

OVERRIDING THE EXPERTS: A FUSION METHOD FOR COMBINING MARGINAL CLASSIFIERS

MARK D. HAPPEL and PETER BOCK

Department of Computer Science

The George Washington University

Washington, DC 20052

mhappel@seas.gwu.edu & pbock@seas.gwu.edu

Received (12 July 2000)

Revised (27 October 2000)

The design of an optimal Bayesian classifier for multiple features is dependent on the estimation of multidimensional joint probability density functions and therefore requires a design sample size that increases exponentially with the number of dimensions. A method was developed that combines classification decisions from marginal density functions using an additional classifier. Unlike voting methods, this method can select a more appropriate class than the ones selected by the marginal classifiers, thus "overriding" their decisions. It is shown that this method always exhibits an asymptotic probability of error no worse than the probability of error of the best marginal classifier.

Keywords: Combination of Classifiers, Pattern Recognition, Curse of Dimensionality

1. Introduction

The use of multiple features by a Bayesian classifier often leads to a reduced probability of error. Unfortunately, the design of an optimal Bayesian classifier for multiple features requires that the class-conditional probability density functions be known. If the density functions are not known *a priori*, they must be estimated from a design sample. The estimation of multidimensional joint probability density functions is often nontrivial and requires a design sample size that, in general, increases exponentially with the number of dimensions.

This paper proposes a method for using multiple classifiers to obtain an acceptable probability of error from a design sample that is too small to permit an adequate estimate of the multivariate class-conditional density functions to be obtained. By using a *supervisory* Bayesian classifier to combine the classification decisions from a set of *marginal* Bayesian classifiers that each use only a subset of the features, an overall probability of error that is at least as good as that of the best of the marginal classifiers can be obtained.

The theoretical performance of such a multiple classifier system is examined in detail below.

1.1 Bayesian pattern classification

Given a set of objects and their corresponding pattern values, one of the fundamental problems of pattern classification is to determine a mapping that can assign an appropriate class label to each pattern in the pattern space. In a Bayesian classifier, the classification decision is made based on the *a posteriori* probabilities that the input is a member of a given class given the input. For a given input pattern X , the *a posteriori* probability for class ω_i , $p(\omega_i | X)$, can be calculated using Bayes' rule: ¹

$$p(\omega_i | X) = \frac{p(X | \omega_i) P(\omega_i)}{\sum_i p(X | \omega_i) P(\omega_i)} \quad (1)$$

The Bayesian decision rule selects the class label which corresponds to the maximum *a posteriori* probability. The class-conditional probability density function $p(X | \omega_i)$ is often referred to as the likelihood function. ¹ The numerator of the right-hand term of (1), consisting of the likelihood function weighted by the *a priori* probability $P(\omega_i)$, will be referred to here as the *weighted likelihood*. Since the sum of the weighted likelihoods (the denominator in the equation above) is positive and common to all of the *a posteriori* probabilities, it can be factored out and the comparison made of the individual weighted likelihoods instead, selecting the class ω_i for which:

$$p(X | \omega_i) P(\omega_i) > p(X | \omega_j) P(\omega_j) \quad \text{for all } j \neq i. \quad (2)$$

In Figure 1, the decision threshold has been identified by a dotted line passing through both the abscissa and the two joint density functions, and the resultant classification regions have been labeled with the boxed letters below the abscissa. Note from Figure 1 that some ω_2 objects have feature values that fall to the left of the decision threshold and would consequently be misclassified as belonging to class ω_1 . The probability of finding an ω_2 object with a feature value that falls in the ω_1 classification region can be found by determining the area under the ω_2 probability density function to the left of the decision threshold. Since all such objects would be erroneously classified, the corresponding area represents the probability of error due to ω_2 objects classified as ω_1 , shown in Figure 1 as ϵ_2 . Likewise, the probability of error due to ω_1 objects classified as ω_2 is shown in Figure 1 to the right of the decision threshold as ϵ_1 . The sum of the two areas represents the probability of error of the classifier. ²

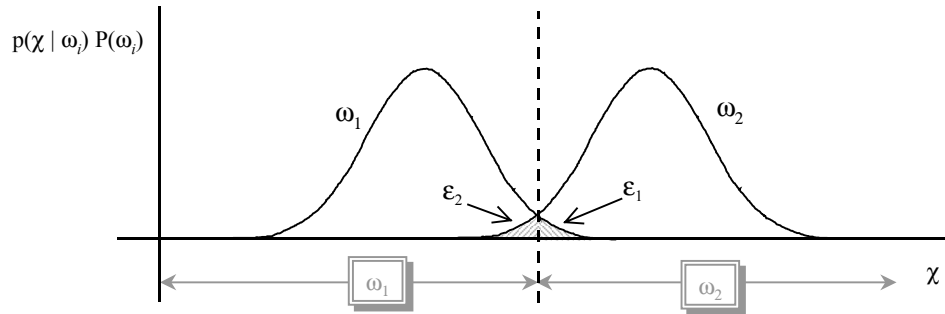
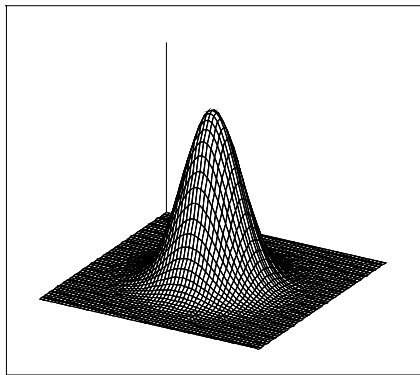


Fig. 1. Decision Threshold and Classification Regions for the Bayesian Classifier

The Bayesian classifier is optimal in the sense that it has the lowest possible probability of error for a given set of probability density functions. If the probability of error attained by a Bayesian classifier is unacceptably high for the requirements of a given problem, a different feature which exhibits better separation between the classes can be sought. Alternatively, two or more features can be used simultaneously to form multivariate joint probability density functions.

The joint probability density function formed from two or more features considered simultaneously can be used by a Bayesian classifier in the same manner as a univariate density function. An example surface plot of a bivariate joint normal distribution formed from two features is shown in Figure 2. When viewed from directly above, the contours of the density function in Figure 2 form concentric circles, as shown in Figure 3.



M

Fig. 2: Example Surface Plot of a Bivariate Normal Joint Density Function

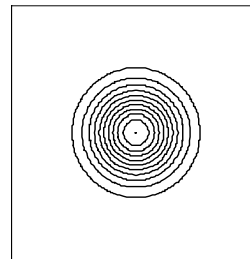


Fig. 3: Example Contour Plot of a Bivariate Normal Joint Density Function

The constituent univariate density functions of the individual features, referred to as *marginal* density functions, can be obtained from the bivariate joint density function by integrating the joint density function with respect to one or the other of the coordinate

axes. In the case of Figure 2, integration yields uncorrelated features with normal marginal density functions of equal variance. The joint density function of Figure 2 and its associated marginal densities are shown in Figure 4 below. Note that the volume under the joint density curve, by definition, is equal to one, as are the areas under the marginal density curves.

