January 20, 2011

Multivariate Analysis Notes

Adrian Bevan*,

These notes have been developed as ancillary material used for both BABAR analysis school lectures, and as part of an undergraduate course in Statistical Data Analysis techniques. They provide a basic introduction to the topic of multivariate analysis.

The content is broken down in to discussions on methods of classifying data in terms of increasing complexity, from a simple cut-based approach, through to the use of decision trees. There is also a section devoted to the topic of how to choose the classification method one should ultimately use.

*a.j.bevanqmul.ac.uk

Chapter 1

Multi Variate Analysis

Consider a data sample Ω described by the set of variables $\underline{x}$ that is composed of two (or more) populations. Often we are faced with the task of trying to identify or separate one sub-sample from the other (as these are different classes or types of events). In practice it is often not possible to completely separate samples of one class A from another class B. There are a number of techniques that can be used in order to try and optimally identify or separate a sub-sample of data from the whole, and some of these are described below in order of increasing complexity. Each of the techniques described has its own benefits and dis-advantages, and the final choice of the optimal solution of how to separate A and B can require subjective input from the analyst. In general this type of situation requires the use of Multi Variate Analysis (MVA).

The simplest approach is that of cutting on the data to improve the purity of a class of events is described in Section 1.1. More advanced classifiers such as Fisher discriminants, neural networks and decision trees are subsequently discussed. The Fisher discriminant described in Section 1.3 has the advantage that the coefficients required to optimally separate two populations of events are determined analytically up to an arbitrary scale factor. The Neural Network (Section 1.4) and Decision Tree (Section 1.5) algorithms described here require a numerical optimisation to be performed. In this context the optimisation process is called training, and having trained an algorithm with a set of data one has to validate the solution. Having discussed several classifier algorithms the concepts of Bagging and Boosting are described as variants on the training process. In the following it is assumed that the data Ω contain only two populations of events. These populations are either referred to as A and B, or as signal and background depending on the context. It is possible to generalise to an arbitrary number of populations.

1.1 Cutting on variables

An intuitive way to try and separate two classes of events is to ***cut*** on the data to select interesting events in a cleaner environment than exists in the whole data set. For example if you consider the case where the data set Ω contains two types of events A and B that are partially overlapping. One possible solution to the problem of separating A from B is to select the events that satisfy $A \setminus B$. If $C = A \setminus B \neq \emptyset$, then this will be a pure sample of interesting events. The pertinent questions are (i) what sacrifice has been made in order to obtain C, and (ii) would it have been possible to reject less data and obtain a more optimal separation of A and B so that we can further study a subset of the data?

What do we mean by making a cut on the data? Consider the data sample above Ω which contains two classes of events: A and B, each of which is described by discriminating variables in n dimensions. If we cut on the value of one or more of the dimensions, then we decide to retain an event e_i that pass some threshold

$$P(e_i \in A) > 0 \tag{1.1.1}$$

and would decide to discard an event if

$$P(e_i \notin A) \text{ is significant.} \tag{1.1.2}$$

There is an element of subjectivity in the second condition. We can think of a cut in some variable x as a binary step function $f(x)$ where in the case of a positive step we may write

$$f(x) = 1 \text{ for } x > X_0, \tag{1.1.3}$$
$$f(x) = 0 \text{ elsewhere.} \tag{1.1.4}$$

and for a negative step, we change the inequality from $>$ to $<$. In order to optimize the cut on x we need to determine what it is we aim to achieve. If we assume our signal events are those of class A, then it follows that we would like to retain as many events of type A as possible, while discarding as many events of type B as possible. If $A \cap B = \emptyset$, then it is possible to determine X_0 by inspection of the distributions. If however $A \cap B \neq \emptyset$, we need to choose what we mean by the term *optimally separating A and B*.

The following lists some of the possible ways to determine X_0 for *optimal* separation.

1. If it is of paramount importance to obtain a pure sample of A with no contamination or dilution from B, then we define X_0 in such a way that satisfies $C = A \setminus B \neq \emptyset$ with as many events passing into C as possible. Practically this usually is achieved at a significant cost in statistics and so will probably not be a sensible criteria for optimisation for many of the situations encountered.

2. We can introduce the notion of the ***significance*** S of the signal content (amount of A) in Ω relative to the background (amount of B). Then we can choose the value of X_0 that results in the greatest significance of signal. A common definition of significance for this scenario is that of

$$S = \frac{N_S}{\sqrt{N_S + N_B}}, \tag{1.1.5}$$

   where N_S is the number of signal events, and N_B is the number of background events that pass a given cut with cut-value $x = X_0$. The motivation for this definition of significance is that we want to compare any hint of a signal found to the statistical uncertainty on the number of events in the data. The underlying logic is that we want to be able to minimise any incorrect claim of a signal that would arise from statistical fluctuations in the background sample. As a result if we compute a numerical value for S, we normally say that the expected significance for a given cut is $S\sigma$, assuming that the denominator corresponds to a Gaussian uncertainty on the total number of observed events.

3. If we are searching for an effect that is expected to be absent from the data, then we may want to optimize in such a way that we minimise the uncertainty on the background estimation (or number of events of type B that will remain in the sample), as this will dominate the uncertainty we obtain on the possible presence of a signal, and hence on any limit we are able to place that rules out the effect we seek.

The previous discussion with regard to making cuts has been based on a single dimension. In the case that all relevant dimensions $\underline{x}$ in Ω are uncorrelated, it is sufficient (and efficient) to optimize the cut values $\underline{X}_0$ one dimension at a time. The values of $\underline{X}_0$ obtained through such a procedure would be optimal. The more general situation encountered is when two or more dimension are correlated. For such cases one would ideally like to simultaneously optimize the values of $\underline{X}_0$, however in practice this is often not practical in terms of time or resources[1]. A possible alternative to this is to iteratively optimize the values of $\underline{X}_0$ one dimension at a time. If on subsequent iterations of the optimisation the value of X_0 obtained for a given dimension does

[1]The number of iterations required to simultaneously optimise m dimensions scales as the number of iterations for one dimension raised to the power of m. This is referred to as the curse of dimensionality as originally noted by Bellman [1961].

not change appreciably, then you will have obtained the optimal cut value for this dimension. In practice it may take several iterations to achieve this when two or more dimensions are correlated.

Example: Given a sample of data with an expected number of 100 signal events over a background of 1000 events, what is the optimal cut value to maximise the significance $S = N_S/(\sqrt{N_S + N_B})$? In order to determine this, we use 10^6 simulated data events for signal and background with known mean and widths that correspond to that expected in the data. These distributions are shown in Figure 1.1, where the signal and background are distributed according to Gaussian PDFs with means of 0.1 and 0.5 and widths of 0.3 and 0.4, respectively. It can be seen that in this case there is a trade off from allowing background to pass the cuts, while retaining a reasonable signal efficiency. The figure also shows the cumulative probability distributions for signal and background, which is equivalent to the efficiency of selecting events for a cut X_0 that rejects higher values of x. The resulting significance distribution for this situation is shown in Figure 1.2 where one can see that $X_0 = 0.34$ would provide optimal separation between signal and background using this method. The significance has a maximum value of 3.8σ for this value of X_0. On further inspection of the figure one can see that for larger values of X_0, there is a drop in significance arising from an increase in background. For smaller values of X_0 there is a drop in significance as signal is removed that would otherwise contribute to a measurement.

