage during the evolution of humans.

By training humans can also be astonishingly good at tasks such as recognizing the species of a bird at a long distance, perhaps by using a combination of features such as the bird's shape and colours, its vocalization and its mode of flight. The human observer's previous knowledge of how common possible bird species are in the current environment at the given time of the year may also be highly useful in identifying the species.

One important task in pattern recognition based on digital images is to try to mimic human pattern recognition by choice of suitable features for recognizing and classifying observed objects. We can divide the field of pattern classification into two disciplines depending on the our previous knowledge of the possible classes. The most well developed discipline is *discriminant analysis* where we assume that we have a given number of classes and that we have a new object that we want to assign to one of these classes. Typically we also assume here that we have a set of objects for which we know the classes. Such a data set, often called a training set, will help us to choose the relevant features of the objects and to design the algorithm for recognizing the class by use of the chosen features. Therefore discriminant analysis is often called *supervised pattern recognition* or *learning with a teacher*.

In the second discipline, called *cluster analysis* we do not assume any prior knowledge of possible classes. However, we will typically assume that we also here have a given data set but without any classification. The data set will be used to find clusters, and the discipline is often referred to as *unsupervised pattern recognition* or *learning without a teacher*.

We will start by discussing discriminant analysis. Several of the sets of images in the previous chapter, the weed seeds in Example 1.2, the weed plants in Example 1.3 and the handwritten digits in Example 1.7 describe problems that call for discriminant analysis.

2.1 Optimal discrimination with two classes and a one feature variable

Suppose that we have two classes ω_1 and ω_2 and a real-valued feature variable X for each object to be classified. Assume that we know how common the two classes are, that is, we know the prior probabilities of the two classes. Assume also that we know the distributions of the feature variable corresponding to the two classes.

For $i = 1, 2$, let π_i denote the prior probability of class ω_i and let f_i be the probability density of X for an observation from class ω_i, or the probability function, $f_i(x) = P(X = x)$, if X is a discrete random variable.

The problem of deciding if an object comes from class ω_1 or ω_2 is to be based on observation of the corresponding feature variable X. Thus we need to specify two disjoint sets A_1 and A_2 with $A_1 \cup A_2 = \mathbb{R}$ and choose class ω_i if $X \in A_i$. To find optimal sets we need further specification corresponding to how costly it is to make different kinds of errors, that is the cost of choosing class ω_1 when ω_2 is true and vice versa. Let us first assume that these cost are equal, and more specifically, that we want to minimize the probability of misclassification.

It turns out that the probability of misclassification is minimized if we use the following rule:

$$\text{choose class } \omega_1 \text{ if } \pi_1 f_1(x) > \pi_2 f_2(x), \tag{30}$$

$$\text{choose class } \omega_2 \text{ if } \pi_1 f_1(x) < \pi_2 f_2(x). \tag{31}$$

To show that a decision rule satisfying (30) and (31) is optimal we note that the probability of misclassification is generally given by

$$\begin{aligned}\Pr(\text{misclassification}) &= \Pr(\omega_1 \text{ true and misclassification}) + \Pr(\omega_2 \text{ true and misclassification})\\ &= \Pr(\omega_1)\Pr(\text{misclassification}|\omega_1) + \Pr(\omega_2)\Pr(\text{misclassification}|\omega_2)\\ &= \pi_1 \int_{A_2} f_1(x)dx + \pi_2 \int_{A_1} f_2(x)dx.\end{aligned}$$

In Figure 26 the set A_1 extends up to a threshold t while A_2 is chosen above t. The probability of misclassification is equal to the area of the coloured region, and it follows that it is minimized precisely when the threshold is the horisontal location of the crossing point of the two curves. Thus the misclassification probability is minimized if A_1 and A_2 are chosen as in (30) and (31). (We note that x-values such that $\pi_1 f_1(x) = \pi_2 f_2(x)$ may be brought to either A_1 or A_2 without affecting the misclassification probability.)