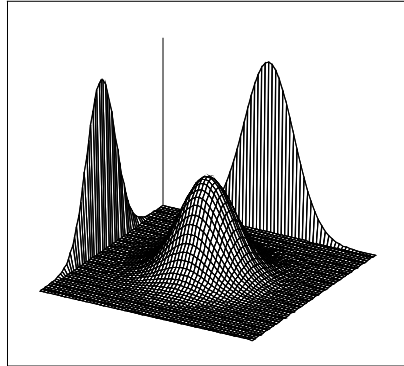


Figure 4: Bivariate Normal Density and Associated Marginal Densities

To better visualize the relationships between marginal and bivariate density functions, a modified contour plot such as the one shown in Figure 5 can be utilized. Here the marginal densities are shown along their respective axes, and a representative contour is plotted such that the corresponding points of equal density are aligned. Note that the contour line shown is only a single representative of the actual contour (see Figure 3); a bivariate normal density actually extends from $-\infty$ to ∞ along both axes.

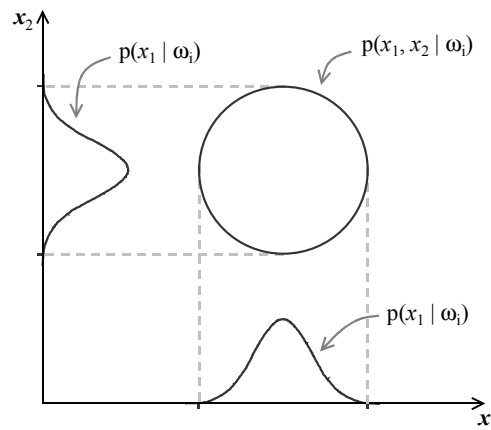


Fig. 5: Modified Contour Plot

The modified contour plot can be used to provide a visual means of comparing the probability of error using all of the available features to the probability of error attainable when using a subset of the available features. In order to make such a comparison, it is useful to extend the decision boundaries from one of the marginal axes into the joint space, thereby dividing the joint space into hyperrectangular *corridors*. The relationship between the marginal and multivariate probabilities of error within a given corridor is made explicit in Lemma 1 below.

Lemma 1. Suppose that a corridor $C\{q_1 \leq x_1 \leq r_1, \dots, q_n \leq x_n \leq r_n\}$ contains a portion of an m -dimensional multivariate complete probability density function $p(x_1, x_2, \dots, x_m)$ and the corresponding portion of an n -dimensional marginal density function $p(x_1, x_2, \dots, x_n)$, $n < m$. Then the hypervolume that lies under that portion of the marginal density function that is contained within the corridor will be equal to the hypervolume that lies under that portion of the complete density function that is contained within the corridor.

Proof. The hypervolume V_C under the complete density function $p(x_1, x_2, \dots, x_m)$ that falls within the given corridor can be found by evaluating the following integral:

$$V_C = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{q_n}^{r_n} \cdots \int_{q_1}^{r_1} p(x_1, \dots, x_n, x_{n+1}, \dots, x_m) dx_1 \cdots dx_n dx_{n+1} \cdots dx_m \quad (3)$$

The marginal density function $p(x_1, x_2, \dots, x_n)$, $n < m$, is defined by:

$$p(x_1, x_2, \dots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, x_2, \dots, x_m) dx_{n+1} dx_{n+2} \cdots dx_m ; \quad n < m \quad (4)$$

The hypervolume V_M under the marginal density function that falls within the given corridor can be found by evaluating the following integral:

$$V_M = \int_{q_n}^{r_n} \cdots \int_{q_1}^{r_1} p(x_1, \dots, x_n) dx_1 \cdots dx_n \quad (5)$$

Substituting (4) into (5) for $p(x_1, x_2, \dots, x_n)$ yields:

$$V_M = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{q_n}^{r_n} \cdots \int_{q_1}^{r_1} p(x_1, \dots, x_n, x_{n+1}, \dots, x_m) dx_1 \cdots dx_n dx_{n+1} \cdots dx_m = V_C \quad (6)$$

which is equal to V_C from (3). Therefore, $V_M = V_C$.

□

When the *a priori* probabilities $P(\omega_i)$ for each class are equal, then a classifier can base its decision solely on the magnitudes of the class-conditional probability density functions (*i.e.*, the likelihoods). However, when the *a priori* probabilities are not equal, it is necessary to compare the weighted likelihoods instead. Lemma 2 below extends the result of Lemma 1 to include weighted likelihoods as well as density functions.

Lemma 2. Suppose that a corridor $C\{q_1 \leq x_1 \leq r_1, \dots, q_n \leq x_n \leq r_n\}$ contains a portion of an m -dimensional multivariate complete weighted likelihood $p(x_1, x_2, \dots, x_m)P(\omega_i)$ and an n -dimensional marginal weighted likelihood $p(x_1, x_2, \dots, x_n)P(\omega_i)$, $n < m$. Then the hypervolume that lies under that portion of the marginal weighted likelihood that is contained within the corridor will be equal to the hypervolume that lies under that portion of the complete weighted likelihood that is contained within the corridor.

Proof. The hypervolume V_C under the complete weighted likelihood $p(x_1, x_2, \dots, x_m)P(\omega_i)$ that falls within the given corridor can be found by evaluating the following integral:

$$\begin{aligned} V_{CL} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{q_n}^{r_n} \cdots \int_{q_1}^{r_1} p(x_1, \dots, x_n, x_{n+1}, \dots, x_m) P(\omega_i) dx_1 \cdots dx_n dx_{n+1} \cdots dx_m \\ &= P(\omega_i) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{q_n}^{r_n} \cdots \int_{q_1}^{r_1} p(x_1, \dots, x_n, x_{n+1}, \dots, x_m) dx_1 \cdots dx_n dx_{n+1} \cdots dx_m \end{aligned} \quad (7)$$

which by (3) is equal to:

$$V_{CL} = P(\omega_i) V_C \quad (8)$$

Similarly, the marginal weighted likelihood hypervolume can be expressed as:

$$V_{ML} = P(\omega_i) V_M \quad (9)$$

Since $V_C = V_M$ by Lemma 1, it follows that:

$$V_{ML} = P(\omega_i) V_M = P(\omega_i) V_C = V_{CL} \quad (10)$$

Therefore, $V_{CL} = V_{ML}$.

