5 Discriminant Analysis

5.1 Introduction

a ▷ **Iris Species Example.** In Chapter 0.0.a it was demonstrated that there are significant differences between the two similar species versicolor and viginica. However, the goal which was stated in the Introduction went further (1.2.a): A rule was sought that would allow the plants to be classified into the three species with as few misclassifications as possible. For the existing data set the classification is known. So, the rule is not needed for these plants, but should allow other **plants to be classified correctly only on the basis of the measurements.** ◁

b **The General Model.** Each observed entity i is characterized by its membership in a class k_i and through the values $X_i^{(j)}$ of the variables. The class k_i determines the distribution of the random vectors $\underline{X}_i$, which we generally denote $\mathcal{F}_{k_i}$,

$$\underline{X}_i \sim \mathcal{F}_{k_i} \ .$$

(Unlike earlier, the class membership is not given with a double index hi for the ith observation in the group h, but instead with the categorial variable k_i.) The distributions $\mathcal{F}_k$, which characterize the classes, are usually given as parametric distributions; we will consider the case of normal distributions more closely.

c In one of the first variants of the model, the **class membership** k_i is interpreted as a **fixed, unknown number**. Earlier we have denoted such numbers as parameters. The situation, however, is different than that of parametric models in the usual sense, since for each observation a new parameter comes in. We call such variables **incidental parameters**.

d In a second version of the model, the **class membership** is itself a **random variable** K_i, and the model includes the g probabilities $P\langle K_i = k\rangle =: \pi_k$ of the memberships to the g groups. The distributions $\mathcal{F}_k$ are then the conditional distributions of $\underline{X}_i$, given $K_i = k$.

e If the distributions $\mathcal{F}_k$ and, in the second variant, the probabilities π_k are known, we can use these to derive rules saying how to **use the values of $\underline{X}$ to determine the class membership k.** This task is also known as **identification analysis**. In application, we usually have to estimate the parameters of the distributions $\mathcal{F}_k$ from the data. For this, a dataset is needed for which not only the variable values $\underline{x}_i$, but also the class memberships k_i are known. This data set is called the **training data set**.

f Here we want to develop the idea using the usual simple model, assuming that classes are comprised of multivariate normally distributed data $\underline{X}_i$, which only differs in expected value $\underline{\mu}_k$, so

$$\underline{X}_i \sim \mathcal{N}_m\langle \underline{\mu}_{k_i}, \mathbf{\Sigma} \rangle \ .$$

How the parameters $\underline{\mu}_k$ and $\mathbf{\Sigma}$ should be estimated from the training data was already stated for $g = 2$ groups in the context of the two sample problem (4.3.c and 4.3.e). In general, the expected values $\underline{\mu}_k$ are estimated through the means

$$\overline{\underline{X}}_k = \tfrac{1}{n_k} \sum\nolimits_{\{i|k_i=k\}} \underline{X}_i$$

The estimation of $\mathbf{\Sigma}$ is

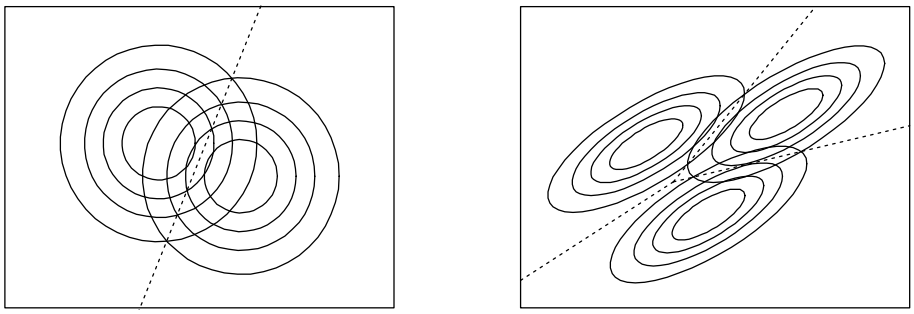
$$\widehat{\mathbf{\Sigma}} = \tfrac{1}{n-g} \sum\nolimits_{k=1}^{g} \sum\nolimits_{\{i|k_i=k\}} (\underline{X}_i - \overline{\underline{X}}_{k\cdot})(\underline{X}_i - \overline{\underline{X}}_{k\cdot})^T = \tfrac{1}{n-g} \sum\nolimits_{i} (\underline{X}_i - \overline{\underline{X}}_{k_i\cdot})(\underline{X}_i - \overline{\underline{X}}_{k_i\cdot})^T \ .$$

We can thus immediately take care of the identification analysis. Here we take the parameters as fixed. The consequences of the imprecision introduced by estimating the parameters will be studied later.

5.2 Classification with Known Distributions

a An observation $\underline{x}_0$ is given for which the class membership k_0 is not known, but is to be determined. On the basis of the observed features, a **decision** should be made between g possible classes. We can also say that the "discrete parameter" k_0 **is to be estimated** from the observation $\underline{x}_0$. As just announced, we will treat the case of normal distributions.

b In the simplest case, the classes follow a multivariate normal distribution, with equal covariance matrices but different expected values, $\underline{X}_0 \sim \mathcal{N}\langle \underline{\mu}_k, \mathbf{\Sigma} \rangle$. Fig. 5.2.b illustrates this model for $\mathbf{\Sigma} = \boldsymbol{I}$ and for general $\mathbf{\Sigma}$.