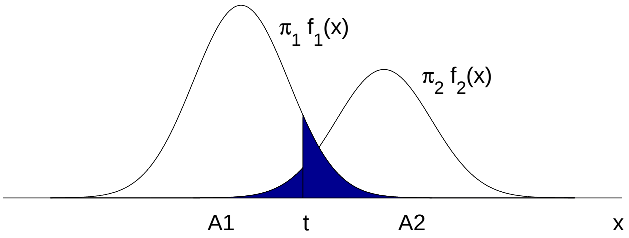

Figure 26: Probability of misclassification is given by the coloured area. The set A_1 where class ω_1 is chosen extends here up to the threshold t, while A_2 is chosen above t.

Example 2.12. *Two-class discriminant analysis with estimated normal densities.*

Suppose that we have a training set with n_1 objects from class ω_1 and n_2 objects from class ω_2. We assume that we have obtained random samples from both classes and that the two samples are independent. We assume further that the variable X is normally distributed with expectation μ_i and variance σ_i^2 in class ω_i, $i = 1, 2$, where we assume that expectations are different in the two classes while the variances may either be assumed to be equal or unequal. Let the observations be denoted X_{im}, $m = 1, \dots, n_i$, $i = 1, 2$. Then it is natural to estimate the expectation in class ω_i by

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{m=1}^{n_i} X_{im}, \quad i = 1, 2. \tag{32}$$

If we make no assumption on equality of the variances we use the variance estimates

$$s_i^2 = \frac{1}{n_i - 1} \sum_{m=1}^{n_i} (X_{im} - \hat{\mu}_i)^2, \quad i = 1, 2, \tag{33}$$

but if we assume variance equality we use the estimate

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{34}$$

for the common variance. □ □

We note that compared to Example 2.12 we have in Example 1.10, where we have classified pixels into soil or plant pixels, a similar but more complicated situation as we here do not have training sets for soil and plant pixels but use the model specified by (10) and (12) for all pixels. Also the proportions of soil and plant pixels are estimated.

2.2 Optimal discrimination with k classes and a d-dimensional feature vector

Suppose now that we have k classes $\omega_i, i = 1, \dots, k$, and a d-dimensional feature vector X for each object to be classified. Let π_i be the prior probability of class ω_i and let f_i be the probability density of X for an observation from class ω_i, $i = 1, \dots, k$. Let us further assume that the cost of assigning an object to class ω_i is $c(i|j)$ when the true class is ω_j. Rather than minimizing the misclassification probability we now want to *minimize the expected cost.*

A decision function for our problem is now specified by a partition of d-dimensional space $\mathbb{R}^d$ into k disjoint sets $A_1, \dots, A_k$ with $\cup_{i=1}^k A_i = \mathbb{R}^d$. If $X \in A_i$ we assign our object to class $\omega_i, i = 1, \dots, k$.

Now it turns out that the expected cost is minimized if the sets A_i satisfy the following condition

$$x \in A_i \;\Rightarrow\; \text{subscript } i \text{ minimizes } \sum_{j=1}^{k} c(i|j)\pi_j f_j(x). \tag{35}$$

If the sum is minimized by several i-values for a certain x-value, then this x-value may be allocated to A_i for any of these i-values.

To show that a decision rule which satisfies (35) is optimal let us consider an arbitrary decision function specified by a a partition $A_1, \dots, A_k$ of $\mathbb{R}^d$. The expected cost for this decision rule may be written

$$\sum_{i=1}^{k} \int_{A_i} \sum_{j=1}^{k} c(i|j)\pi_j f_j(x)dx,$$

from which it follows that a decision rule satisfying the condition (35) is optimal.

Let us now assume that all misclassifications have the same cost, and that the cost of a correct decision is zero. Our criterion then implies that we shall *minimize the probability of misclassification*, and it is not difficult to see that we shall prefer class ω_i to class ω_j if

$$\pi_i f_i(x) > \pi_j f_j(x) \tag{36}$$

similar to what we found previously for the case with two classes and one feature variable.

2.3 Normally distributed feature vectors, linear and quadratic discrimination

A d-dimensional random (column) vector X is said to be $\mathrm{N}(\mu,C)$, that is have a d-dimensional normal distribution with expectation vector μ and covariance matrix C, if X has the d-dimensional density function

$$f_X(x) = \frac{1}{(2\pi)^{d/2}(\det C)^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)), \tag{37}$$

where $\det C$ denotes the determinant of the matrix C.