□

1.2 The curse of dimensionality

Bayesian classifiers provide optimal classification results for a given pattern vector, provided that the class-conditional density functions and *a priori* probabilities for each class are known. Unfortunately, it is often the case that these density functions are not known in advance and must therefore be estimated from sample data. A wide variety of parametric and nonparametric techniques for estimating density functions exist and can be used for the design of Bayesian classifiers.^{2,3,4} The simplest nonparametric technique for density estimation is the histogram, which records the relative frequency with which samples fall within a given range (bin width), thereby providing an estimate of the average magnitude of the probability density function for points falling within the bin.

There are several problems which tend to complicate the use of histograms for the estimation of multivariate density functions, including a dramatic increase in the number of required histogram bins that can occur as the number of dimensions is increased. Since the required number of bins rises exponentially with an increasing number of dimensions, it should be apparent that the resources required to store and analyze such a histogram may quickly exceed what is practical.³ This and other related problems collectively contribute to what has become known as "the curse of dimensionality".⁵ The curse of dimensionality leads to an interesting paradox: for situations in which the optimal Bayesian classifier performance is insufficient for d dimensions, it may not be possible in practice to attain better classification performance using $d+1$ dimensions, even though the theoretical Bayesian performance should increase. From the preceding discussion, it is apparent that a method for obtaining an improvement in the classification performance for the d -dimensional Bayesian classifier without requiring the estimation of $d+1$ dimensional density functions would prove useful.

A promising line of research is based upon the creation of a single "group" decision from the decisions of multiple classifiers.⁶ The performance of a combination of classifiers has been found to be superior to that of a single classifier in many situations. It is intuitively appealing to imagine combining several, lower-dimensional Bayesian classifiers in such a way as to provide a lower error rate than any one of them alone can achieve, and perhaps even to approach the error rate attainable with a higher-dimensional classifier. Current strategies for obtaining group decisions include dynamic classifier selection,⁷ voting or weighted voting,⁸ Bayesian techniques,^{9,10} Dempster-Shafer evidence theory,¹¹ and stacking methods.⁶ Other strategies include the Behavior-Knowledge Space (BKS) approach, which combines handwriting recognition classifiers based on different classification algorithms to obtain improved performance.¹²

None of the classifier combination methods reported above appear to be suitable for combining the results of classifiers utilizing marginal density functions into "pseudo-complete" classification decisions. Voting methods are severely constrained by the classification performance of the member classifiers, which suffers when feature subsets are used. Bayesian and Dempster-Shafer methods assume conditional independence between the features. The next section describes a method capable of combining marginal classifiers that is not constrained by these limitations.

2. Solution Method

When two features x_1 and x_2 are available, they can be used simultaneously by a bivariate classifier, whose block diagram is shown in Figure 6. By using two features, the bivariate classifier is often able to achieve a significantly better classification performance than a comparable univariate classifier that is relying on either of the same two features. Unfortunately, the bivariate classifier will also require a larger training sample size than the univariate classifier to accurately estimate the class-conditional probability density functions.

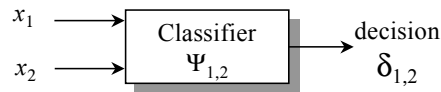


Fig. 6: Bivariate Bayesian Classifier

The method proposed here is to use the marginal decisions as features, thereby forming a new pattern vector. An additional classifier, called a supervisory classifier, can then be used to classify the pattern of subset decisions and generate a new classification decision. A block diagram of this architecture is shown in Figure 7. Note that this is essentially a stacking architecture, formed in this case from multiple marginal classifiers and a single supervisory classifier. The intention is to allow all of the classifiers to be implemented from a single common Bayesian building block. A version of this architecture has been previously implemented and shown empirically to improve recognition of printed characters.¹³

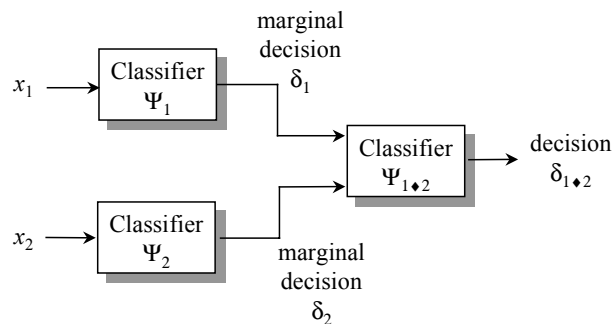


Fig. 7: Stacking Architecture

In order to clarify the operation of this method, a detailed example is presented in the next section.

2.1 An example

The example presented here involves three classes, labeled "M", "B", and "G". Two features, x_1 and x_2 , are used to classify points in the feature space. The joint probability density functions $p(x_1, x_2 | M)$, $p(x_1, x_2 | B)$, and $p(x_1, x_2 | G)$ are all normally-distributed, and the *a priori* probabilities $P(M)$, $P(B)$, and $P(G)$ are equal. The marginal density functions $p(x_1 | M)$ and $p(x_2 | M)$ are normal, of equal variance, and uncorrelated, as are $p(x_1 | B)$, $p(x_2 | B)$, $p(x_1 | G)$ and $p(x_2 | G)$. The density functions are shown in a modified contour plot in Figure 8.

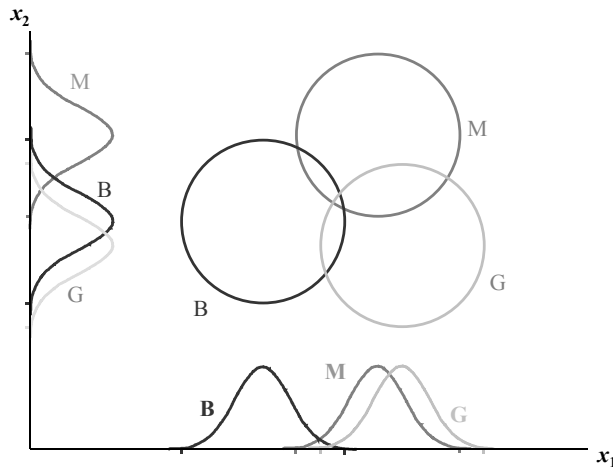


Fig. 8: Probability Density Functions for Two-Class Example

Using the architecture of Figure 7, feature x_1 is classified by marginal classifier Ψ_1 , while feature x_2 is classified by marginal classifier Ψ_2 . The marginal classifiers make a Bayesian decision based only on the marginal density functions for the single feature that they each receive. The decision surface of each classifier is shown in Figure 9, and the (one-dimensional) decision regions are shown along each feature axis.