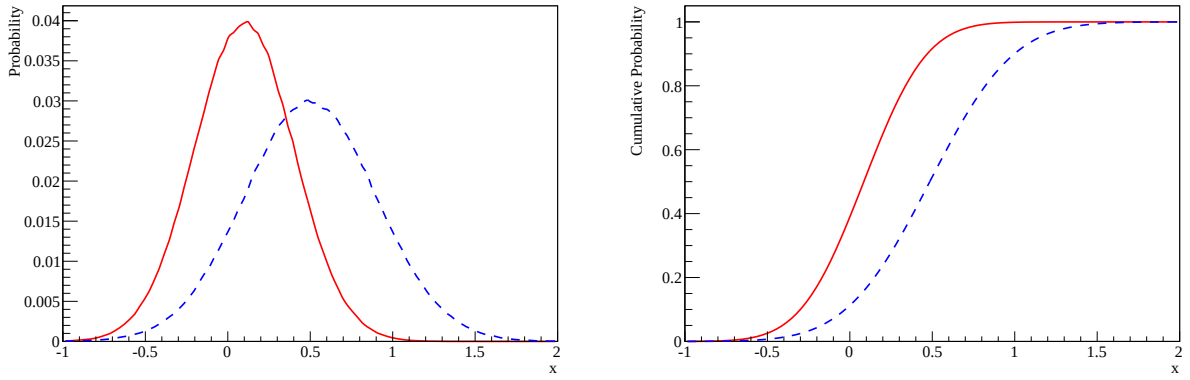

Figure 1.1: (left) The distribution in x of simulated (dashed) signal and (solid) background events, and (right) the cumulative probability distributions summing up from left-to-right for the cut based optimisation example described in the text.

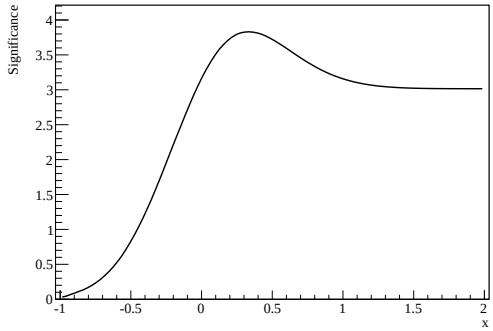

Figure 1.2: The significance computed as $N_S/(\sqrt{N_S + N_B})$ for the cut based optimisation example described in the text.

1.1.1 Optimisation of cuts as a precursor to further analysis

It should be noted that if the aim of a cut based selection of events is to then use those events with a more complicated algorithm such as a fit based optimization as discussed in Section **??**, or with an MVA like those described in the remainder of this chapter, then it doesn't make sense to optimize cut values as described above. The end result of your experiment will be the result of analysis with that more sophisticated technique - and so it is that which should be used to determine what is the *optimal* measurement to make. If one does optimize with a cut based approach, and then to perform a more sophisticated data analysis, the end result will be less precise than if one applied a loose set of cuts, then performed the subsequent data analysis. It is important to stress that control samples of data or simulated data should be used in the optimisation process so that the result you obtain is not biased. Practically this can be achieved through objective experimentation or via a blind analysis technique.

1.2 Bayesian classifier

Bayes theorem states that for data A given some theory B

$$P(B|A) = \frac{P(A|B)}{P(A)}P(B). \tag{1.2.1}$$

The terms used in this theorem have the following meanings

$P(B\|A)$	is called the ***posterior probability*** and it represents the probability of the theory or hypothesis B given the data A.
$P(A\|B)$	is the probability of observing the data A given the theory or hypothesis B.
$P(B)$	is called the ***prior probability***. This is the subjective part of the Bayesian approach and it represents our degree of belief in a given theory or hypothesis B before any measurements are made.
$P(A)$	the probability of obtaining the data A, this is a normalization constant to ensure that the total probability for anything happening is unity. More generally $P(A)$ is a sum over possible outcomes or hypotheses B_i, i.e. $P(A) = \sum_i P(A\|B_i)P(B_i)$.

Sometimes it is not possible to compute the normalisation constant $P(A)$, in which case it is possible to work with the following proportionality derived from Eq. (1.2.1)

$$P(B|A) \propto P(A|B)P(B), \tag{1.2.2}$$

and compute ratios instead of absolute probabilities. The scenario of hypothesis comparison can be extended generally to a classification problem. Given some data Ω that can be tested against a set of classifications given by H, where the i^{th} classification is given by H_i. For each event ω_j in the data set we can compute the probability $P(\omega_j|H_i)$ that the event is of the i^{th} classification. The most probable hypothesis is given by

$$P_{max}(\omega_j|H_i) = max\left[P(\omega_j|H_i)\right], \tag{1.2.3}$$

i.e. the largest value of $P(\omega_j|H_i)$ for all i is used to identify the classification for an event. Such a classifier is referred to as a Bayesian classifier.

Example: Consider the situation where one is interested in identifying three categories of event: (i) Interesting I and in need of detailed study, (ii) possibly interesting PI at some level, and (iii) not interesting NI.

If one can calculate $P(\omega_i|H_i)$, where $i = I, PI, NI$ for an event, then it is possible to compute

$$P_I = P(\omega_i|I), \tag{1.2.4}$$
$$P_{PI} = P(\omega_i|PI), \tag{1.2.5}$$
$$P_{NI} = P(\omega_i|NI). \tag{1.2.6}$$

If the largest probability for event ω_i is P_I, one will classify the event as interesting and in need of further study. Similarly if the largest probability is P_{PI} or P_{NI}, the event would be classified as possibly interesting or not interesting, respectively. A more specific example utilising Bayesian classifiers is discussed in Section 1.7.1.

1.3 Fisher Discriminant

Fisher's linear discriminant (or ***Fisher discriminant***) is a linear combination of the variables $\underline{x}$ to form a single classifier output $\mathcal{O}$ given by

$$\mathcal{O} = \sum_{i=1}^{n} \alpha_i x_i + \beta, \tag{1.3.1}$$
$$= \underline{\alpha} \cdot \underline{x} + \beta. \tag{1.3.2}$$

The sum is over the number of dimensions n in the classification problem. In order to make use of Eq. 1.3.2 in practice we need to determine the weight coefficients α_i, or equivalently the weight vector $\underline{\alpha}$. The value of β does not affect the separation between data types, it adjusts the overall central value of the resulting Fisher distribution, and in the following discussion this parameter will be set to zero.