An important special case in discrimination is to assume that the d-dimensional feature vector X has a multivariate normal distribution $\mathrm{N}(\mu_i,C_i)$ in class ω_i, $i = 1,\ldots,k$. Sometimes the covariance matrices are assumed to be equal, that is

$$C_i = C, \quad i = 1,\ldots,k. \tag{38}$$

Let us first assume that the covariance matrices are all equal to C and that we want to minimize the probability of misclassification. A computation from (36) and (121) shows that if $X = x$ is observed we shall prefer class ω_i to ω_j if

$$(\mu_i - \mu_j)^T C^{-1}(x - \frac{1}{2}(\mu_i + \mu_j)) > \ln\frac{\pi_j}{\pi_i}. \tag{39}$$

We note that (39) is linear in x and this case is therefore often called *linear discrimination.*

Let us now find a corresponding rule without the assumption (38). It follows from (36) and (121) that we shall prefer class ω_i to ω_j if

$$\begin{aligned}\frac{1}{2}x^T(C_j^{-1} - C_i^{-1})x + (\mu_i^T C_i^{-1} - \mu_j^T C_j^{-1})x + \frac{1}{2}(\mu_j^T C_j^{-1}\mu_j - \mu_i^T C_i^{-1}\mu_i)\\ > \ln\frac{\pi_j(\det C_i)^{1/2}}{\pi_i(\det C_j)^{1/2}}.\end{aligned} \tag{40}$$

We see that the border between the two regions in d-dimensional space where we should or should not prefer ω_i to ω_j is given by a quadratic surface. When we allow the covariance matrices for the classes to vary we therefore talk about *quadratic discrimination* compared to the linear discrimination referred to above.

Example 2.13. *k-class discriminant analysis with estimated normal densities.*

Suppose that we have a training set with n_i objects from class ω_i, $i = 1, \ldots, k$. From all the classes we assume that we have obtained independent random samples of objects. We assume further that the vector X is normally distributed with expectation vector μ_i and covariance matrix C_i in class ω_i. Let the observations vectors be denoted X_{im}, $m = 1, \ldots, n_i$, $i = 1, \ldots, k$. Then it is natural to estimate the expectation vector in class ω_i by

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{m=1}^{n_i} X_{im}, \quad i = 1, 2. \tag{41}$$

If we make no assumption on equality of the covariance matrices we use the covariance matrix estimates

$$\hat{C}_i = \frac{1}{n_i - 1} \sum_{m=1}^{n_i} (X_{im} - \hat{\mu}_i)(X_{im} - \hat{\mu}_i)^T, \quad i = 1, \ldots, k. \tag{42}$$

If we assume equality of the covariance matrices we use instead the estimate

$$\hat{C} = \frac{1}{\sum_{i=1}^{k}(n_i - 1)} \sum_{i=1}^{k} (n_i - 1)\hat{C}_i \tag{43}$$

for the common covariance matrix C. □ □

2.4 Error rate estimation. Resubstitution and cross-validation

An important issue in discriminant analysis is to estimate the rates of misclassification errors. One simple type of error estimates, often called *resubstitution error-rate estimates*, is obtained by directly computing the observed error rates in the training set for the chosen allocation rule.

However, the resubstition error-rates are typically too optimistic as the objects used to evaluate the error rates are also used in the choice of the discriminator including estimation of parameters in the discriminator. Particularly if the discriminator is complicated, for instance if it contains many parameters, we can grossly underestimate the error-rate corresponding to classification of a new object.

One way of avoiding the bias of resubstitution error rates is to divide the available data into one training set and one evaluation set, for instance, by using half of the data for estimation and half of it for evaluation. One critisism of this procedure is that it may seem wasteful if data are scarce.

Nowadays one often uses resampling methods for evaluation of error rates. One such method is *k-fold cross-validation.* Then we divide the data set consisting of n objects into k equal or approximately equal groups, often by random choice of which objects that should go into group $j, j = 1, \ldots, k$. Then we fix j temporarily and use all objects except those in group j to estimate parameters and compute error average rates for all objects in group j. This procedure is repeated for all groups and we finally average error rates also over groups to get overall error rate estimates. Often $k = 5$ or $k = 10$ is recommended.

