In statistical pattern recognition, we recognize that features may be measured with error and that some of the features are useful for identification of the class while others are not. Our goals are then to obtain useful sets of features and to use these features in such a way that the identification is as accurate as possible—that is, that the number of classification errors (or probability of misclassification) are as small as possible. If there is an object that is to be classified on the basis of a feature vector $\mathbf{x}$, into one of g possible classes ($c_1, c_2, \ldots, c_g$), then the probability that the object is classified into class i when $\mathbf{x}$ is observed can be described by $P(c_i|\mathbf{x})$. This probability is obtained by Bayes' theorem, which we shall now derive. We start from the "theorem on compound probabilities" [Felle57]. This states that if there are random variables x and y, then

$$P(x\&y) = P(y|x)P(x) \tag{7.4}$$