Given the data set Ω and the knowledge of which elements in Ω are of class A and which are of class B we can compute the mean and variance of $\underline{x}$ for the two classes. These are $\underline{\mu}_{A,B}$ and $\underline{\sigma}^2_{A,B}$ where the subscript indicates the event type. Using Eq. 1.3.2 we can also compute the mean M and variance Σ^2 of the Fisher distributions for the two classes of data

$$M_{A,B} = \alpha^T \mu_{A,B} = \sum_i \alpha_i \mu_{A,B}, \tag{1.3.3}$$
$$\Sigma^2_{A,B} = \alpha^T \sigma^2_{A,B} \alpha = \sum_i \sum_j \alpha_i \sigma^2_{ij\,A,B} \alpha_j, \tag{1.3.4}$$

where we now revert to matrix notation to avoid having to explicitly write out the summations involved. In order to maximixe the separation between A and B we want to maximise the difference between M_A and M_B, while at the same time minimize the sum of the variances of the two output distributions. These requirements are expressed in the ratio

$$J(\alpha) = \frac{[M_A - M_B]^2}{\Sigma^2_A + \Sigma^2_B}, \tag{1.3.5}$$

where the squared sum of the mean values of the Fisher distribution for the two classes is

$$[M_A - M_B]^2 = \left[\sum_{i=1}^{n} \alpha_i(\mu_A - \mu_B)_i\right]\left[\sum_{j=1}^{n} \alpha_j(\mu_A - \mu_B)_j\right], \quad (1.3.6)$$

$$= \sum_{i,j=1}^{n} \alpha_i(\mu_A - \mu_B)_i(\mu_A - \mu_B)_j\alpha_j, \quad (1.3.7)$$

$$= \alpha^T B\alpha, \quad (1.3.8)$$

where the matrix B is introduced to represent the separation *between* the classes of events based on mean values. The the sum of the Fisher distribution variances is

$$\Sigma_A^2 + \Sigma_B^2 = \alpha^T\sigma_A^2\alpha + \alpha^T\sigma_B^2\alpha, \quad (1.3.9)$$

$$= \alpha^T W\alpha, \quad (1.3.10)$$

where the matrix W is the sum of covariance matrices *within* the classes. Thus

$$J(\alpha) = \frac{\alpha^T B\alpha}{\alpha^T W\alpha}. \quad (1.3.11)$$

The optimal separation between classes A and B can be found by minimising J with respect to the weight coefficients α, therefore by satisfying the condition

$$\frac{\partial J(\alpha)}{\partial\alpha} = 0. \quad (1.3.12)$$

One can show (for example see Cowan [1998]) that the minimum is found when

$$\alpha \propto W^{-1}(\underline{\mu}_A - \underline{\mu}_B), \quad (1.3.13)$$

so we are able to compute the weights α if we are able to determine the mean values $\mu_{A,B}$, and invert the matrix W. As the coefficients are determined up to some proportionality, we don't have a unique solution for the set of weights, but have a family of related solutions. This method implicitly assumes that the matrix W can be inverted. If W is singular, then one either has to change the input dimensions to produce a non-singular W matrix, or alternatively use a different classification method.

If we so wish, we can extend the form of Eq. 1.3.2 by scaling or offsetting the input data to lie within a specified range. Furthermore it is possible to scale or offset the computed $\mathcal{O}$ as desired if you want to relocate the mean value or change the range over which the classifier outputs are computed for the data. On doing this the separation between types A and B will remain optimal as defined by the Fisher algorithm.

Example: Consider the situation where we have a data sample comprising two types of events: signal (S) and background (B), each described in two dimensions that are independent: x and y. We want to compute a set of Fisher discriminant coefficients α to separate out S from B so that we can further analyse a clean sample of the signal events. From the data sample we have, we are able to compute

$$\mu_S = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}, \text{ and } \sigma_S = \begin{pmatrix} 0.3 & 0.0 \\ 0.0 & 0.2 \end{pmatrix}, \quad (1.3.14)$$

for the signal, and

$$\mu_B = \begin{pmatrix} 1.0 \\ 1.2 \end{pmatrix}, \text{ and } \sigma_B = \begin{pmatrix} 0.4 & 0.0 \\ 0.0 & 0.5 \end{pmatrix}, \quad (1.3.15)$$

for the background. The distributions of the signal and background data are shown in Figure 1.3. There are regions of the signal that are background free, and similarly there are regions of the background data that are signal free. The objective is to obtain an optimal separation between the two classes of events.

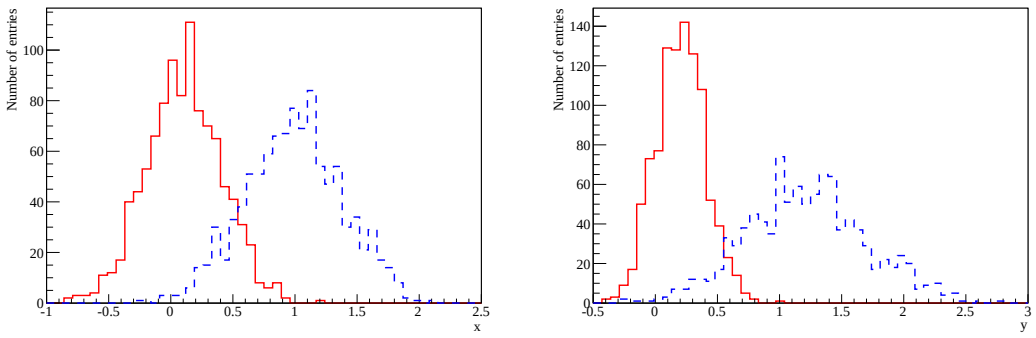

Figure 1.3: Distributions of (left) x and (right) y for (solid) signal and (dashed) background events for the example described in the text.

Given this information we can compute the difference in mean values of S and B as

$$(\mu_S - \mu_B) = \begin{pmatrix} -0.9 \\ -1.0 \end{pmatrix}, \tag{1.3.16}$$

and W is given by

$$W = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.29 \end{pmatrix}, \tag{1.3.17}$$

where the off-diagonal terms are zero as x and y are uncorrelated for both signal and background. From Eq. 1.3.13 we can determine the weight vector up to some arbitrary scale factor to be

$$\alpha = \begin{pmatrix} -3.6 \\ -3.44 \end{pmatrix}. \tag{1.3.18}$$

Figure 1.4 shows the output fisher distribution $\mathcal{O}$ obtained using the weights computed for this example. The separation between signal and background distributions in terms of $\mathcal{O}$ is better than the separation either with x or with y when one compares with the distributions in Figure 1.3. The signal distribution appears on the right hand side of the figure as a result of the convention adopted in Eq. (1.3.13) where background means are subtracted from signal ones.

1.3.1 Choice of input variables

Often we have a choice of input variables or dimensions that we want to use to separate between classes of events. Some common sense should be used when doing this, as for example if you introduce a dimension where A and B are almost completely overlapping with similar distributions, that dimension will have essentially no weight in the final Fisher discriminant that you compute. In turn you may decide that it is not worth including that variable in your classifier.