Example 2.14. *Handwritten digits. Digits 1 and 2*

We use the same data as in Example 1.11 with one small modification consisting of standardization of the two moment features by linear transformations so that they get average zero and varince one. We now use both liner and quadratic discrimination and get, respectively, the linear and elliptic boundaries shown in Figure 27. We also computed the resubstitution and 5-fold cross-validation errors for the liner and quadratic discrimination models. It turned out that all four error rate estimates were identical and equal to 15 %. □

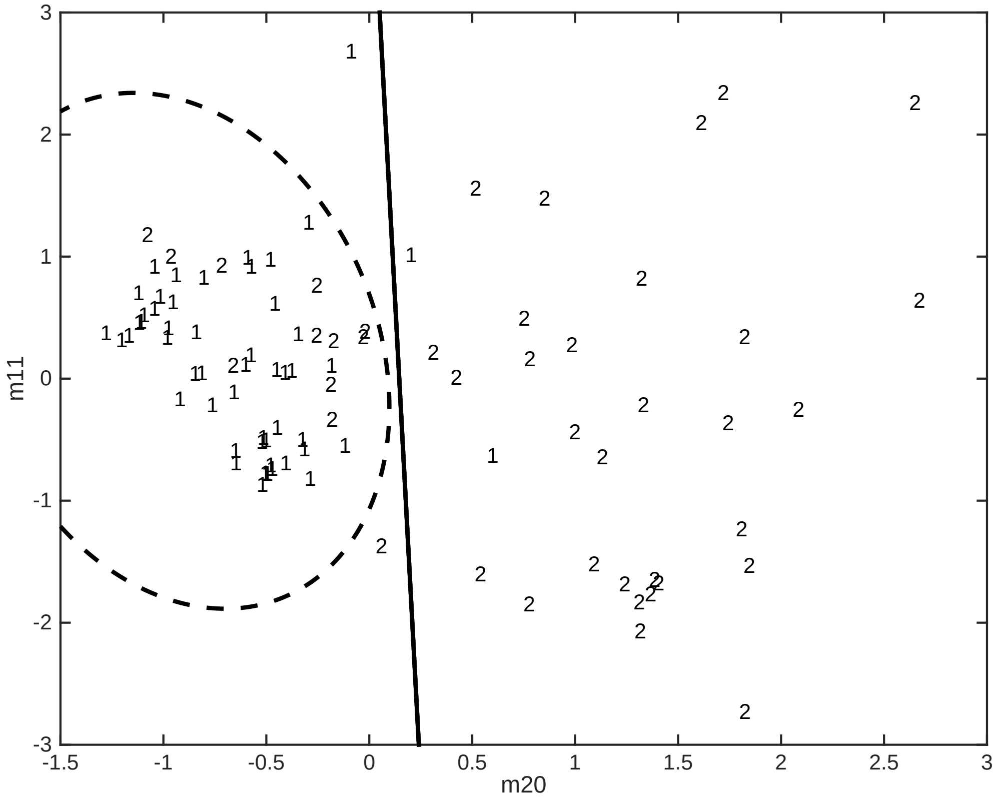

Figure 27: Plot of standardized moments μ_{11} versus μ_{20} for handwritten digits digits 1 and 2 among the the first 400 digits in the MNIST data base together with the class boundaries corresponding to linear and quadratic discrimination.

Example 2.15. *Handwritten digits. Moment features*

We use the first 8000 digits in the MNIST database, see Example 1.7, and consider discrimination between the 10 types of digits by use of all central moment features μ_{pq} in (28) with $p + q \leq K$. We computed the resubstitution and the 10-fold cross-validation error estimates for all $K \leq 13$, see Figure 28. Note that both for the linear discrimination full drawn curves and for the quadratic discrimination dashed curves the resubstitution errors are smaller than the cross-validation errors. For the linear discrimination the cross-validation minimum error is 12.3 % for order 12 and for the quadratic discrimination the cross-validation minimum error is 9.6 % for order 7.