The extension of the Bayesian decision surfaces from a given marginal classifier Ψ_m divides the joint space into hyperrectangles called *corridors*. Consequently, each corridor represents that portion of the joint space that corresponds to marginal decision δ_k made by classifier Ψ_m . The intersection of corridors from multiple classifiers further divides the space into (hyper)rectangular subdivisions called *partitions*. By convention, a partition θ is labeled in the form $\delta_1\delta_2$ according to the corresponding marginal decisions δ_1 and δ_2 (i.e., the leftmost symbol will correspond to the marginal decision made from feature x_1). A partition represents a portion of pattern space whose associated patterns share common marginal decisions.

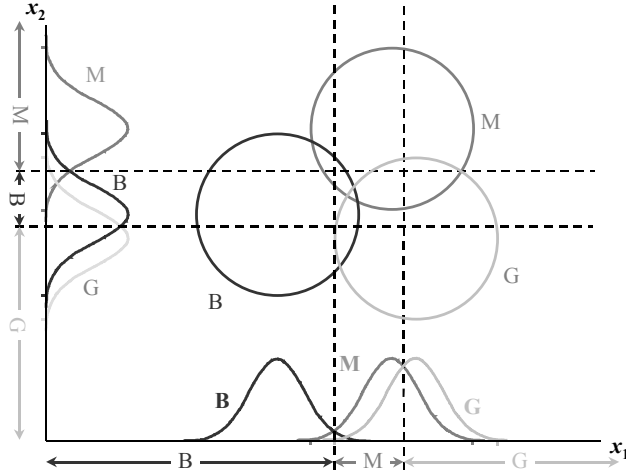


Fig. 9: Partitioning of the Joint Space

The marginal classifiers (Ψ_1 and Ψ_2 in Figure 7) receive features in pattern space Π and transform them into marginal decisions in partition space ρ on the basis of the class-conditional pattern probability density functions $p(x_j | \omega_i)$. The supervisory classifier ($\Psi_{1,2}$ in Figure 7) receives the marginal decisions in partition space ρ and transforms them to classification decisions in decision space Δ . To accomplish this, the supervisory classifier requires the class-conditional partition probabilities $P(\theta_k | \omega_i)$. The supervisory classifier can use η bins to construct its estimate of $P(\theta_k | \omega_i)$ for each class. If η is appreciably less than the number of bins that would have been required to estimate the joint probability density functions $p(x_1, x_2 | \omega_i)$, then the supervisory classifier will probably require a smaller design sample size than would have been required by a corresponding bivariate classifier using $p(x_1, x_2 | \omega_i)$.

The supervisory classifier will select the class ω_i within partition θ_k for which $P(\theta_k | \omega_i)$ is the largest. $P(\theta_k | \omega_i)$ is equal to the volume of the class-conditional joint probability density function $p(x_1, x_2 | \omega_i)$ that is contained within the boundaries of partition θ_k . Consequently, the class selected by the supervisory classifier will be that class whose probability density function $p(x_1, x_2 | \omega_i)$ covers the most volume within the partition in question.

This can be clearly seen for the current example in Figure 10. Focusing on the upper right partition (GM), it is readily apparent that the area under the class M bivariate density function that falls within partition MM is much larger than the area under the class B or G bivariate density functions that also falls within partition MM. (It is important to remember that the curves M and B of Figure 10 are merely representative contours of the bivariate density functions, but the contours are useful for comparing the respective volumes.) Consequently, a pattern that falls within the partition is more likely to correspond to an object from class M than one from classes B or G.

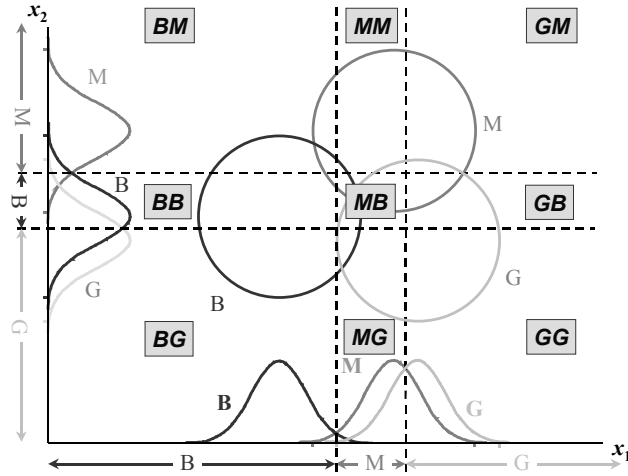


Fig. 10: Partition Labels Based on Marginal Classifications

Special attention should be paid to the center partition (MB), which is shown in more detail in Figure 11. It is readily apparent that the G class has the most volume under its bivariate density function and therefore would be chosen by the supervisory classifier. However, the marginal classifiers chose classes M and B respectively. Thus, in this case the supervisory classifier has overridden the advice of the marginal classifiers with the result that a lower probability of error is obtained. This *override* behavior is not exhibited by more common classifier combination techniques, such as majority voting, and is a key element in the improved performance from this architecture.

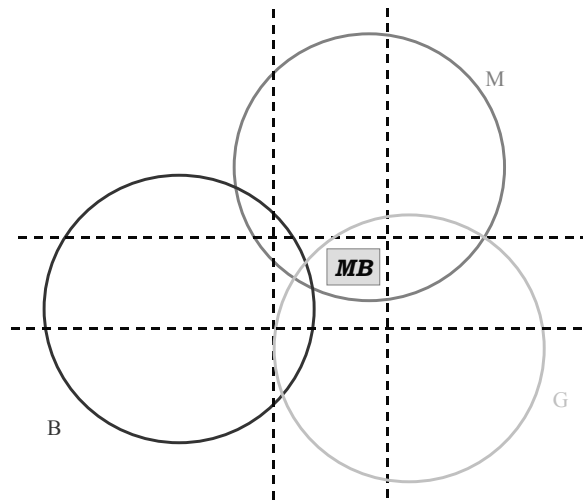


Fig. 11: Decision Override in Partition MB

Similarly, the decision that would be made by the supervisory classifier for the other partitions can be predicted from Figure 10 and have been labeled in Figure 12 within the appropriate partitions. Note that the resultant decision surfaces, as shown in Figure 13, are not identical to the decision surfaces from either of the marginal classifiers, but have instead been formed from a combination of them. Figure 13 also shows the resultant probability of error, which is less than that of either marginal classifier, as crosshatched volumes under the respective density functions.