(i) (ii)

Figure 5.2.b: Models (i) for two groups with $\mathbf{\Sigma} = \boldsymbol{I}$ and (ii) for three groups with general covariance matrix

In the first case, there is a "most natural" classification rule: The observation $\underline{x}_0$ is assigned to the class for which the distance to the "class center" $\underline{\mu}_k$ is the smallest. For general $\boldsymbol{\Sigma}$ it makes sense to measure the distance via the Mahalanobis distance $d\langle \underline{x}_0, \underline{\mu}_k; \boldsymbol{\Sigma} \rangle$.

c **Two Classes, Equal $\boldsymbol{\Sigma}$.** In the case of two classes, the rule is especially simple. An observation is then assigned to the second class if the difference of the squared distances $d^2\langle \underline{x}_0, \underline{\mu}_1; \boldsymbol{\Sigma} \rangle - d^2\langle \underline{x}_0, \underline{\mu}_2; \boldsymbol{\Sigma} \rangle$ is positive. The difference can be written as

$$\begin{aligned} d^2\langle \underline{x}_0, \underline{\mu}_1; \boldsymbol{\Sigma} \rangle - d^2\langle \underline{x}_0, \underline{\mu}_2; \boldsymbol{\Sigma} \rangle &= (\underline{x}_0 - \underline{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\underline{x}_0 - \underline{\mu}_1) - (\underline{x}_0 - \underline{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\underline{x}_0 - \underline{\mu}_2) \\ &= 2(\underline{\mu}_2 - \underline{\mu}_1)^T \boldsymbol{\Sigma}^{-1} \underline{x}_0 + \underline{\mu}_1^T \boldsymbol{\Sigma}^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \boldsymbol{\Sigma}^{-1} \underline{\mu}_2 \ . \end{aligned}$$

The last two terms do not depend on $\underline{x}_0$, so together they form a constant. The first is of the form $\underline{\beta}^T \underline{x}_0$.

Together these two parts thus form a linear function

$$h\langle \underline{x}_0 \rangle = \alpha + \underline{\beta}^T \underline{x}_0 \ ,$$

which serves for discriminating between the classes according to the rule

$$\widehat{k}\langle \underline{x}_0 \rangle = \begin{cases} 1 & \text{if } \ h\langle \underline{x}_0 \rangle < 0 \\ 2 & \text{if } \ h\langle \underline{x}_0 \rangle > 0 \end{cases} \ .$$

The function h is called the **discriminant function** – more precisely: Fisher's Linear Discriminant Function, since it was created by this gentleman.

d For practical application, we have still have to estimate the parameters from training data for which the class is known (5.1.e). If we use the above (5.1.f) named estimations, we get **Fisher's Linear Discriminant Analysis** for two groups.

e ▷ Fig. 5.2.e shows that the values of the discriminant function with estimated parameters achieves a fairly good separation of the two similar species in the iris example. We will come back to the incorrect classifications. ◁

f **Logistic Regression.** The linear discriminant function reminds us of a linear regression model. In regression, we want to be able to "predict" a continuous variable of interest Y from the values of the explanatory variables $\underline{x}$. Here we would like to determine the class membership k. The difference is only that the variable of interest is no longer continuous, but two-valued or binary. In fact, in this section we haven't treated it as a random variable. We now want to rectify this.

A binary random variable Y with possible values 0 and 1 is characterized by the probability $\pi = P\langle Y = 1 \rangle$. In the present context this probability should depend on $\underline{x}$. The probability that an observation with feature values $\underline{x}$ belongs to class 2 – which is what $Y = 1$ means – is larger if the observation lies nearer to the mean point $\underline{\mu}_2$ of group 2 and further from the center $\underline{\mu}_1$ of group 1. It makes sense to set the probabilities for $Y = 1$ and $Y = 0$ proportional to the corresponding probability densities $f_k\langle \underline{x} \rangle$ ($k = 2$ and $= 1$, respectively). Then,

$$\log \left\langle \frac{P\langle Y = 1 \rangle}{P\langle Y = 0 \rangle} \right\rangle = \log \left\langle \frac{f_2\langle \underline{x} \rangle}{f_1\langle \underline{x} \rangle} \right\rangle = h\langle \underline{x} \rangle \ .$$

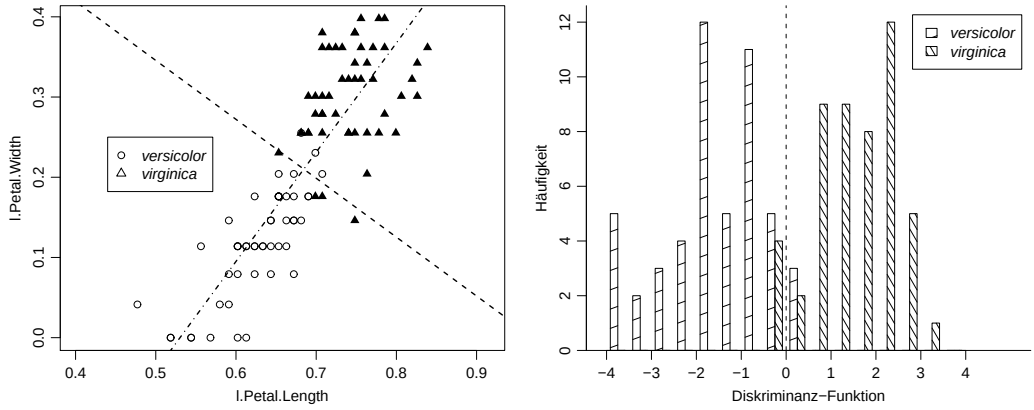


Figure 5.2.e: Values of the discriminant function with estimated parameters for two iris species

In the case of the normal distribution with equal covariance matrices we get directly the difference of the Mahalanobis distances, which according to 5.2.c is a linear function $\alpha + \underline{\beta}^T \underline{x}$ of $\underline{x}$.

This is the model of **logistic regression**. The ratio "probability : Counter probability is also known as **odds**, commonly used in English, especially by people who bet on horses. The logistic regression model expresses the "log odds" as a linear function with the input variables or regressors $X^{(j)}$.

g The methodology of logistic regression leads to a direct estimation of α and $\underline{\beta}$, which is not based on estimation of the parameters $\underline{\mu}_k$ and $\boldsymbol{\Sigma}$. Specifically, in the model, as in classical multiple linear regression **no assumptions about the regressors** $X^{(j)}$ are made. The strong assumption that the data of the two classes have a multivariate normal distribution, not to mention equal covariance matrices, is thus not used. We only use the assumption that the "log odds" are a *linear* function of the explanatory variables $X^{(j)}$. From linear regression we know, however, that many relationships that initially do not appear linear can be modeled with this approach. Key words are **transformations** and interactions.