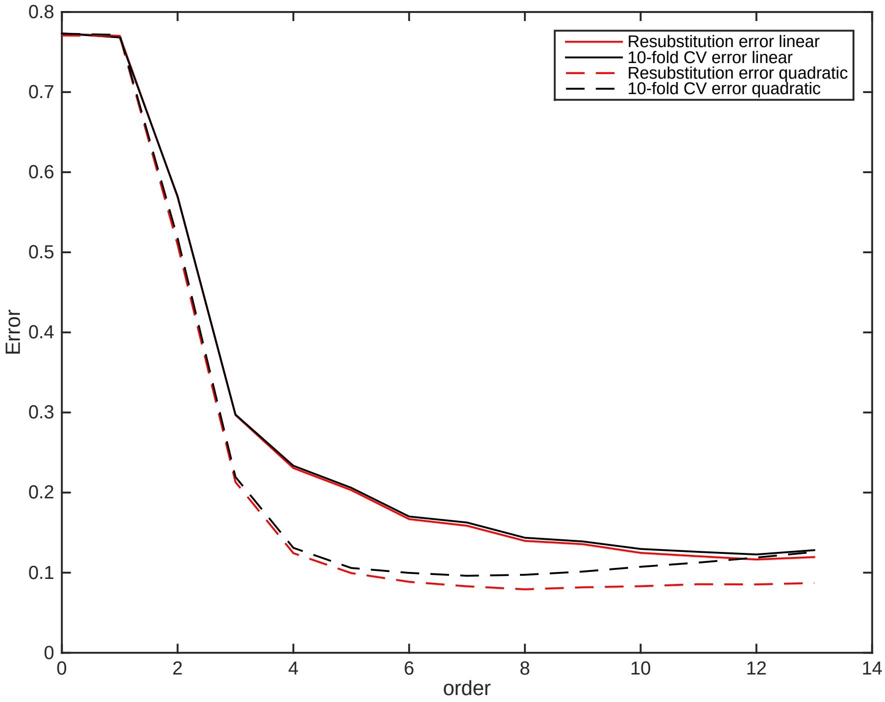

Figure 28: Plot of error probabilities for linear discrimination, full drawn curves, and quadratic discrimination, dashed curves. Resubstitution error curves are in gray and cross-validation error curves are in black. Order K on the horizontal axis means that all moments μ_{pq} with $p + q \leq K$ are used as features to discriminate between the digits.

□

2.5 Nearest neighbour classifaction

Suppose that we have a distance function $\delta(x, x')$ between feature vectors x and x'. Examples of distance functions for d-dimensional feature vectors are the Euclidean distance

$$\delta(x, x') = (\sum_{i=1}^{d} (x_i - x'_i)^2)^{1/2} \tag{44}$$

and $\delta = 1 - r$, where r are is the correlation

$$r(x, x') = \frac{\sum_{i=1}^{d} (x_i - \bar{x})(x'_i - \bar{x}')}{(\sum_{i=1}^{d} (x_i - \bar{x})^2)^{1/2} \ (\sum_{i=1}^{d} (x'_i - \bar{x}')^2)^{1/2}} \tag{45}$$

where $\bar{x}$ and $\bar{x}'$ are the arithmetic means of the vectors x and x'.

A useful discrimination method is the *m-nearest neighbour* rule, which proceeds as follows. Suppose we have a training set for which we know the correct classification. For a new observation we find the m nearest neighbours in the the training set, and we classify the new observation by majority voting among these nearest neighbours.

Example 2.16. *Handwritten digits. Nearest neighbour discrimination*

We use the same data as in Example 2.14. The m-nearest neighbour classications with m=3 and 5 are shown in Figure 29. We also computed the resubstitution and 5-fold