A corollary of the method is that if the mean value of the distribution of events in a given dimension is the same for both classes A and B, but the shapes of the two distributions are rather different, by definition

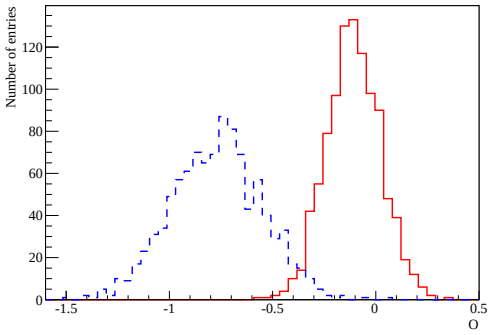

Figure 1.4: The Fisher discriminant output distribution $\mathcal{O}$ for (solid) signal and (dashed) background events for the example described in the text.

the corresponding weight α_i will be zero (this follows from Eq. 1.3.13). In such cases it makes sense to transform the variable somehow in order to make sure that the mean values of the distributions for A and B are different. One possible way to do this if you have with a common mean value for types A and B, where events are distributed differently for the two types, is to take the modulus of that variable as an input to the Fisher discriminant. The resulting distribution will not be symmetric about a common mean for A, and the variable will in turn have a greater contribution to the separation between the two classes of event.

1.4 Artificial Neural Networks

There are many variants on the concept of artificial neural networks. These are all built upon complex structures assembled from individual perceptrons (See Section 1.4.1). The type of neural network that is most commonly used in physics applications is that of a multi-layer perceptron (MLP) (See Section 1.4.2). The MLP is an ensemble of layers of perceptrons used in order to try and optimally separate classes of events. Typically there are n dimensions input to the network and only a single output, however it is also possible to configure a network with multiple outputs. Only single output MLPs are discussed here. An important, and often overlooked aspect to the use of neural networks is that of validation. After describing the MLP, there is a discussion on training methods in Section 1.4.3, and the issue of validation is discussed in Section 1.4.4.

1.4.1 Perceptrons

The fundamental building block of a neural network is the perceptron. The ***perceptron*** is an algorithmic analogy of a neuron. There are n inputs to the perceptron, these provide an impulse for the perceptron to react to. The perceptron has a pre-defined action which in turn performs some function and finally gives a response in the form of an output (See Figure 1.5). The simplest type of perceptron is a binary threshold perceptron. This takes an n dimensional input in the form of an event e_i described by the vector $\underline{x}_i$. Given $\underline{x}_i$, the perceptron is used to compute some output y_i using a so-called ***activation function***, and if $\mathcal{O}$ is above threshold the response y_i is one. If $\mathcal{O}$ is below threshold the response y_i is zero. The binary threshold perceptron algorithm is

$$\mathcal{O} = \underline{w} \cdot \underline{x}_i + b, \tag{1.4.1}$$

$$\begin{aligned} y_i &= 1 \text{ if } y_i > 0, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{1.4.2}$$

The vector $\underline{w}$ corresponds to the set of weights used to separate classes of events from each other, and b is a constant offset used to tune the binary perceptron's threshold value. If we think about what the perceptron is actually doing, one can see that we are defining a plane in an n-dimensional space as $\underline{w} \cdot \underline{x}_i + b$, and then accepting all events that occupy space on one side of this plane. The events on the other side of the plane are rejected. In order to optimally select interesting events using a single perceptron we need to determine the parameters $\underline{w}$ and β. So for each perceptron there are $n + 1$ weights (or n weights if you set β to zero) to determine. Training is discussed in more detail in Sections 1.4.3 and 1.4.4.

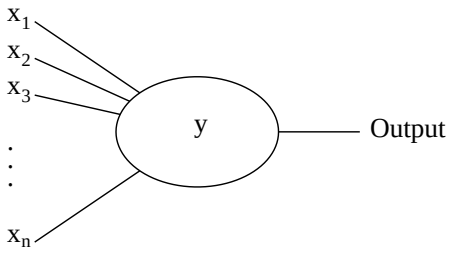

Figure 1.5: A single perceptron with n input values, an activation function y, and a single output.

The n-dimensional binary threshold perceptron is similar to the n-dimensional cut based event selection described in Section 1.1. Both algorithms are making a cut in the problem space. Conceptually this is equivalent to performing a cut-based selection as discussed in Section 1.1 when the weights are chosen such that they are aligned with the axes. Having noted this point it follows that a single perceptron won't be able to discriminate between classes of events any better than optimally cutting on data to separate them.

The function given in Eq. 1.4.2 is called the activation function of the binary threshold perceptron. In practice we are not restricted to a single type of activation function, and we are able to try other options. The following types of activation function are commonly used in perceptrons

- n-dimensional binary threshold function given by Eq. 1.4.2.
- A sigmoid (or logistic) function given by

$$y = \frac{1}{1 + e^{\underline{w} \cdot \underline{x}_i + \beta}}, \tag{1.4.3}$$

  which is a smoothly varying function with output values in the range of 0 to +1.
- The hyperbolic tangent: $y = \tanh(\underline{w} \cdot \underline{x}_i)$ which is a smoothly varying function with output values in the range -1 to $+1$.
- The radial function: $y = e^{-\underline{w} \cdot \underline{x}_i}$ which is a smoothly varying function between zero and one.

Figure 1.6 shows example distributions of the aforementioned activation functions. By using a smoothly varying activation function, as opposed to the binary threshold function described in detail previously, we are able to finely tune the decision as to whether an event is signal like or not in terms of a continuous variable. Another way of thinking about this is that it is possible to consider an event to be a little like signal or background, without having to make a hard judgment as to whether the event is definitely signal or background. This can be useful when sample distributions overlap in data as is often the case. One can think of the use of a continuous activation function as a blurred cut in parameter space for events that lie on the boundary between classification as type A or type B, compared to the hard cut that would be imposed by the binary threshold function.

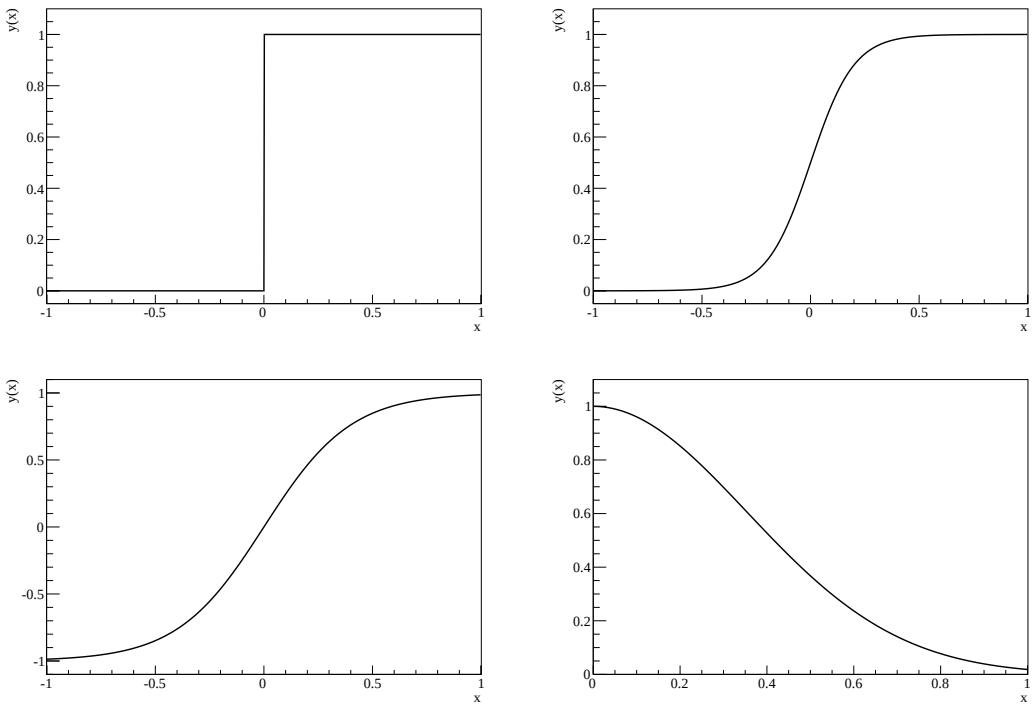

Figure 1.6: Example distributions of the (top left) binary, (top right) sigmoid, (bottom left) hyperbolic tangent, and (bottom right) radial activation functions.