Thus, in practice, **logistic regression is preferable to linear discriminant analysis.**

h **Multiple Classes.** In the case of multiple classes with multivariate normally distributed features $X^{(j)}$ and equal covariance matrices $\boldsymbol{\Sigma}$, as mentioned (5.2.b), we seek the class with the minimal squared Mahalanobis distance
$d^2\langle \underline{x}_0, \underline{\mu}_k; \boldsymbol{\Sigma}\rangle = (\underline{x}_0 - \underline{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\underline{x}_0 - \underline{\mu}_k)$. For two classes, the classification rule can be expressed with the linear discriminant function h. Can some analogous be found for more classes?

The simplest case is, as before, if $\boldsymbol{\Sigma} = \boldsymbol{I}$. The Mahalanobis distance is then the usual (Euclidean) distance. For three dimensional data and three groups, the geometrical

representation helps with clarification. To determine to which of the three class centers a point $\underline{x}_0$ lies nearest, we can project the point onto the plane that goes through the three centers. How far the point is from the plane has no influence on which class it will be assigned to – despite this, it is not completely meaningless for the classification, as we will discuss later. – So, for three groups, two dimensions suffice to determine the classification. Analogously, for g classes, the g centers lie in an (at most) $g-1$ dimensional subspace, and this space is sufficient for the classification. (If fewer than g variables are available, the entire m dimensional space is required.)

Such a subspace is determined by $g-1$ vectors. We can, for example, choose the vectors $\underline{\mu}_k - \underline{\mu}_1$, $k = 2, 3, ..., g$. Better suited are vectors that are perpendicular (orthogonal) to each other, since then the distances, which are decisive for the classification, then remain the same. If we use these as the basis of the space, which means if we project the observations onto such vectors and use the projections as coordinates, then the usual distance in the subspace can thus be used for the class assignment.

We handle the case of a general covariance matrix $\boldsymbol{\Sigma} \neq \boldsymbol{I}$ by turning back to the just discussed special case: We first transform observations $\underline{X}$ with a matrix $\boldsymbol{B}$ so that for the transformed observations the covariance matrix becomes $= \boldsymbol{I}$.

i For g classes, there are $g-1$ **discriminant functions**, that determine the space in which the class membership is determined.

They are, however, still not uniquely determined with this essential property, since we can arbitrarily orthogonally transform within the subspace. Conventionally, they are determined so that they form the principal components of the class centers in the subspace.

j ▷ In the **iris species example** there are in total 3 species to differentiate. This leads to 2 discriminant functions, whose values are shown in Fig. 5.2.j . The discriminant functions are chosen so that the estimated covariance matrix within the classes (analogous to 5.1.f) is the identity matrix. Thus, the separation lines between the classes are given by the mean perpendiculars to the lines between the centers. ◁

The coefficients $\underline{\widehat{\beta}}$ of the discriminant functions are

	Sepal leaves		Petal leaves	
D.f. 1	8.70	9.07	-20.779	-3.529
D.f. 2	-9.85	-15.18	-0.713	0.313

The first discriminant function essentially forms a contrast between the length measurements of the sepal leaves and the width of the petal leaves, the second between the width of the sepal leaves and the two length measurements of the petal leaves.

k **Multinomial Regression.** As the case of two classes leads to the logistic, discriminant analysis for multiple classes leads to multinomial regression, which again waives assumptions about the distribution of the variables $X^{(j)}$ and is thus preferable in practice.

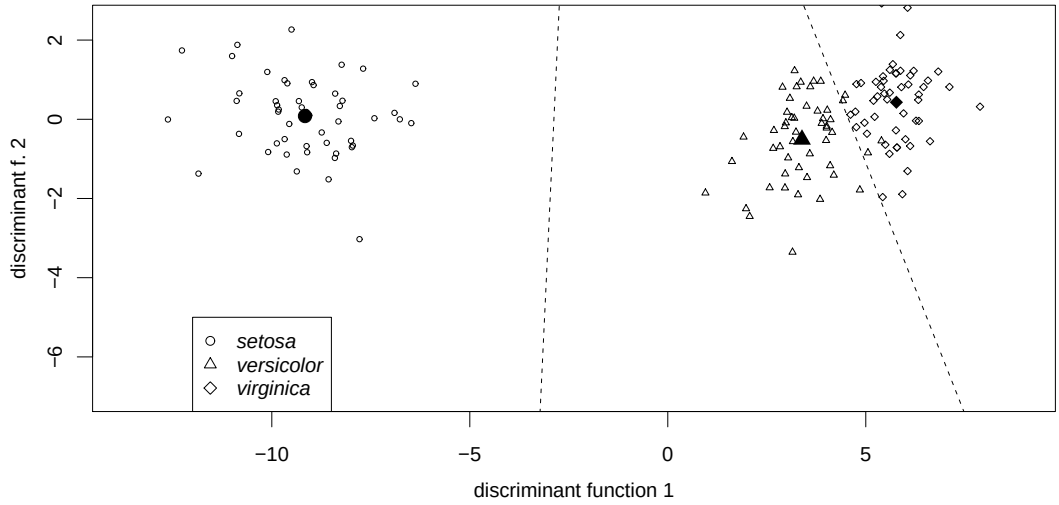


Figure 5.2.j: The two discriminant functions in the iris species example. The means of the groups are shown with filled symbols.

l **Model Choice.** In linear regression, the following steps under the keyword **residual analysis** are important:

- Checking model assumptions,
- Selecting input (or explanatory) variables,
- Modeling nonlinear dependencies or interactions.

In the linear discriminant analysis discussed here, the assumption of the multivariate normal distribution must be checked. Transformations can help to avoid deviations as much as possible. For discriminant analysis it is also reasonable to remove variables. So, model selection is generally necessary here too.