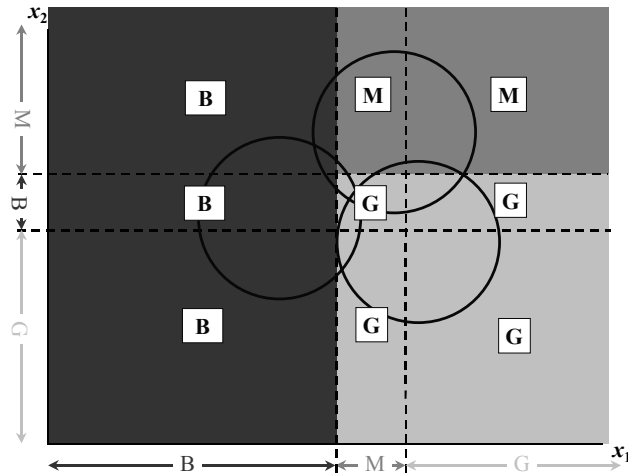


Fig. 12: Partition Classifications

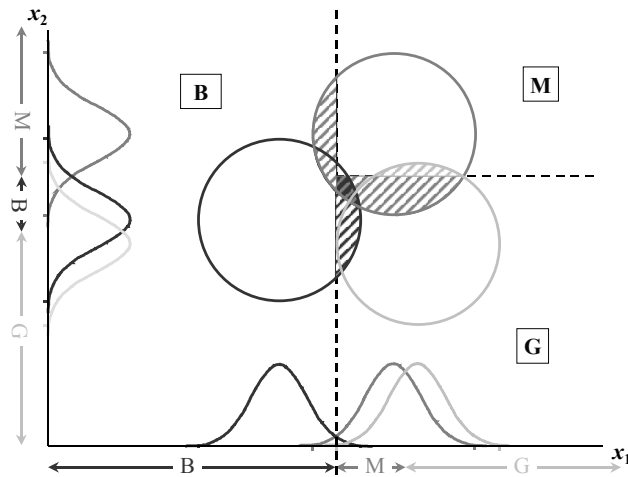


Fig. 13: Resultant Probability of Error

3. Theoretical Performance

The performance of a Bayesian classifier is dependent on the accuracy of the class-conditional probability density function estimates on which it is based. As the size of the available design sample increases, the accuracy of the density function estimates increases and the probability of error consequently decreases toward the *asymptotic* probability of error. Note that the asymptotic probability of error for the ideal Bayes classifier is the Bayes error ϵ_β .

It is possible to determine upper and lower bounds on the asymptotic performance of the proposed method without making assumptions as to the form or parameters of the likelihoods. In the asymptotic case, it has been noted previously that the Bayes error ϵ_β is optimal, and therefore forms the lower bound on the proposed method's error. The upper bound can be determined with respect to the probabilities of error of the associated marginal classifiers.

3.1 Proof

Let ϵ_α denote the lowest probability of error attained by any of the m associated marginal classifiers:

$$\epsilon_\alpha = \min (\epsilon_1, \epsilon_2, \dots, \epsilon_m) \tag{11}$$

It is shown below in Theorem 1 that ϵ_α is the upper bound for the asymptotic probability of error of the proposed method.

Theorem 1. Suppose that d features x_i , $1 \leq i \leq d$, are used to discriminate between patterns belonging to N classes ω_j , $1 \leq j \leq N$, for which the class-conditional probability density functions are known. Then $\epsilon_\beta \leq \epsilon_H \leq \epsilon_\alpha$.

Proof: The volume which lies under the joint weighted likelihood for class ω in partition θ can be found by performing a multiple integration, over all d dimensions, of the joint weighted likelihood within the partition boundaries:

$$V_{\omega\theta} = \int_{\theta} \dots \int p(x_1, x_2, \dots, x_d) dx_1 dx_2 \dots dx_d \tag{12}$$

This volume is significant in that it is proportional to the probability that a pattern drawn from class ω will fall within the chosen partition. If the classifier has chosen a different class for partition θ (i.e., $\delta_\theta \neq \omega$), then the volume $V_{\omega\theta}$ represents the probability that a pattern from class ω will be incorrectly classified (i.e., the contribution of class ω to the probability of error within partition θ).

The portion of the classifier's probability of error H that lies within partition θ , H_θ , is equal to the sum of the volumes under all of the weighted likelihoods that correspond to the classes not selected by the classifier:

$$H_\theta = \sum_{\omega \neq \delta_\theta} V_{\omega\theta} \quad (13)$$

The portion of the classifier's probability of error that lies within corridor K is equal to the sum of the probabilities of error for all of the partitions within the corridor:

$$H_K = \sum_{\theta \in K} H_\theta = \sum_{\theta \in K} \sum_{\omega \neq \delta_\theta} V_{\omega\theta} \quad (14)$$

The total probability of error for the classifier H will be equal to the sum of the probabilities of error of all of the corridors:

$$H = \sum_K \sum_{\theta \in K} \sum_{\omega \neq \delta_\theta} V_{\omega\theta} \quad (15)$$

A marginal density function is obtained by integrating the joint density function over some subset of the available dimensions, yielding a lower-dimensional probability density function:

$$p(x_1, x_2, \dots, x_f) = \int \dots \int p(x_1, x_2, \dots, x_d) dx_{f+1} dx_{f+2} \dots dx_d ; \quad f < d \quad (16)$$

The volume (or area, in the case of a univariate marginal density function) which lies under the marginal weighted likelihood for class ω in corridor K can be found by performing a multiple integration, over f dimensions, of the marginal weighted likelihood within the corridor boundaries:

$$A_{\omega K} = \int \dots \int_K p(x_1, x_2, \dots, x_f) dx_1 dx_2 \dots dx_f ; \quad f < d \quad (17)$$

The portion of the marginal classifier's probability of error M that lies within corridor K , M_K , is equal to the sum of the volumes (or areas) under all of the weighted likelihoods that correspond to the classes not selected by the marginal classifier:

$$M_K = \sum_{\omega \neq \delta_K} A_{\omega K} \quad (18)$$

Lemmas 1 and 2 show that the hypervolume that lies under that portion of the marginal density function that is contained within the corridor is equal to the hypervolume that lies under that portion of the complete density function that is contained within the corridor. This implies that

$$A_{\omega K} = V_{\omega K} \quad (19)$$

(18) and (19) provide a relationship between the marginal probability of error M_K and the class volumes within corridor K :

$$M_K = \sum_{\omega \neq \delta_K} A_{\omega K} = \sum_{\omega \neq \delta_K} V_{\omega K} \quad (20)$$

This result can be extended to the partition volumes:

$$M_K = \sum_{\omega \neq \delta_K} A_{\omega K} = \sum_{\omega \neq \delta_K} V_{\omega K} = \sum_{\theta \in K} \sum_{\omega \neq \delta_K} V_{\omega \theta} \quad (21)$$

Note that the volumes $V_{\omega \theta}$ that constitute the joint volume equivalent to the marginal probability of error M_K correspond to all of the classes other than the one selected by the marginal classifier. This is equivalent to:

$$M_K = \sum_{\theta \in K} \sum_{\omega \neq \delta_K} V_{\omega \theta} = \sum_{\theta \in K} \left[\left(\sum_{\omega} V_{\omega \theta} \right) - V_{\delta_K \theta} \right] \quad (22)$$

In other words, the marginal probability of error is equal to the sum of all of the class volumes within the corridor minus the selected class volume within the corridor.