1.4.2 Multi-layer Perceptron

A ***neural network*** is a combination of perceptrons, each with n inputs. It is possible to have a single perceptron to govern the output of the network, which would combine the decisions made by each of the input nodes into a single output. Usually the output would be a continuous number between either zero and one or -1 and $+1$ to indicate if an event e_i was signal like ($\mathcal{O} = +1$) or not ($\mathcal{O} = -1$ or 0 depending on the activation function).

In general a ***multi-layer perceptron*** is more complicated than this picture, and there will be a single input layer connected to the output node via one or more hidden layers. Figure 1.7 shows an MLP with n input nodes, one hidden layer of n nodes, connecting to a single output node. Each of the input nodes has n inputs, and the output of each of these nodes is transmitted to all of the nodes in the next layer. As each perceptron has at least n weight parameters to determine, if there are several hidden layers and n is large, the number of parameters to determine rapidly increases. For m perceptrons, in an input layer, each with an n dimensional input, feeding o perceptrons in a hidden layer, and a single output perceptron, then number of parameters to determine in order to compute the output of the MLP is: $n \times m$ for the input layer, $m \times o$ for the hidden layer, and o for the output node. So there would be a total of $(n + o) \times m + o$ parameters to determine. This assumes that the activation function for each node depends only on factors of $\underline{w} \cdot \underline{x}_i$. So if one has ten inputs, with a hidden layer of ten nodes, and a single output layer, the number of weights to compute is 210. Such a network would be described as having a $10 : 10 : 1$ configuration in shorthand. Even for a $5 : 5 : 1$ MLP, one would have of the order of 55 parameters to determine. Such flexibility in the configuration of a network means that a lot of care needs to be taken to ensure that the trained set of *optimal* weights is not fine tuned on fluctuations in training samples. A method of determining the weight parameters is discussed in more detail in Section 1.4.3, and Section 1.4.4 discusses the importance and main issues of validating the computed weight parameters.

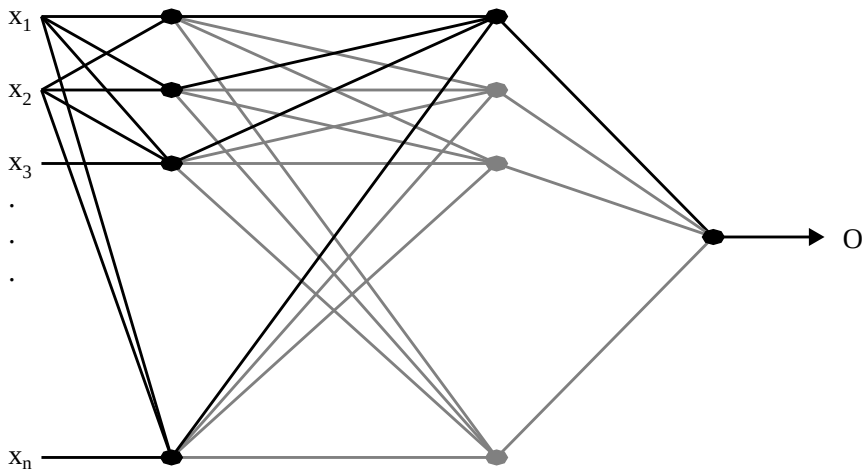

Figure 1.7: A Multi-Layer Perceptron with n input values, one hidden layer of n nodes, and a single output.

1.4.3 Training an MLP

The process of determining the weight parameters for neural network is called ***training***. There are several steps involved in training a MLP which are as follows

1. Define an algorithm to assign an error to a given set of weights.
2. Define the procedure for terminating training, based on the computed error, or other information.
3. Guess an initial set of weights to test the classification process.
4. Evaluate the error defined in step (1) for a given set of data containing (preferably) equal numbers of target types for signal and background.
5. Determine a new set of weights based on mis-classified events.
6. Iterate the last two steps until the convergence criteria defined in step (2) has been reached.
7. Validate the weights obtained via this procedure (See Section 1.4.4).

The case of a single perceptron

The ***error*** assignment for a single perceptron is based on the ability to correctly classify if an event e_i is of the appropriate type. For example signal events should be classified as signal, and background events should be classified as background. In real world problems there is an overlap between the signal and background, which is exactly the reason that leads us to try and use a complicated classifier to distinguish between the two samples of events.

If the signal classification (class A) is type= 1, and the background classification (class B) is assigned type= 0 (or -1 depending on the activation function) for an event e_i, then we can define an error on the output of a perceptron ϵ_i as

$$\epsilon_i = \frac{1}{2}(t_i - y_i)^2, \tag{1.4.4}$$

where t_i is the true target type for the event, and y_i is the output of the perceptron. The value of y_i computed for an event will depend on the set of weight vectors used in the computation, and on the form of the activation function chosen for the perceptron. The misclassification of the event is given by $t_i - y_i$, however we want to be able to sum up error terms, and so it is conventional to square this difference to maintain a positive definite quantity. Similarly the factor of $1/2$ is also conventional.

If there are N events in the data sample Ω, then the total error from a single perceptron will be given by

$$E = \sum_{i=1}^{N} \epsilon_i \tag{1.4.5}$$

$$= \frac{1}{2}\sum_{i=1}^{N}(t_i - y_i)^2. \tag{1.4.6}$$

Having computed the error on the event classification it is desirable to be able to compute a new set of weight vectors that are closer to the optimal set than the initial guess. If the initial weight vector is $\underline{w}_m$, then we want to compute a new weight vector

$$\underline{w}_{m+1} = \underline{w}_m + \Delta\underline{w}, \tag{1.4.7}$$

such that $\underline{w}_{m+1}$ is closer to the minimum of the total error function than $\underline{w}_m$. If we consider the E versus $\underline{w}$ parameter space, then we can estimate the derivative of E with respect to $\underline{w}$ as

$$\frac{\Delta E}{\Delta \underline{w}} \simeq \frac{\partial E}{\partial \underline{w}}, \tag{1.4.8}$$

so

$$\Delta E \simeq \Delta\underline{w}\frac{\partial E}{\partial \underline{w}}. \tag{1.4.9}$$

If we choose $\Delta\underline{w}$ such that it depends on the rate of change of E with respect to the weight vector we can ensure that we take a small step toward the minimum if

$$\Delta\underline{w} = -\alpha\frac{\partial E}{\partial \underline{w}}, \tag{1.4.10}$$

where α is a small positive parameter called the ***learning rate***. Thus the total change in error ΔE given by

$$\Delta E \simeq -\alpha\left(\frac{\partial E}{\partial \underline{w}}\right)^2, \tag{1.4.11}$$

which is always a negative quantity by construction. The functional form of E is given by Eq. 1.4.6, so once the activation function is defined, hence the $\underline{w}$ dependence of E has been chosen, it is possible to compute $\Delta\underline{w}$ and hence ΔE for a data sample. This method of iteratively computing weight vectors is often called the gradient descent method or the Δ Rule.