These questions become even more similar to the case of usual linear regression if, instead of linear discriminant analysis, **logistic regression** or multinomial regression, respectively, is used. As mentioned, we do not need a multivariate normal distribution for the variables $X^{(j)}$ and so the model development is as free as in the usual linear regression. We don't go into any more detail in this chapter.

m **General Distribution.** How should the idea of linear discriminant analysis be generalized for **other assumed distributions**? To estimate k_0, we can fall back on the idea of maximum likelihood. So, if the $\mathcal{F}_k$ have density f_k, $\widehat{k}_0$ should be the k with the maximal density $f_k\langle \underline{x}_0 \rangle$,

$$\widehat{k}_0 = \text{argmax}_k \left\langle f_k\langle \underline{x}_0 \rangle \right\rangle \ .$$

n For the case of multivariate normally distributed data with **unequal covariance matrices** $\mathbf{\Sigma}_k$ a simple rule arises: The value $-2\log\langle f_k \rangle$ equals the squared Mahalanobis distance plus a term that depends on $\mathbf{\Sigma}_k$. Instead of maximizing f_k, we determine

$$\arg\min_k \left\langle d^2\langle \underline{x}_0, \underline{\mu}_k; \mathbf{\Sigma}_k \rangle + \log\langle \det\langle \mathbf{\Sigma}_k \rangle \rangle \right\rangle \ .$$

As in the case of equal covariance matrices, it is useful to illustrate in the case of $g = 2$ classes, how the regions look, in which an observation $\underline{x}_0$ is assigned to one or the other of the classes. The border is given by

$$\begin{aligned} d^2\langle \underline{x}, \underline{\mu}_2; \hat{\mathbf{\Sigma}}_2\rangle - d^2\langle \underline{x}, \underline{\mu}_1; \hat{\mathbf{\Sigma}}_1\rangle - c &= \\ (\underline{x} - \underline{\mu}_2)^T \hat{\mathbf{\Sigma}}_2^{-1}(\underline{x} - \underline{\mu}_2) - (\underline{x} - \underline{\mu}_1)^T \hat{\mathbf{\Sigma}}_1^{-1}(\underline{x} - \underline{\mu}_1) - c &= 0 \end{aligned}$$

with $2c = \log\langle\det\langle \hat{\mathbf{\Sigma}}_2\rangle\rangle - \log\langle\det\langle \hat{\mathbf{\Sigma}}_1\rangle\rangle$. This can be expressed analogously to the calculation in 5.2.f in the form $\underline{x}^T(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}_2)\underline{x} + \underline{\beta}^T\underline{x} + \alpha = 0$, so as a quadratic equation in $\underline{x}$. We therefore call this classification method **quadratic discriminant analysis**.

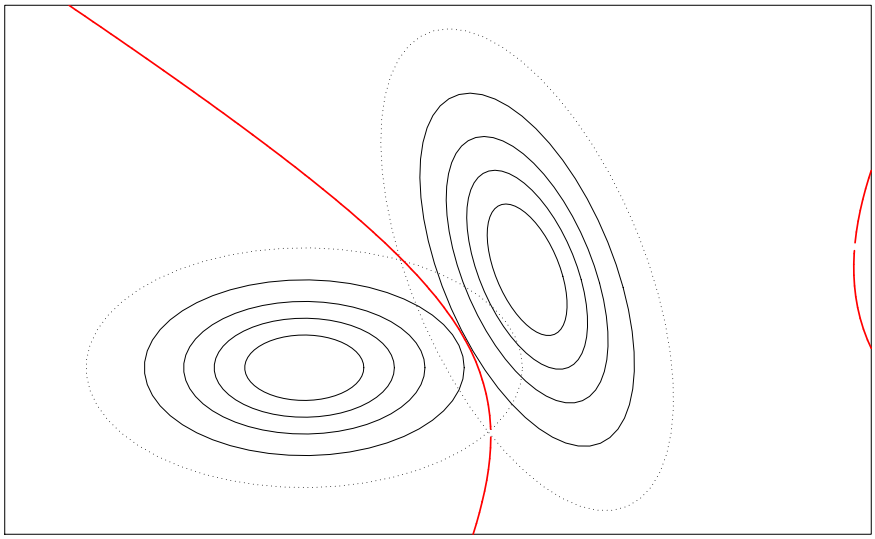


Figure 5.2.n: Two groups with unequal covariance matrix, with (only approximately calculated) quadratic separation lines of the classification

In two dimensions, the separation line between the areas in the classification into two groups is a hyperbola (Fig. 5.2.n). The curve on the right edge of the figure shows that far to the right the density of the "group on the left" becomes larger than that of the group with the closer center. For a reasonable classification this is of questionable value. Both densities are so small there that an observation appearing there would be very unlikely for both groups.

o In practice, as before, we must estimate the parameters. Since for each class there is now an entire covariance matrix $\hat{\mathbf{\Sigma}}_k$ to estimate, this is only reasonable if we have a lot of training data. Therefore, linear discriminant analysis is often better, even if the assumption of equal covariance matrices is not justifiable, as long as the covariance matrices are not too extremely different.

5.3 Error Rates

a In the introduction (1.2.c), diagnostic tests in medicine were discussed that serve to divide patients into sick and healthy in respect to a certain illness. In this case there are two types of errors with clearly different consequences: If healthy people are classified as sick by the test, this leads to unnecessary worry and, if there is no more precise clarification that can reveal the error, to unnecessary treatment. This is usually not as bad as the opposite error: If a sick person is declared to be healthy, a vital treatment can be missed. It therefore makes little sense to total the errors and come up with a single "error rate".

b **Sensitivity and Specificity.** You probably are familiar with the confusing medical jargon: If a test classifies you as sick, then the test result is called *positive*. Healthy people that are classified as sick are therefore called **false positives**, and sick people who are declared to be healthy are the **false negatives**. There are two reasonable ways to express their counts: If we refer to the overall number of positive or, respectively, negatives, we get the

$$\text{False positve rate} = \frac{\text{Number of false positives}}{\text{Number of positives}},$$

$$\text{False negative rate} = \frac{\text{Number of false negatives}}{\text{Number of negatives}}.$$

For patients that get a "positive" result, the first rate represents the **conditional probability** that they are healthy despite the test result.