Similarly, (14) can be rewritten as:

$$H_K = \sum_{\theta \in K} \sum_{\omega \neq \delta_\theta} V_{\omega \theta} = \sum_{\theta \in K} \left[\left(\sum_{\omega} V_{\omega \theta} \right) - V_{\delta_\theta \theta} \right] \quad (23)$$

It can be seen that both the marginal probability of error M_K and the proposed method's probability of error H_K are made up of a sum of volumes, and that within each partition these volumes differ by the volume associated with the marginal classification decision and the volume associated with the proposed method's classification decision. Note that the proposed method makes its classification decision within a given partition by selecting the class with the largest class volume. Therefore:

$$V_{\delta_0\theta} \geq V_{\omega\theta} ; \omega \neq \delta_0 \quad (24)$$

Consequently,

$$V_{\delta_0\theta} \geq V_{\delta_\kappa\theta} \quad (25)$$

The Bayes error is optimal, in that it always yields the lowest achievable probability of error. (25) shows that, within a given partition, the proposed method's error ϵ_H is always smaller than any marginal classifier's error ϵ_m (and consequently, the smallest marginal error ϵ_α). Since this is true in all the partitions of the joint space,

$$\epsilon_\beta \leq \epsilon_H \leq \epsilon_\alpha \quad (26)$$

□

3.2 Effects of class separation

Theorem 1 states that the proposed method's probability of error ϵ_H is constrained to a value that falls between the optimal Bayesian error ϵ_β and the smallest marginal error ϵ_μ . However, the theorem does not provide guidance as to which classification problems will yield a probability of error ϵ_H that approaches the Bayesian probability of error ϵ_β and which will yield much higher ones. An examination of the behavior of the classifier for a family of related classification situations can provide some additional insight.

As discussed in Section 1.1, the probability of error in a Bayesian classification problem arises from class-conditional probability density functions which overlap each other. One interesting experiment is to vary the separation between the classes' density functions by changing the mean value of one class' density function in such a way as to cause that density function to gradually pass through the other class' density function. (Of course, this is not a realistic classification problem, in which the density functions are a function of the features selected and whose means occur at fixed values. This is done only so that the effect of separation of the means can be examined independently of other possible variations, such as differences between the variances of the densities.) The intent is to examine the manner in which ϵ_β , ϵ_α , and ϵ_H vary as the separation between the densities is varied.

A modified contour plot for an example problem is shown in Figure 14. The two features, x and y , are independent in the case of the "B" class but are dependent in the case of the "M" class. Note that the magnitudes of the bivariate densities is shown within the corresponding contours. For this particular example, the mean of the y marginal density function for class B will be held constant, while the mean of the x marginal density function will be gradually increased. This will have the effect of shifting the class B bivariate density function horizontally through the M bivariate density function.

As the class B bivariate density function passes through the class M bivariate density function, it is possible to calculate the probabilities of error ϵ_β , ϵ_α , and ϵ_H for each position of the class B density's mean. A plot of the resultant values is shown in Figure 15.

Note that the probability of error ϵ_y of the y marginal classifier is constant; since the means are being held constant along the y axis, the overlap of the densities does not vary and the probability of error does not change. Conversely, the probability of error ϵ_x of the x marginal classifier increases from zero, when the means are sufficiently separated that there is no overlap of the density functions, to a maximum value when the two densities are completely overlapped (the fact that the value of ϵ_x is equal to that of ϵ_y at the point of maximum overlap is coincidental).

Like ϵ_x , the optimal Bayesian probability of error ϵ_β increases as the separation between the means of the density functions for the two classes decreases. The asymmetry on either side of the zero separation point is due to the asymmetry of the class M bivariate density function. As noted previously, ϵ_β is the minimum probability of error that can be achieved by any classification algorithm for the given features.