Back propagation: Training a MLP

When one moves from a single perceptron to a MLP, the error assignment algorithm is more complicated. One has to assign some importance to different contributions to the final network output, and where necessary to work back from the output layer to the input layer to modify the choice of weights. Back propagation is a generalisation of the gradient descent algorithm discussed in Section 1.4.3. The weight determination for the input layer of perceptrons is based on Eq. (1.4.7). The remainder of this section discusses the error contributions coming from nodes in the hidden and output layers.

If we consider a single connection between an input layer node and a hidden layer node, or a hidden layer node and output node, then there is weight assigned to this connection: w_{jk}, where the the indices j and k refer to the j^{th} first layer node and the k^{th} second layer node respectively. The back propagation method uses (the event index i is suppressed in the following)

$$\Delta w_{jk} = \alpha x_{\text{input}} \Delta t \frac{\partial E}{\partial \underline{w}}, \tag{1.4.12}$$

where x_{input} is the input target type (which may be a correct or an incorrect assignment), α is the learning rate, and Δt is a target type error contribution that depends on the type of node. If one has a hidden node, then that hidden node will connect to j other nodes in the previous layer, and so we need to sum up the contribution to all of those nodes when we compute Δt, hence

$$\Delta t_{HIDDEN} = \sum_{\text{j}\, nodes} (t_j - y_j) w_{jk}, \tag{1.4.13}$$

and if one considers the output node, then there is only one connection relevant for the error classification

$$\Delta t_{OUTPUT} = (t_j - y_j). \tag{1.4.14}$$

Thus the total error contribution ϵ_i for an event is given by

$$\epsilon_i = \Delta t_{INPUT} + \Delta t_{HIDDEN} + \Delta t_{OUTPUT}, \tag{1.4.15}$$

where the contribution Δt_{INPUT} can be defined in a similar way to Δt_{HIDDEN}.

As with the case of training a single perceptron, having determined the error for an ensemble of events, given an initial assumed set of weights, one can iterate and estimate a new set of weights. This process follows an analogous procedure to that outlined in Section 1.4.3. This method is a generalisation of the Δ Rule, so again it works on the concept of error minimisation through gradient descent. A detailed description of the back propagation method can be found in Rojas [1996].

1.4.4 Training validation for a Neural Network

Training validation is discussed as a sub-section in its own right to highlight the importance of this topic. It is not sufficient to assume that a computed set of weights for a network are correct. Having obtained what is assumed to be a reasonable set of weights, it is necessary to perform cross checks to ensure that the solution is not tailored to statistical fluctuations in the data used to compute the weights. There are many local minima that could be found through the minimisation of E with respect to the weight parameters – so how can one determine if the minimum obtained is really the global minimum, or if it is one of the local minima?

The problem arises as the MLP with a given set of weights w has a total classification error E as computed for some training data sample Ω_{train}. This training sample is a reference where target types of each event,

either as class A or as class B are known beyond doubt. In practice we will want to apply the MLP to a classification problem using a different data sample comprising real data $\Omega_{\rm data}$ where the target type is not certain. How do we know that the MLP will behave reasonably when applied to $\Omega_{\rm data}$? If we have sufficient training data then we can construct a statistically distinct set $\Omega_{\rm validate}$ that is equivalent to $\Omega_{\rm train}$ in all respects, but satisfies $\Omega_{\rm train} \cap \Omega_{\rm validate} = \emptyset$. If the MLP gives the same total error for both $\Omega_{\rm train}$ and $\Omega_{\rm validate}$, then it is reasonable to expect the MLP to behave as expected when we apply it to $\Omega_{\rm data}$. Hence to ensure that we have not fine tuned the weights of the MLP, we need to check the total error E obtained from the network using $\Omega_{\rm train}$, and then compute the total error E' obtained when the network is applied to $\Omega_{\rm validate}$. When $\partial E/\partial \underline{w}$ has reached a minimum, and both E and $E - E'$ are sufficiently small, we can assume that the weights computed for the MLP are not fine tuned on $\Omega_{\rm train}$ and that the training has converged. Hence we can use the network with confidence on a real data set. In order to determine if E is sufficiently small we have to set an ***error threshold*** δ by hand.

Typically we use either pure reference data samples that resemble the classes we are trying to separate, or Monte Carlo simulated data for $\Omega_{\rm train}$ and $\Omega_{\rm validate}$. While the number of events of class A or B used in training can be different, it is generally better to use equal numbers of both types of events in training.

If we reflect upon the large number of weight parameters that have to be determined when we train a neural network, the next logical question is *"How much data do we need to use when training a given network"*. There has been some discussion on this in the Literature, for example it has been noted that for a MLP with a single hidden layer, with W weight parameters that need determining and an error threshold of δ, then you should use more than W/δ events in the training sample [Baum and Haussler, 1989]. For more complicated networks this number is multiplied by a factor of $\ln(N/\delta)$, where N is the number of nodes in the network.

1.5 Decision Trees

The concept of a ***decision tree*** (DT) is derived from that of an optimal cut based selection of events. As with the previously described methods, the aim of the DT is to separate classes of events with as small a misclassification error as possible. If one has n dimensions describing classes A and B, then the root node of a DT uses the optimal set of dimensions required to separate A and B with some cut on $\underline{x}_i$. In general, the resulting sub-samples of events will contain both classes, so it is possible to consider further subdivision using an optimal combination of dimensions. This iterative process can be continued until such time as one is able to classify A and B with a satisfactory error rate. Figure 1.8 shows a schematic of a DT. Each of the nodes in the tree will split the data set into an A-like and a B-like part. As a result, the lowest level of the tree will contain a number of A-like and B-like parts. In other words, this layer contains sub-sets of Ω that are either mostly A or mostly B, each with a small misclassification error. As the optimal dimensions are used at each step to separate A and B it is quite possible that some dimensions will be used more than once while others are never used to classify events. The decision process at each node is equivalent to that of a cut based algorithm. The additional flexibility of a DT of many nodes compared to a cut based optimisation means that the algorithm has more flexibility (hence power) to separate A from B.