For assessing a method, it is more meaningful to relate the results to the true status of the patients. We stress the correct results and choose the terms

$$\text{Sensitivity} = \frac{\text{Number of sick positives}}{\text{Number of sick}}$$

$$\text{Specificity} = \frac{\text{Number of healthy negatives}}{\text{Number of healthy}}$$

The sensitivity measures the (conditional) probability that a sick person will be classified as such, while the specificity gives the (conditional) probability that a healthy person will not get a false alarm. These two values characterize the"selectivity" of the medical tests. They are not influenced by the proportion of sickness in the population, the "**prevalence**". In contrast, the false positive rate is smaller if the prevalence increases, while the rate of false negatives gets larger.

c **Variable Thresholds.** For the case of 2 classes, linear discriminant analysis and logistic regression determine a linear discriminant function $h\langle \underline{x} \rangle = \widehat{\alpha} + \underline{\widehat{\beta}}^T \underline{x}$. The classification is done with a comparison with a constant c that, according to 5.2.c, is 0. There are good reasons to set the threshold otherwise

- if one of the classes is more frequent than the other and
- if costs associated with a misclassification turn out to be different: To mistakenly classify a sick person as healthy can be fatal, whereas classifying a healthy person as sick (until closer study) carries less weight. (If therapies may or may not be successful, the assessment may be reversed.)

Such concepts are clarified in **decision theory**, see section 5.4.

Often it is simplest to only determine the values of the discriminant function and determine the thresholds according to a pragmatic application-based point of view. For example, in an advertising campaign, the size of a mailing can be determined in advance, and we would serve (or bother) the corresponding number of address data with the highest scores. For choosing the threshold, we often use the number of expected misclassifications, as we'll discuss in the following.

d **Sensitivity and Specificity Curves.** Via the choice of the threshold c for the discriminant function, the sensitivity can be arbitrarily increased at the cost of the specificity – in the extreme case, we would declare everyone to be sick to be sure not to miss anyone! Coversely, by raising c we can raise the specificity. We would then avoid worrying healthy people, but would miss recognizing sick people. It is reasonable to consider both values as a function of the threshold c. Fig. 5.3.d shows this for the vein constriction example.

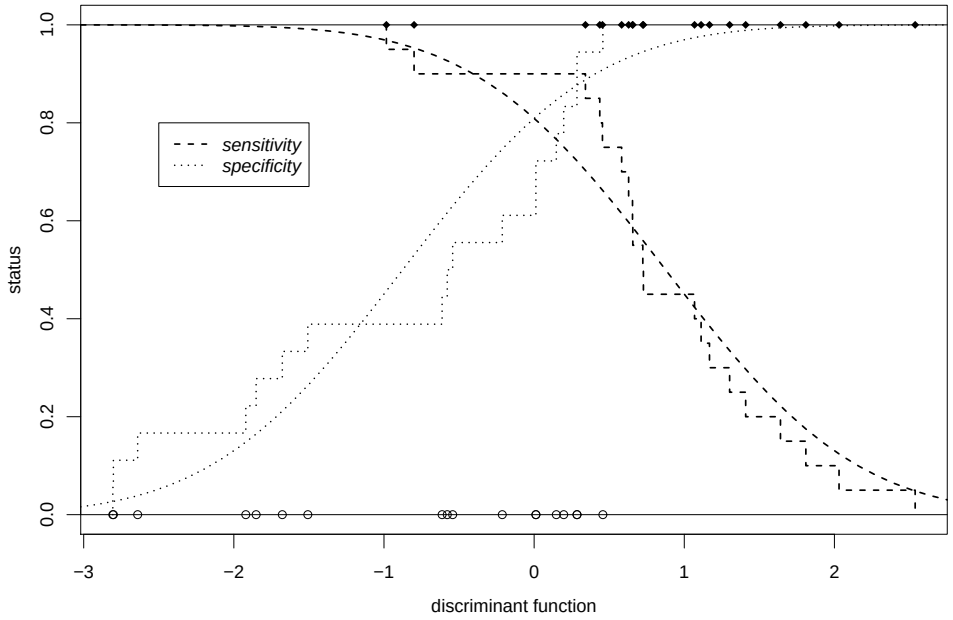


Figure 5.3.d: Sensitivity and specificity in the vein constriction example. The step curves represent the empirical error rate, the smooth curves stand for the rates corresponding to the estimated model.

The sensitivity and specificity curves are a very informative characterization of a method. They allow

- the threshold to be chosen on the basis of a reasonable compromise between the two criteria,
- the decision to be refined: We can introduce two thresholds c_0 and c_1, between which we declare a region of doubt in which further clarification must be made. Then, only for values $< c_0$ does a conclusion of "healthy" result, and for values $> c_1$, a conclusion of "sick" .

e **Error Rates.** For determining the sensitivity and the specificity – or the corresponding error rates – there are multiple methods, which are discussed in the following. We stick with the case of two classes and consider a fixed assignment rule $\widehat{k}\langle \underline{x} \rangle$, for example the linear discriminant function with a fixed threshold c. The error probability which corresponds to the sensitivity is the probability

$$Q_{k\ell} = P\Big\langle \widehat{k}\langle \underline{X} \rangle = \ell \mid \underline{X} \sim F\langle \underline{\theta}_k \rangle \Big\rangle$$

with $k = 1$ and $\ell = 2$, and vice versa for the specificity.

f **Appearant error rate** . The simplest estimation of the error rate comes from the relative error frequency in the training data,

$$Q_{k\ell}^{app} = \#\{i \mid \widehat{k}\langle \underline{X}_i \rangle = \ell \ , \ k_i = k\} \ / \ n_k \ ,$$

where n_k is the number of observations in the group k, so with $k_i = k$, (and $\#\{i \mid ...\}$ is the *number* of is, for which ... holds).

For an overall assessment of the rule, we can combine the two error rates Q_{12}^{app} und Q_{21}^{app}. We do this with weights which correspond to the number of observations in the two classes, and thus obtain the simple overall error rate

$$Q^{app} = (\#\{i \mid \widehat{k}\langle \underline{X}_i \rangle = 2 \ , \ k_i = 1\} + \#\{i \mid \widehat{k}\langle \underline{X}_i \rangle = 1 \ , \ k_i = 2\})/n \ .$$

g **Theoretical Error Rate.** If we know the distributions $F\langle \underline{\theta}_k \rangle$ – including parameters θ_k – then we can calculate the theoretical error rate.