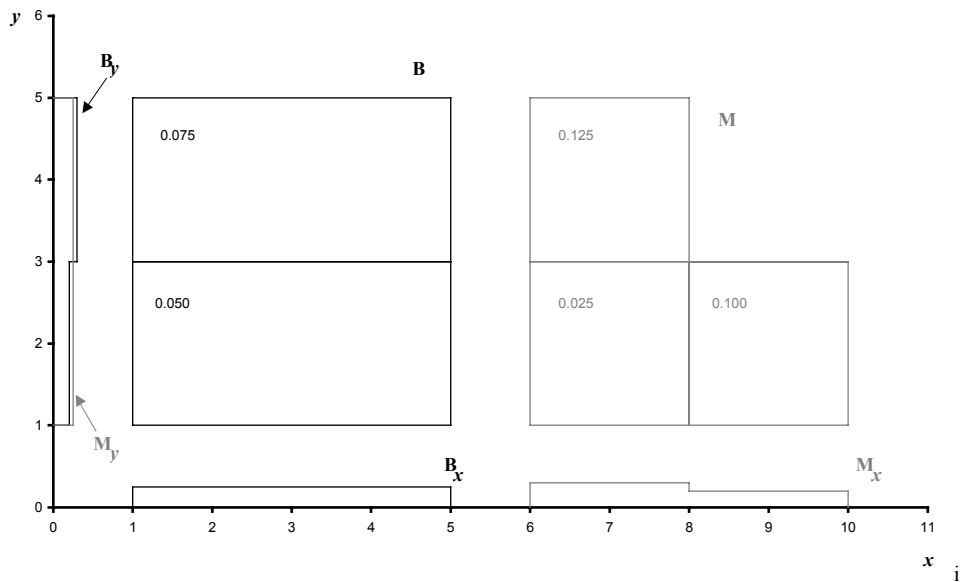


Fig. 14: Sample Classification Problem

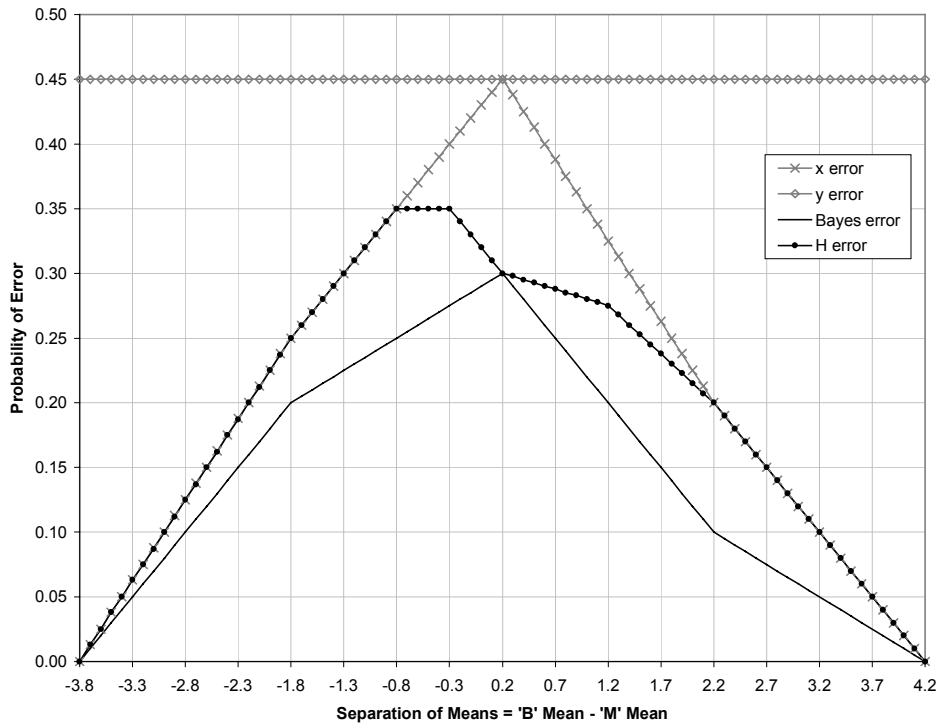


Fig. 15: Probability of Error as a Function of the Separation of the Means

The proposed method's probability of error ϵ_H is equal to that of the x marginal classifier when the separation between the bivariate means is relatively large. However, as the separation is reduced, and the x marginal error correspondingly increases, a point is reached at which ϵ_x increases faster than ϵ_H , resulting in a lower probability of error ϵ_H than that of either of the two marginal classifiers. As the separation continues to be reduced, ϵ_H increases more slowly or decreases until, at the point at which the means of the two bivariate density functions are equal, ϵ_H is equal to the optimal value ϵ_β (the fact that the value of ϵ_H is equal to that of ϵ_β at the point of maximum overlap is also coincidental).

Examination of the partition classifications at each point in Figure 15 helps to explain the behavior of ϵ_H as the separation between the means changes. Figure 15 can be divided into five regions of interest, based on the particular arrangement of partition classifications that occurs within a given region. The regions are shown in Figure 16.

The first region to examine is the region where the separation is less than -0.8 (labeled Region 1 in Figure 16). Throughout this region, ϵ_H is equal to ϵ_x . A careful examination of the partition classifications reveals the arrangement of partition classifications shown in Figure 17.

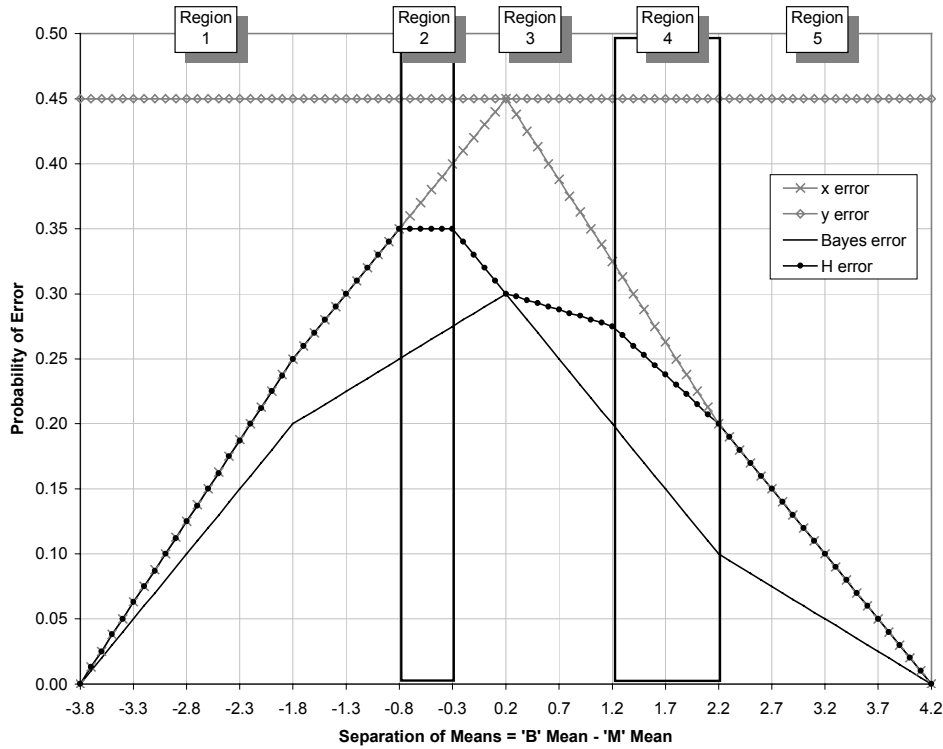


Fig. 16: Partition Classification Arrangements for Sample Classification Problem

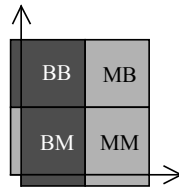


Fig. 17: Partition Classification Arrangement Occurring in Region 1

Likewise, an examination of Region 5, where the separation is greater than 2.2, reveals a similar arrangement, which is shown in Figure 18. Note that, although the two partitions classified as M are to the right in Figure 17 and to the left in Figure 18, the x marginal classification regions are also reversed. Consequently, the partitions with the same partition label (e.g., BM) in Figures 17 and 18 are also classified as having the same class label (e.g., BM is classified as B), so the arrangements are equivalent. It can be shown¹⁴ that such an arrangement will always result in a probability of error such that ϵ_H is equal to ϵ_x , which agrees with the observed probability of error in Region 1 and Region 5.

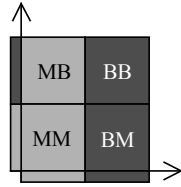


Fig. 18: Partition Classification Arrangement Occurring in Region 5

The partition classification arrangement within Region 2 of Figure 16 is somewhat more complex. When the separation between the means of the density functions falls within this region, the x marginal density functions are arranged such that several decision points occur along the x axis, forming eight partitions as shown in Figure 19 (the white partition is one throughout which both density functions are equal to zero). However, it should be noted that each partition label is shared by two partitions; with two classes and two features, only four unique partitions are possible. By the definition of the proposed method, the volumes under the density functions corresponding to a given class which lie in multiple partitions with the same label are added, and the Bayesian decision rule applied to the sum from each class. When this is done for separations within Region 2, Figure 19 is reduced to the arrangement shown in Figure 20 (the order of the x marginal classifications is arbitrary).