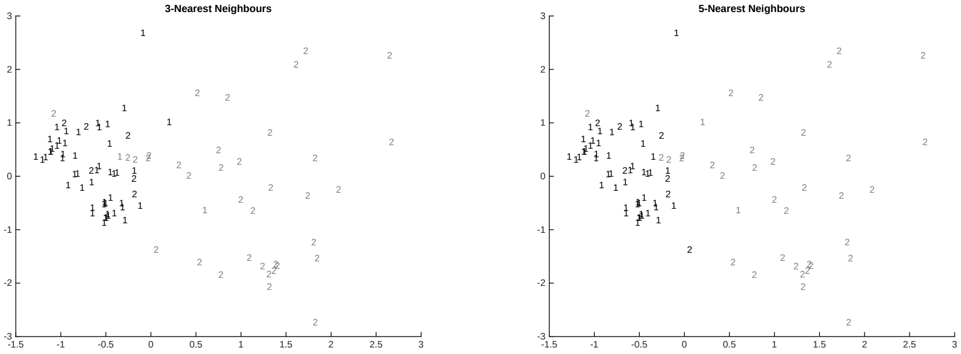

Figure 29: Plot of standardized moments μ_{11} versus μ_{20} for handwritten digits digits 1 and 2 among the the first 400 digits in the MNIST data base together classifications from m-nearest neighbour classification for $m = 3$ and $m = 5$. Digit colours indicate classification: black digits are classified as 1 and grey digits are classified as 2.

cross-validation errors for nearest m-neighbour methods with m ranging from 1 to 10. the result is shown in Figure 30. The minimum crossvalidated error is obtained for $m = 5$ and equals 12 %. □

2.6 Selection of features

If we have a large number of possible features it is useful to make a selection of features. One often used method is *forward selection* where we start by choosing the single feature which gives the smallest error rate. Then we add that feature of the remaining ones which together with the first chosen feature gives the best performance. The procedure is continued a suitable number of steps. If one uses cross-validation error rate estimates, we typically find that the error rates first decrease when we add new variables but then a minimum is obtained and after that the error rate increases due to overfitting.

In *backward selection* we start by including all features. Then we eliminate one feature so that the resulting error rate is as small as possible. The procedure is iterated a suitable number of steps.

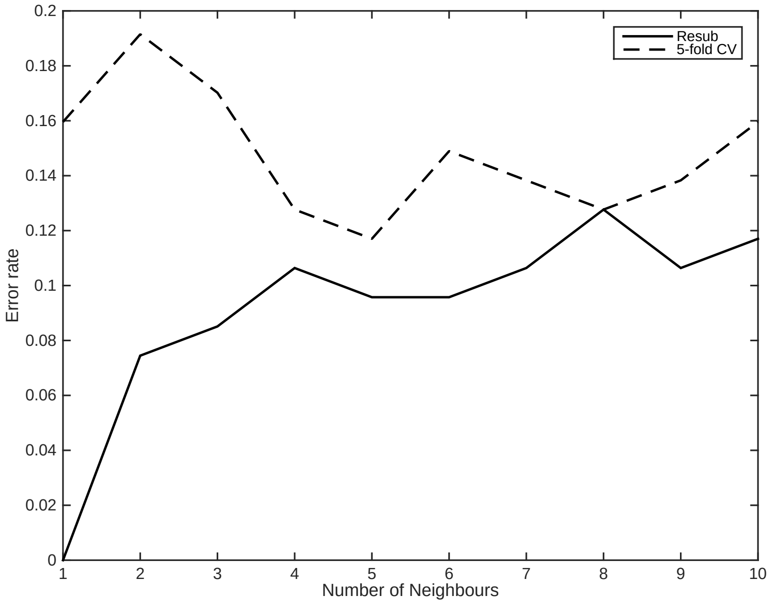

Figure 30: Plot of resubstitution and 5-fold cross validation error estimates for m-nearest neighbour classications for $m = 1, \ldots, 10$.

2.7 Cluster analysis, k-means clustering

Suppose that we have collected a number of colonies of bacteria of a type that has not been studied before but which we want to order in classes corresponding species or subspecies. That is, we want to construct a taxonomy for these bacteria. Instead of an individual bacterial particle the natural unit here is a homogeneous colony of bacteria.

One possible procedure would be to measure a number of variables, say d for each individual colony and to see if these variables tend produce clusters in d-space. Let X denote the d-dimensional vector of measurements, and let $f(x)$ denotes the corresponding probability density (or probability function if X is discrete). Corresponding to k classes we would then expect that f could be written as a mixture,

$$f(x) = \sum_{i=1}^{k} p_i f_i(x), \tag{46}$$

where f_i denotes the probability density in the ith class, and p_i the proportion of the ith class.