A possible simple case is the model $\underline{X}_i \sim \mathcal{N}_m\langle \underline{\mu}_{K_i}, \boldsymbol{I} \rangle$ with $\underline{\mu}_1 = [0,0]^T$, $\underline{\mu}_2 = [\Delta, 0]^T$, cf. Figure 5.2.b (i). The obvious classification rule says $\widehat{k}\langle \underline{x} \rangle = 1$, if $x^{(1)} < 0$, and otherwise $\widehat{k}\langle \underline{x} \rangle = 2$. The error rates for this rule can be calculated for the model simply: They are both equal, $Q_{12} = Q_{21} = Q = \Phi\langle -\Delta/2 \rangle$. This result is also true for general $\underline{\mu}_k$ and $\boldsymbol{\Sigma}$ if we set $\Delta^2 = (\underline{\mu}_2 - \underline{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\underline{\mu}_2 - \underline{\mu}_1)$.

h If we don't know the parameters θ_k, we can plug in the estimators based on the training data and get a "**parametrically estimated error rate**" $\widehat{Q} = \Phi\langle -\widehat{\Delta}/2 \rangle$.

i It is plausible that this error rate gives a **too optimistic result** since the rule has been optimally fitted on the training data. (An analogous discovery is made in multiple regression, where we see that the residuals have a somewhat smaller scatter than the random deviations.) The same holds for the empirical error rate.

This difficulty can tackled with two ideas that are also useful in other situations and are thus fundamental.

j **Test Data.** The first is basically simple: We determine the error rate with help of "new" data, which was not used for the estimation of the classification rule. To differentiate them from the training data, we call such observations "test data". At the beginning of an analysis, we can decide to randomly divide the existing data into training data and test data so that at the end the usefulness of the results can be realistically estimated. This is certainly a good strategy if we have a data source that gives large numbers of observations, as is typically the case in data mining.

k **Cross Validation.** In many applications, the data set is limited, and it would be unreasonable to use only a part of it for the estimation of the rule – only to be able to correctly assess its precision later. (* Here, correctly means "without systematic error". If the test data set is small, the random errors prevail in the estimation of the precision!) Here, a refined version of the previous idea helps:

If we do not use an observation X_i for the estimation of the decision rule, the probability of a misclassification remains intact. We thus leave out the ith observation and derive the rule with the remaining $n-1$ training data. Now we determine whether the rule correctly assigns the ith observation. If we don't shy away from the effort, we can do this for each observation. If we now count the number of misclassifications, there is no systematic optimism any more.

The classification rule that is determined without the ith observation is denoted as $\widehat{k}_{[-i]}$. Then the estimated error rate is

$$Q_{cv1} = (\#\{i \mid \widehat{k}_{[-i]}\langle \underline{X}_i \rangle = 2 \text{ and } K_i = 1\} + \#\{i \mid \widehat{k}_{[-i]}\langle \underline{X}_i \rangle = 1 \text{ and } K_i = 2\} \; .$$

* For the calculation of the rules with the ith observation we usually don't have to repeat the entire calculation effort. There are so-called "update"formulas.

The idea can be altered so that not just one, but multiple observations are be left out to redetermine the rule, and then the misclassifications are determined for all excluded observations, and this step is repeated with other sets of exclusions a large number of times.

L **Literature:** Rencher (1998), Sec. 6.4

5.4 * Decision Theory

a In some situations in which observations should be classed, we know something about the probability of the individual classes. In the iris flowers example it could be that we know the frequency of occurrence of the three species in a region under study. Then, in cases of doubt, we would be tend to assign a plant to the more frequent of the two species in question.

This situation is described by a model in which the **class membership is a random variable** K (5.1.d). We now understand the distributionin $\mathcal{F}_k$ for the classes k as **conditional distribution** of the observation $\underline{X}_i$, given that it belongs to class k. For a complete model we additionally need the probabilities $P\langle K_i = k\rangle$ of the classes. This is summarized in formulas as

$$(\underline{X}_i \mid K_i = k) \sim \mathcal{F}_k \; , \qquad P\langle K_i = k\rangle = \pi_k \; .$$

The distribution of $\underline{X}_i$ for unknown classes is then a so-called **mixture distribution**. If the distributionin $\mathcal{F}_k$ of the classes have density f_k, then the density of the mixture distribution is

$$f\langle \underline{x}\rangle = \sum\nolimits_k \pi_k f_k\langle \underline{x}\rangle \; .$$

b **Bayesian Approach.** Now, to which class should an observation with feature values $\underline{x}_0$ be assigned? In the language of the model, we ask about the random variable K_0, if $\underline{X}_0$ is given. The conditional probabilities of $K_0 = k$, given the features $\underline{X}_0 = \underline{x}_0$, can be calculated

$$P\langle K_0 = k \mid \underline{X}_0 = \underline{x}_0\rangle = \frac{dP\langle K_0 = k \cap \underline{X}_0 = \underline{x}_0\rangle}{dP\langle \underline{X}_0 = \underline{x}_0\rangle} = \frac{\pi_k f_k\langle \underline{x}_0\rangle}{\sum_\ell \pi_\ell f_\ell\langle \underline{x}_0\rangle} \ .$$

(The notation dP means that it does not concern probabilities, but really probability densities.) We have already found such a formula in the introduction. From the conditional probabilities of an event B, given each of the events A_k, the probabilities of the A_k, given B, were calculated. That formula was equivalent to the one just given and was called **Bayes' theorem**.

In this context, the probabilities π_k of the classes are called **a priori probabilities** or **prior probabilities** and the conditional probabilities $P\langle K_0 = k \mid \underline{X}_0 = \underline{x}_0\rangle$, as **a posteriori probabilities** or **posterior probabilities** – the former are valid *before* we know the feature values $\underline{x}_0$, the latter, *afterwards.*

c By this, we have sketched the **basic scheme of Bayesian statistics**: It starts from the assumption that we know something about the unknown variable K before we get observations, and that this prior knowledge can be expressed as a probability distribution, the "prior distribution". This prior knowledge may stem from earlier studies or from a subjective assessment. It reflects the state of knowledge before an observation is made (or becomes known). By obtaining an observation (or many), the knowledge increases, and applying Bayes' theorem, the new state of knowledge can be calculated as the posterior distribution (probability of K.