Similarly, the partition classification arrangement that occurs in Region 4 is also equivalent to the arrangement shown in Figure 20. It can be shown¹⁴ that, for such arrangements, $\epsilon_H \leq \epsilon_x$, which concurs with the plots of ϵ_H and ϵ_x in Figure 16.



Fig. 19: Partition Classification Arrangement Occurring in Region 2

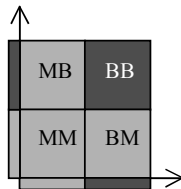


Fig. 20: Equivalent Partition Classification Arrangement for Region 2

In Region 3, the separation between the means of the density functions is minimal, and the probabilities of error are correspondingly large. Notice, however, that it is within this region that the proposed method most closely approaches the optimal error ϵ_β . Examination of the partition classifications reveals the arrangement shown in Figure 21. Significantly, within the partition (MM) in which both marginal classifiers have chosen class M, the proposed method has chosen class B, overriding the unanimous decision of the marginal classifiers. This decision override leads to a significantly reduced probability of error, as can be seen in Figure 16. Note that a voting classifier would have followed the unanimous choice of the marginal classifiers and chosen class M, resulting in a higher probability of error.

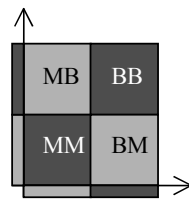


Fig. 21: Partition Classification Arrangement Occurring in Region 3

Finally, it should be noted that, in this case, the Naive Bayesian classifier (a Bayesian multivariate classifier that assumes conditional independence between the features) never achieves a probability of error lower than that of the x marginal classifier. The key to understanding this behavior lies in recognizing that the Naive Bayesian classifier is not using the class-conditional bivariate density functions shown in Figure 14, but rather, as a result of its assumption of conditional independence, is constructing a different set of bivariate density functions by multiplying the values of the respective marginal density functions. This results in the bivariate functions shown in Figure 22. While the class B bivariate density function is the same (which is to be expected, since the features were independent for class B), the class M bivariate density function is considerably different.

4. Conclusions

The design of an optimal Bayesian classifier for multiple features is dependent on the estimation of multidimensional joint probability density functions and therefore requires a design sample size that increases exponentially with the number of dimensions. It has been shown above that it is possible to combine the results of two marginal classifiers and obtain a probability of error that is less than that of either of the marginal classifiers. Further, it has been shown that this method always demonstrates an asymptotic probability of error that is greater than or equal to the probability of error of the optimal multivariate Bayesian classifier and less than or equal to the probability of error of the marginal classifier with the lowest probability of error. Current efforts are directed toward

determining the method's performance for design samples much smaller than those needed to obtain the asymptotic performance reported here.

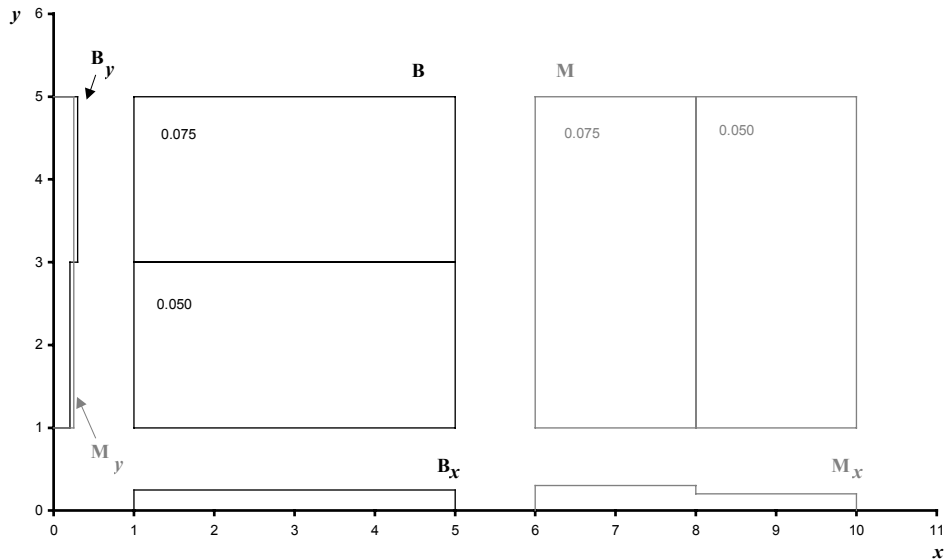


Figure 5-20: Bivariate Density Functions Used By Naive Bayesian Classifier

References

- [1] Duda, R., and P. Hart. *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons, Inc. (1973).
- [2] Fukunaga, K. *Introduction to Statistical Pattern Recognition* (2nd ed.). Boston: Academic Press, Inc. (1990).
- [3] Bishop, C. *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press. (1995).
- [4] Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons. (1992).
- [5] Bellman, R. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press. (1961).
- [6] Dietterich, T. *Machine Learning Research: Four Current Directions*. *AI Magazine*, (Winter 1997) 97-136.
- [7] Woods, K., W. Kegelmeyer Jr., and K. Bowyer 1997. *Combination of Multiple Classifiers using Local Accuracy Estimates*. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19/4 (1997) 405-410.

- [8] Lam, L., and C. Suen. *Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance*. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 27/5 (1997) 553-568.
- [9] Bloch, I. *Information Combination Operators for Data Fusion: A Comparative Review With Classification*. IEEE Transactions on Systems, Man and Cybernetics, Part A, 26/1 (1996) 52-67.
- [10] Xu, L., A. Krzyzak, and C. Suen. *Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition*. IEEE Transactions on Systems, Man, and Cybernetics 22/3 (1992) 418-435.
- [11] Buede, D., and P. Girardi. *A Target Identification Comparison of Bayesian and Dempster-Shafer Multisensor Fusion*. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 27/5 (1997) 569-577.
- [12] Huang, Y., and C. Suen. *A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals*. IEEE Transactions on Pattern Analysis and Machine Intelligence 17/1 (1995) 90-94.
- [13] Happel, M., and P. Bock. *The Classification of Symbolic Concepts Using the ALISA Concept Module* in Proceedings of the Ninth International Symposium on Artificial Intelligence (ISAI-96), Cancun, Mexico (1996).
- [14] Happel, M., and P. Bock. *Analysis of a Fusion Method for Combining Marginal Classifiers*. Multiple Classifier Systems, eds. J. Kittler and F. Roli, Springer-Verlag, Berlin (2000).