Let n denote the number of colonies observed, and let X_j, $j = 1, \ldots, n$, denote our observed d-dimensional vectors. The basic problem in cluster analysis can then be formulated as estimation of the number k of classes and also the functions f_i, $i = 1, \ldots, k$, on the basis of our observations $X_1, \ldots, X_n$. Note that this problem is much more complicated than the problems previously discussed in this chapter as we neither know the number of classes, nor which observations that belong to the different classes.

One procedure that is often used is k-means clustering. Consider d-dimensional observations and let us for simplicity regard Euclidean distances between observations. We assume that there are k classes and choose first randomly k cluster centers among the observations X_j, $j = 1, \ldots, n$. Then we alternate between two types of steps. In the *observation allocation step* we suppose that we have cluster centers $C_i, i = 1, \ldots, k$, and allocate each observation to the closest cluster center. In the *cluster center recomputation step* we compute new cluster centers as averages of all observations allocated to each cluster. We alternate between the two types of steps until there are no changes. Typically we will also repeat the procedure a number of times with different (randomly chosen) starting cluster centres and finally choose the clustering which has the the minimal total sum of within cluster square distances to cluster centres.

Example 2.17. *Handwritten digits. Cluster analysis*

We use the same data as in Example 2.14 but now we cluster them by k-means clustering with $k = 2$, 3 and 4. The results are shown in Figure 31.

□

2.8 Case studies

2.8.1 Weed seed identification

In Peterson (1992) weed seed identification was studied with 25 images of seeds for each of 40 species.

A large number of possible features were investigated and with 25 features an optimal cross-validation error rate of 2.3% was found.

2.8.2 Weed plant identification

Andersson (1998) studied identification of plants at an early stage of carrot and seven weed species. With 27 images for each of the eight plant species a cross-validation error rate of about 16% was found with 7 or 8 features.

2.8.3 Comparison of discrimination methods for microarray data

In Dudoit, Friedyland and Speed (2000) different discrimation methods are compared for classification of tumors based on gene expression data from three datasets available on the Internet. In particular, the nearest neighbour method is found to perform well in these examples. The number of neighbours is here determined by cross-validation.

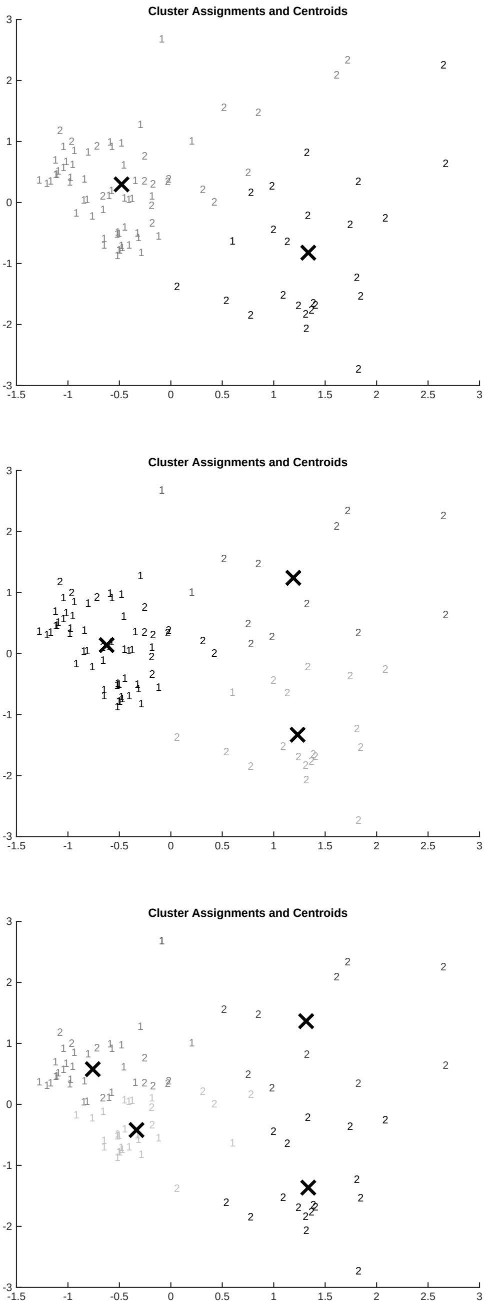

Figure 31: Results from k-means clustering with $k = 2$, 3 and 4 of the same data as used in Example 2.14.