This scheme can be and is applied to the parameters of any given parametric model. The parameter is no longer regarded as a fixed, unknown number, but as a random variable, for which a prior distribution is postulated. Based on an observation or a whole sample Bayes' theorem is used to obtain a posterior distribution for the parameter.

This type of reasoning, called "Bayes statistics", has found wide applications in many areas.

d **Bayes' Classification Rule.** probabilities define a precise description of the knowledge about class membership. A natural classification rule is to assign the observation $\underline{x}_0$ to the class with highest posterior probability resulting in

$$\widehat{k}_0 = \arg\max_k \left\langle P\langle K_0 = k \mid \underline{X}_0 = \underline{x}_0\rangle \right\rangle = \arg\max_k \left\langle \frac{\pi_k f_k\langle \underline{x}_0\rangle}{\sum_\ell \pi_\ell f_\ell\langle \underline{x}_0\rangle} \right\rangle \ .$$

Since the denominator is the same for all classes, the rule can be simplified,

$$\widehat{k}_0 = \arg\max_k \langle a_k\langle \underline{x}_0\rangle\rangle \quad \text{mit} \quad a_k\langle \underline{x}\rangle = \pi_k f_k\langle \underline{x}\rangle \ .$$

In case of equal prior probabilities π_k, this rule coincides with the rule that appeared natural in the "non-Bayesian" type of reasoning. The general case demonstrates the simplest way of taking the prior knowledge about the probabilities into account when making decisions.

e* **Expected Error Rate.** Up to here we have characterized rules as "natural". Mathematicians – and many others – prefer more precise statements. In general, we certainly strive for a classification rule that has a low error rate. Thus, we want to minimize the expected error rate

$$P\langle\widehat{k}\langle\underline{X}_0\rangle \neq K_0\rangle = 1 - P\langle\widehat{k}\langle\underline{X}_0\rangle = K_0\rangle = 1 - \sum\nolimits_k \pi_k P\langle\widehat{k}\langle\underline{X}\rangle = k \mid K = k\rangle \ .$$

It is not difficult to prove mathematically that Bayes' decision rule is optimal in this sense if the assumptions hold.

f* **Cost of a misclassification.** As mentioned above, the error consisting of falsely declaring a sick person as healthy is often more severe than the opposite error to judge a healthy person as sick – even more so if the latter error can be uncovered by further tests. If "good" and "bad" debitors of a bank are to be distiguished by a classification rule, it is less costly to assign a good client to the bad group – leading to a more close surveillance of his behavior – than to classify a client which finally stops payments as good, missing the possibility to take counter measures in time. Quite generally, a missed alarm is more expensive than a false alarm.

In these cases, one can try and determine a cost value for the possible errors. Let $c_{k\ell}$ be the cost of estimating class $\widehat{k} = k$ for a case with true class $K = \ell$. In addition, different benefits can be defined for the different correct decisions.

With these specifications, it is obvious to seek the rule that minimizes the expectation of the cost. The optimal rule again has the structure $\widehat{k}_0 = \arg\max_k\langle a_k\langle\underline{x}_0\rangle\rangle$, but this time with $a_k\langle\underline{x}\rangle = -\sum_\ell c_{k\ell}\pi_\ell f_\ell\langle\underline{x}\rangle$. If $c_{k\ell} = 0$ for $k = \ell$ and $= 1$ otherwise, we get back to the preceding case.

These considerations form the starting point of a topic called **decision theory**, for which Wald layed the basis back in 19??.

g **Basic Scheme.** All the rules mentioned above have the same structure: One determines "affinities" $a_k\langle\underline{x}\rangle$ and assigns $\underline{x}$ to the class with the highest affinity.

For practice, it is often useful to not only give the decision $\widehat{k}\langle\underline{x}\rangle = \arg\max_k\langle a_k\langle\underline{x}\rangle\rangle$ but also the $a_k\langle\underline{x}\rangle$ itself. With this we get a different picture:

- If similarly large affinities exist, $a_\ell\langle\underline{x}_i\rangle \approx a_{\hat{k}}\langle\underline{x}_i\rangle$, then the decision is uncertain.
- If even the maximal affinity is small, the observation does not fit in any class.

h* The $a_k\langle\underline{x}\rangle$ don't have to be interpretable as probabilities – in fact, in 5.4.f, they aren't. If we want to make this possible formally – even for a_k which can be negative – we can form $p_k\langle\underline{x}\rangle = c^{a_k\langle\underline{x}\rangle}/\sum_{k'} c^{a_{k'}\langle\underline{x}\rangle}$. For $c = e$ this rule is known as "**softmax**" . Note: If the rule is applied on probabilities, it changes these. The basis c plays a role.

5.5 * Further Methods in Discriminant Analysis

a The earlier discussion depends on the assumption of the multivariate normal distribution. For large data sets, we should be able to find procedures that react better to subtleties of the distribution. Here some of such procedures should be briefly mentioned.

b **Nearest Neighbors.** A procedure that in principle allows for arbitrarily shapeless "territories" of the individual classes is based on finding, for a new observation to be classified, the $\ell \geq 1$ nearest neighbors from the training data and determining their class membership. The membership of the new observation is then determined by

"majority vote" among these neighbors. As a refinement, we could weight the "votes" corresponding to their proximity to the new observation.

```
R> library(class) ; knn(...)  ; knn1(...)
```

c **Neural Networks** are suitable for modeling a general regression problem with "input"variables $X^{(j)}$ and one or more variables of interest or "output" variables $\underline{Y}^{(k)}$. They do this by creating a type of circuit, in which additional "switch nodes" – or, in statistical terms, latent variables – are introduced.

The most common variant is the "one hidden layer feed-forward neural network", which is schematically represented in Fig. 5.5.c. The model has the form

$$Y^{(k)} = g_k \left\langle \alpha_k + \textstyle\sum_\ell w_{\ell k}\, \widetilde{g}_\ell \left\langle \widetilde{\alpha}_\ell + \sum_j \widetilde{w}_{j\ell} X^{(j)} \right\rangle \right\rangle \ .$$

For g and $\widetilde{g}$ the logistic function is usually used.