The function used to separate data into A and B-like parts is that of Eq. 1.4.2, where the number of dimensions used for a given node is that required to provide optimal separation (i.e. not all dimensions have to be used to make the decisions at all of the branching points in a tree). As a result there are between 1 and n weight parameters to determine per node in the tree. A corollary of this is that a decision tree with m nodes will have between m and $n \times m$ weight parameters to determine. While the weight parameter scaling issues of DTs are not as severe as those for a neural network, it follows that the issues discussed above with regard to training validation of weights for neural networks are also serious issues for DTs. Two techniques that can be used to improve the stability of the trained DTs are Boosting (Section 1.5.1) and Bagging (Section 1.5.2).

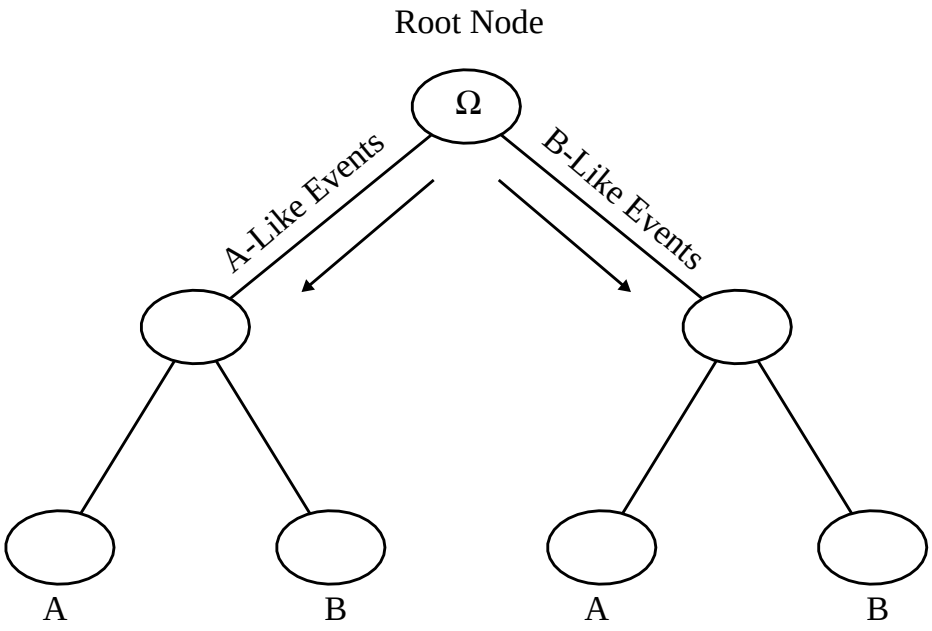

Figure 1.8: A DT with a root node above two layers of nodes that further sub-divide the data sample into pockets of A and B-like events.

1.5.1 Boosting

The aim of training a DT with a ***boosting*** algorithm (referred to as a boosted DT or BDT) is to re-weight events in favor of those that are mis-classified. The logic is that subsequent training iterations will try and correctly classify those events that were previously mis-classified. When boosting a DT one typically re-weights events with a factor $\alpha = (1-\epsilon)/\epsilon$, where ϵ is the error rate. Having re-weighted events, the total weight of the data sample is renormalised so that the sum of event weights used is constant for all iterations. In general a BDT is a more stable classification algorithm than a DT.

1.5.2 Bagging

Bagging is an alternative (or additional) method for improving the stability of a DT to that of boosting. This method involves sampling sub-sets of data from Ω, and then performing many different training cycles for the DT one for each sub sample of data. The ultimate set of weights used will be the mean value obtained for the ensemble of DTs. If the data sample Ω is not sufficient to provide statistically distinct sub-sets of data one can oversample Ω, and use each event many times for different training cycles. This re-sampling method reduces the susceptibility of a DT to statistical fluctuations.

1.6 Choosing an MVA technique

There are a number of factors that should be considered before choosing a particular MVA to separate between classes of events. Some of these factors are logical and based on taking the best classifier to do the

job, other factors are subjective and are based on the understanding of the analyst, or indeed the use case of the MVA. If a classifier will be used to provide an end decision on how probable it is that a given element is of class A or B, then your decision to use that classifier might differ from that made if you intended to use the classifier in a fit based minimization problem, or indeed as an input to another MVA algorithm.

When assessing the logical input required to understand what is the best classifier we want to understand how well class A is separated from class B in our data. There are number of ways to do this, however it can often be instructive to compute curve of the efficiency of class A vs. class B. In this case we would consider the best classifier to be the one that has the maximum efficiency of one class while minimising the efficiency of the other.

Consider the example from an experiment where we have signal and background classes for A and B. There are many input variables that distinguish between the classes based a set of quantities that can be calculated. These are the n dimensions that we will use to try and classify the data. The single output variable from a classifier is then the quantitative information that we have to decide if one algorithm is better than another at separating signal from background. Figure 1.9 shows the distribution of signal versus background efficiency for these test data. The better the event classification, the closer it will pass to the bottom right hand corner. An extreme example of this is the case of being able to identify a sample of pure signal, where the curve will pass through the point $(1, 0)$.

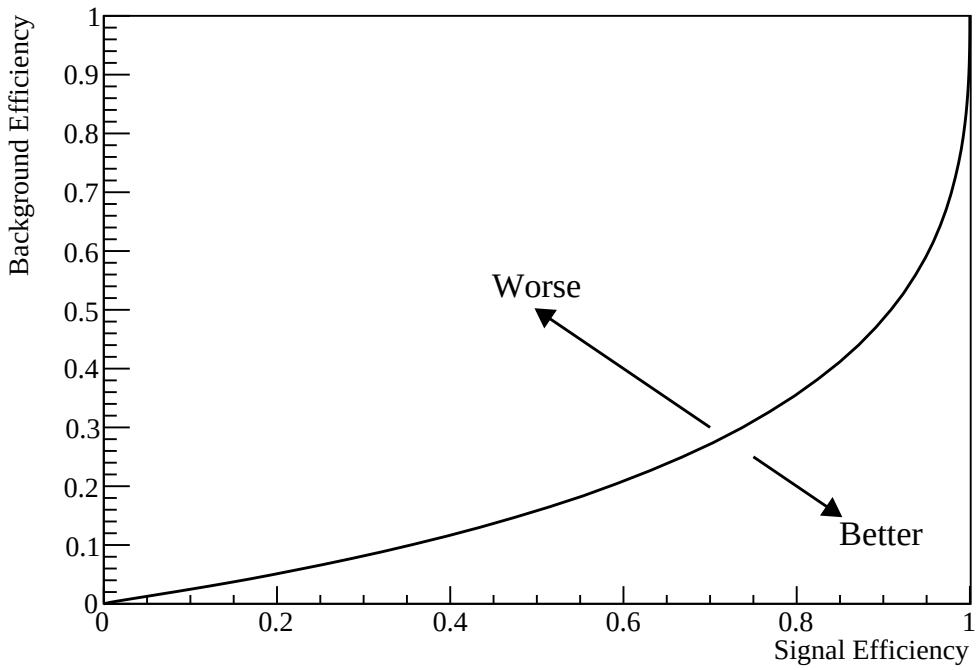

Figure 1.9: The distribution of signal versus background efficiency for a classifier output from an experiment. The better the event classification, the closer the curve will pass to the bottom right hand corner.