2.9 Exercises

Images and data sets for the exercises below may be found from the course home pages.

Exercise 2.1. Fisher's Iris data, a classical data set. One of the famous data sets in statistics is Fisher's Iris data, used in Fisher (1936), where discriminant analysis was introduced. Consider the data in Table 1 with four variables measured for 50 plants of each of three *Iris* species. The data were assembled by E. Anderson, cf. Anderson (1935), and analysed in detail by Fisher (1936).

(a). Draw scatter plots for all 150 observations and all six pairs of variables. Alternatively, if you do not have access to a computer, draw scatter plots for subsets with, say, 5 plants from each species, and for, say, two pairs of variables.

(b). Find the best linear discriminators using all four variables for discrimination between all pairs of the three species. Alternatively, without a computer, describe with formulas how the computations are made. Under what assumptions is this discrimination method optimal.

(c). Find the best quadratic discriminators using all four variables for discrimination between all pairs of the three species. Alternatively, without a computer, describe with formulas how the computations are made. Under what assumptions is this discrimination method optimal.

(d). Find the optimal combination of two variables for discriminating between the three species. Alternatively, without a computer, describe with formulas how the computations are made.

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7	3.2	4.7	1.4	6.3	3.3	6	2.5
4.9	3	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4	1.3	6.3	2.9	5.6	1.8
5	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5	3.4	1.5	0.2	4.9	2.4	3.3	1	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5	2	3.5	1	6.5	3.2	5.1	2
4.8	3.4	1.6	0.2	5.9	3	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3	1.4	0.1	6	2.2	4	1	6.8	3	5.5	2.1
4.3	3	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5	2
5.8	4	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3	4.5	1.5	6.5	3	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6	2.2	5	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4	1.3	5.6	2.8	4.9	2
4.6	3.6	1	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5	3	1.6	0.2	6.6	3	4.4	1.4	7.2	3.2	6	1.8
5	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3	5	1.7	6.1	3	4.9	1.8
5.2	3.4	1.4	0.2	6	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1	7.2	3	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1	7.9	3.8	6.4	2
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.1	5.4	3	4.5	1.5	6.1	2.6	5.6	1.4
5	3.2	1.2	0.2	6	3.4	4.5	1.6	7.7	3	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.1	1.5	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3	1.3	0.2	5.6	3	4.1	1.3	6	3	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4	1.3	6.9	3.1	5.4	2.1
5	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4	1.2	5.8	2.7	5.1	1.9
5	3.5	1.6	0.6	5	2.3	3.3	1	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3	1.4	0.3	5.7	3	4.2	1.2	6.7	3	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3	5.2	2
5.3	3.7	1.5	0.2	5.1	2.5	3	1.1	6.2	3.4	5.4	2.3
5	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3	5.1	1.8

Table 1. Measurements of four variables (in cm) of flowers for 50 plants each of three *Iris* species, from Fisher (1936).

Exercise 2.2. Weed seeds. Consider the weed seed images of *Rumex crispus* and *Rumex thyrsiflorus* from Figures 5 and 6 in Example 1.2 or a subset of these 25 plus 25 images.
(a). Compute the areas of the seeds and the convexity of them for the images considered.
(b). How well can you discriminate between the two species by use of the feature convexity and linear discrimination?
(c). How well can you discriminate between the two species by use of the feature convexity and quadratic discrimination?
(d). How well can you discriminate between the two species by use of the features convexity and area and linear discrimination?
(e). How well can you discriminate between the two species by use of the features convexity and area and quadratic discrimination?

Exercise 2.3. Weed plants. Consider images of carrot and weed plants such as those described in Example 1.3. Choose two or more species and see well you can discriminate between them by suitably chosen featuers. Compare with the results found by Andersson (1998).

2.10 Literature on pattern recognition

A good introductory text on statistical pattern recognition is Fukunaga (1990). Many algorithms, including neural networks, are describe in Ripley (1996) which also contains an extensive list of references. A highly useful review of clustering methods with particular emphasis on applications with image data is given in Jain, Murty and Flynn (1999).