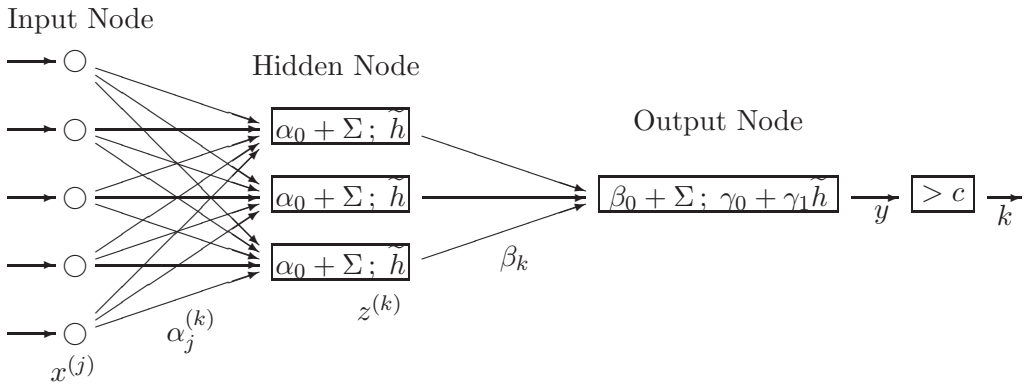


Figure 5.5.c: Plan of a neural network with a "hidden layer" of nodes.

For discriminant analysis the output $\underline{Y}$ still must be transformed into a classification. For 2 classes an individual Y can be compared with a threshold as with linear discriminant analysis. For more than two classes, the corresponding number of $Y^{(k)}$ can be introduced, and K can be defined as $K = \arg\max_k \langle Y^{(k)} \rangle$.

Neural networks are seen by engineers as universal tools for flexible modeling of any kind of input-output relationship. For statisticians there are three points of concern:

- The danger of overfitting to the data is large. We must therefore be careful that the number of the estimated parameters remains small in relationship to the number of observations which are used for their estimation.
- The model gives no direct concrete representation or interpretation; it remains a "black box".
- If new input data lies outside of the region of the input data of the training data set, the classification becomes completely unreliable.

```
R> library(nnet) ; nnet(...)
```

d **Classification and Regression Trees (CART).** The idea of a classification tree is simple and corresponds to the classical identification keys for plant species: The observations are divided into two groups on the basis of of an appropriate variable, such that the split separates the two given classes as well as possible. Then each group is split again with help of an appropriate variable and proceeds thus, until in each group, insofar as possible, only one class is present. Thus gnerates a "decision tree"}.

`R> library(tree) ; tree(...)` or `R> library(rpart) ; rpart(...)`

e **Boosting** is the name of a recipe that creates a better procedure from a (too) simple classification method through "recycling":

1. Estimate the rule with the simple method. This produces the classification $K^{(0)}\langle \underline{x}_i \rangle$
2. Determine the incorrectly classed observations. Estimate the rule again, with larger weights for these observations. This determines the classification $K^{(1)}\langle \underline{x}_i \rangle$

Repeat this step several times, until the incorrect classifications no longer change. Finally, define as the new rule the classification that results from a "weighted majority vote" among all the classifications generated along the way.

A disadvantage of this procedure is that the achieved classification rule is not directly interpretable, since it is no longer obvious which variables are effective for classification. That is also true for the following procedure, which is based on a similar idea.

Literature: Friedman, Hastie and Tibshirani (2000)

f **Bagging** is a second idea that improves a simple classification method. As the longer name "bootstrap aggregating" expresses, the simple rule is determined many times by means of bootstrap. The improved procedure is determined by majority voting.

L **Literature:** Ripley (1996) Treats all except the last two methods fairly in detail and with a focus on practical application. Sometimes not precise.

5.S S-Functions

a **Linear Discriminant Analysis.** A linear discriminant analysis is carried out with the function `lda` from the package `MASS`,

```
> library(MASS)
> t.r  <- lda(Species~.,data=iris)    or
> t.r  <- lda(x=iris[,1:4], grouping=iris[,5])
```

In the first case we provide a formula `groups ~ x1 + x2 + ...` , where `groups` is the grouping factor and x_i the continuous X variables. (The short formula `Species~.` is a shortened notation for the case that we want to use all variables except for the one to the left of the ~ sign.) The argument `data` indicates as usual the data frame form which these variables will be taken. In the second version, `x` is a data frame or a matrix which encompasses the X variables, and `grouping` equals the variable indicating the classes.

Variations are chosen by setting addtional arguments:

`prior`	prior probabilities π_i
`CV = T`	estimation of error rates by cross validation
`method`	robust estimation methods

The result, an `lda` object, lists the components

`$counts`	number of observations in each class
`$means`	class means
`$scaling`	β coefficients of the discriminant function(s); α is regrettably not given.
If `CV` is `TRUE`:	
`$class`	class membership according to cross validation
`$posterior`	posterior probabilities.

b **Graphical output.** `plot(t.r)` displays the values of the discriminant function(s), by a histogram in the case of two classes, a scatterplot for three classes, and a scatterplot matrix for more classes.

c The **identification**, that is the determination of the most plausible class for arbitrary (new) obervations, is obtained by

```
> predict(object=t.r, newdata)
```

Here, `object` is the result of `lda` (an `lda` object) and `newdata` equals the observations to be classified, the default being the training data used for generating `object`.

d **Logistic Regression.** A logistic regression is fitted by the function `glm` setting the argument `family=binomial`,

```
> t.r <-  glm(Species~., data=iris[51:150,], family=binomial)
```

`t.p  <- predict(t.r, newdata, type="response")` then yields the posterior probabilities, and the classification is obtained by classifying the observations with `t.p<0.5` into the second group.

e* **Further Methods.** With `qda` ("q" for quadratic), a quadratic discriminant analysis is produced. `predict` then chooses the corresponding method for identification.

The package `mda` contains the functions `mda` (mixed) and `fda` (flexible discriminant analysis).