The fact that an algorithm gives optimal separation for the specified set of discriminating variables used, does not necessarily mean that this is the most optimal solution to the problem. In order to ascertain this we would have to compare classifiers for all possible combinations of input discriminating variables and all possible classifier algorithms. Where it is impractical to perform tests with all of the combinations, one should endeavor to test as many as is reasonable before converging upon a candidate classifier to use in the analysis of data. Having identified such a candidate, the next step in the process is to consider any subjective factors as discussed in the following that may influence the decision of an algorithm or number of dimensions

to use to classify the data.

When it comes to addressing the subjective input required to understand which classifier is the best, we should remember that there is no magic recipe to help us. However there are a number of factors that should always be adhered to:

1. Simplicity can be a key factor in determining which classifier is used. The clarity in understanding what is happening to your data in order for a given event to be classified as one type or another, or indeed to be able to easily explain what you are doing to a colleague should not be underestimated.
2. Only use a method that you understand. If you use algorithms that you do not fully understand, you may find that you have a better separation between classes using a given method, but you run a risk of having over-trained the algorithm without realising it, or falling foul of some other pathological behavior. The only way to limit such a risk to acceptable levels is to adhere to an abstinence policy of only using algorithms that you understand.
3. Where appropriate, always ensure that sufficient data are used to train and validate an algorithm. There is a necessary trade off between the desire to use as much data as possible as an input to an algorithm, and ensuring that you can validate that your resulting classifier does not suffer from over-training or some other pathology when checked against a statistically independent sample. If you fail to validate a classifier that requires training, then you should not be tempted to use that classifier for anything beyond an educational exercise.
4. Think carefully about the shape of the output classifier in the context of how you wish to use it. For example if you are using this as an input to a fit based optimisation:
   - Are you able to easily parameterize the target shapes of the classifier?
   - If the classifier is a highly irregular, or peaked shape, is there a $1:1$ mapping that you can apply in order to retain the separation power of the classifier, but obtain a distribution that can be parameterised or used as an input variable in a fit?

The quantitative and subjective inputs discussed above all play a role when we want to understand which classifier is the best for solving our problem. The discussion in this section is relevant for any MVA technique, not just the algorithms that have been described in detail here. Each problem that you are faced with will have its own unique set of quantitative and subjective factors that must be considered in order to choose which classifier is the best for a given problem. If in doubt, there is little lost in opting for the simplest algorithm. The cost in doing so is usually some loss of precision in a measurement, however sometimes this can be considered acceptable if the gain in clarity is a subjective factor that carries significant weight for your particular problem.

1.7 Case Studies

1.7.1 SPAM Filtering

Many e-mails in circulation are SPAM, and so rather than wade through such mail in order to identify those of interest, it is desirable to be able to automatically identify mails that could potentially be considered SPAM and have them moved to a separate mail folder prior to deletion. This problem is common place, and the issue of SPAM filtering using a Bayesian algorithm with inputs from both text and domain information was first raised by Sahami et al. [1998].

Here the issue is simple: We want to be able to retain all of our legitimate e-mail to read, without having to select and delete every single SPAM e-mail. One thing that is worth noting is that the penalty of making a type I error and classifying a legitimate e-mail as a SPAM mail is far greater than the penalty of making a

type II error and classifying a SPAM mail as a legitimate mail. Following the example proposed by Sahami et al. [1998], one can use a Bayesian classifier to compute a SPAM filter. The filter considered in the following is simpler than that discussed in the reference.

An event in this context corresponds to an e-mail, including the header information, the subject, and text body. Each unique element of information in an e-mail corresponds to a dimension of the problem space, and one can assign the frequency of information occurring as a value for that particular dimension. Thus one has an n dimensional vector $\underline{x}$. For example a simple e-mail greeting may include the salutation "Dear Mr(s) Smith", therefore the dimension associated with the word "Dear" would have a value of one, and so on. Having re-arranged an e-mail $\omega_i(\underline{x})$ into this vector of information, the next step is to compute $P(\omega_i(\underline{x})|LEGITIMATE)$ and $P(\omega_i(\underline{x})|SPAM)$, or some ratio as discussed in Section **??**. In order for this to happen one has to assign a significance of a given dimension in the computation of the probability that the event is legitimate or SPAM. A possible metric to use for this is

$$P(\omega_i|H_i) = \prod_{k=1}^{m} P(x_{ik}|H_i), \tag{1.7.1}$$

where x_{ik} is the value of the k^{th} dimension for the i^{th} e-mail. In other words the probability of an e-mail being legitimate (or SPAM) is given by the product of the probabilities of all elements of the e-mail being consistent with a legitimate (or SPAM) mail. In making this assertion one is assuming that all elements of the e-mail are independent of each other, so for example the individual words in the phrase "Dear Mr(s) Smith" are uncorrelated. In order to determine the $P(x_{ik}|H_i)$ for a given word or combination of words, one needs to have access to a sample of training data, where the target type is know with certainty. Given this it is possible to compute the effectiveness of a Bayesian classifier to correctly classify legitimate and SPAM e-mail. As mentioned at the start of this section, this example is a simpler method than that proposed by Sahami et al. [1998]. In fact the classification of text based information is an active field of research with very practical ramifications that goes beyond the use of Bayesian classifiers. The interested reader might wish to consult Srivastava [2009] for an overview of this topic.

1.8 Summary

The main points introduced in this section are listed below.

1. It is often possible to combine information from a multi-dimensional space of discriminating variables into a single output that can be used to separate two or more classes of data. There are a number of algorithms available in the Literature that can address this problem, each with their own benefits and limitations.
2. The algorithms discussed here are as follows
   - Cut-based selection (Section 1.1).
   - Bayesian classifier (Section 1.2).
   - The Fisher linear discriminant (Section 1.3).
   - Artificial neural networks (Section 1.4).
   - Decision trees, DT (Section 1.5).
3. Classification algorithms that require an iterative training process in order to be used, must be validated in some way. Failure to validate training could result in the inappropriate use of a non-optimal classifier. Section 1.4.3 discussed training, and validation techniques are reviewed in Section 1.4.4.
4. There is no right or wrong solution to the question of which classifier to use in a particular problem. If in doubt, the simplest solution to the problem that is well behaved and well understood could be an adequate choice.

5. In general the most optimal solution to a problem should be adopted to classify events. However, if the user does not understand an algorithm, they should be discouraged from applying the results of a 'black box' to their problem for the simple case that there may be pathologies in that solution which have been overlooked.

Bibliography

E. Baum and D. Haussler. *Neural Comp.* **1** *151-160*, 1989.

R. Bellman. *Adaptive Control Processes: A guided tour.* Oxford University Press, 1961.

G. Cowan. *Statistical Data Analysis.* Oxford University Press, 1998.

R. Rojas. *Neural Networks: A Systematic Introduction.* Springer, 1996.

M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. *Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05*, 1998.

M. Srivastava, A. Sahami. *Text Mining: Classification, Clustering, and Applications.* Chapman and Hall/CRC, 2009.