The NeuroBayes Neural Network Package

M.Feindt[a b], U.Kerzel [b *]

[a]Phi-T Physics Information Technologies GmbH, 76139 Karlsruhe, Germany (`www.phi-t.de`)

[b]Institut für Experimentelle Kernphysik, University of Karlsruhe, Germany

Detailed analysis of correlated data plays a vital role in modern analyses. We present a sophisticated neural network package based on Bayesian statistics which can be used for both classification and event-by-event prediction of the complete probability density distribution for continuous quantities. The network provides numerous possibilities to automatically preprocess the input variables and uses advanced regularisation and pruning techniques to essentially eliminate the risk of overtraining. Examples from physics and industry are given.

1. Introduction

Neural networks are inspired by a simple model of how the brain works in nature: A neuron "fires" if the stimuli received from other neurons exceed a certain threshold. In neural networks, this is described by $x_j^n = g\left(\sum_k w_{jk}^n \cdot x_k^{n-1} + \mu_j^n\right)$ where $g(t)$ is a sigmoid function and the constant μ_j^n determines the threshold. Thus the output of node j in layer n is given by the weighted sum of all nodes in layer $n-1$. Network training is then understood as the process of minimising a *loss function* by iteratively adjusting the weights w_{jk}^n such that the deviation of the actual network output from the desired output is minimised. Popular choices for the loss function are the sum of quadratic deviations or a measure of the entropy.

Neural networks are superior to other methods because they are able to learn correlations between the input variables, can incorporate information from quality variables (e.g. the return code of a certain algorithm) and do not require that all input variables are filled for each event. The latter is of particular importance when e.g. parts of a detector cannot be read out for each considered candidate.

2. The NeuroBayes Neural Network

2.1. Overview

The NeuroBayes neural network package is a highly sophisticated tool to perform multivariate analysis of correlated data. A three-layered feedforward neural network is combined with an automated preprocessing of the input variables. Users can choose from a wide range of options to optimally prepare both individual and all variables for the network training. Advanced regularisation and pruning techniques ensure a small resulting network topology where all non-significant weights and nodes are removed.

The package is split into two parts: The NeuroBayes Teacher and the NeuroBayes Expert. The Teacher uses the training dataset (simulated events or historic data) provided by the user, performs the requested preprocessing steps and trains the network to learn the complex relationships between the input variables and the training target. The statistical significance of each network weight and node is evaluated automatically during network training to ensure that only significant parts of the network topology remain. After the training the user is provided with a set of control plots to verify that the training has been successful and all information is stored in the NeuroBayes Expertise. The Expertise is then used by the NeuroBayes Expert analysing the data of interest.

*speaker

2.2. The Bayesian Approach

NeuroBayes uses Bayesian statistics to incorporate *a priori* knowledge. The conditional probability to observe B when A has already been observed is given by $P(B|A) = \frac{P(B \cap A)}{P(A)}$ (correspondingly for $P(A|B)$). Since $P(B|A) = P(A|B)$, this can be combined to *Bayes' theorem*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

This theorem is extremely important due to the interpretation A=theory and B=data: $P(data|theory)$ is then the likelihood, $P(theory)$ the Bayesian prior and $P(theory|data)$ represents the *a posteriori* probability. $P(data)$ is called the evidence. Incorporating Bayesian statistics in the NeuroBayes package prevents unphysical predictions. This can be illustrated by considering the measurement of the lifetime of a particle. The true distribution of the number of particles at a given time time t is given by $f(t) \propto \exp(-t/\tau)$. However, limited detector resolution impairs the measurement which can be modelled by smearing the true distribution by a Gaussian distribution as illustrated in figure 1: Although the true distribution (i.e. the projection on the x-axis) is never negative, values smaller than zero are obtained in the measured distribution (i.e. the projection on the y-axis). Denoting the true quantity t and the measured value by x, a typical measurement approximates $f(x|t) = f(x|t)$ which gives good results in case of good experimental resolution and far away from physical boundaries as illustrated in the upper right corner of figure 1. However, this is not the case close to physical borders as illustrated in the lower left part of the figure: While the measured distribution $f(x|t)$ can become negative, the true lifetime is always positive (semi-)definite. NeuroBayes takes this *a-priori* knowledge of the marginal distribution $f(t)$ into account and the thus never yields unphysical results.

2.3. NeuroBayes tasks

The NeuroBayes neural network package can be used both for **classification** and **shape reconstruction**. In the case of classification NeuroBayes is used to separate between two different

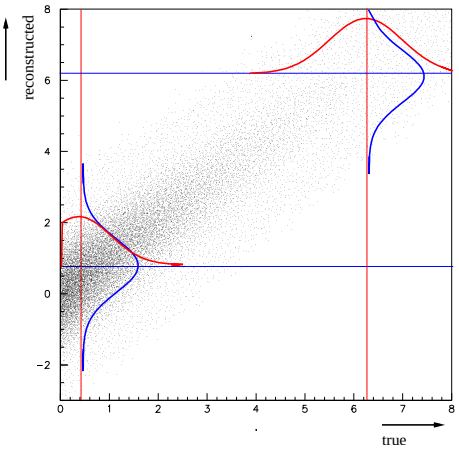

Figure 1. Illustration of the Bayesian approach: The figure shows the distribution of points following an exponential smeared by a Gaussian distribution to simulate limited detector resolution.

classes, i.e. it is determined whether a given candidate belongs to class A or class B. For example, NeuroBayes can be used to distinguish between electrons and other particles reconstructed in the detector or if a given jet of hadronising particles contains a B meson or not.

NeuroBayes also allows the prediction of complete probability density distributions $f(t|\vec{x})$ for a single multi-dimensional measurement $\vec{x}$ when used for shape-reconstruction. This feature can be used to e.g. determine the energy of an inclusively reconstructed B-hadron on a per candidate basis using for example the median of the predicted density distribution. The knowledge of the full distribution provides much more information than for example the mean value (which is determined in a standard regression analysis): The width of the distribution can be interpreted as an estimate of the uncertainty of the prediction. In particular, non-Gaussian behaviour can be correctly taken into account. This also allows selecting only high quality candidates with low uncertainties for the subsequent analysis by e.g. rejecting candidates with a large width of the pre-

dicted distribution.

2.4. Preprocessing

Preprocessing the input variables prior to network training plays a vital role in the analysis of multidimensional correlated data. This can be illustrated by a simple two-dimensional example: A hiker is to find the deepest valley in the Swiss Alps starting from a high mountain. Once very shallow valleys are rejected by a first glance, the hiker starts descending - however, without further tools he has no means to determine whether the next valley is deeper than this one. Preprocessing the input variables thus corresponds to finding the optimal starting point for the subsequent network training. Users can choose from a wide range of possible preprocessing options. Each preprocessing option is applied to either a specific input variable or to all input variables. A few options used in many applications from the extensive list of possible options are highlighted below. All variables can be normalised and (linearly) decorrelated such that the covariance matrix of the thus obtained new set of input variables is given by the unit matrix. Binary or discrete variables are automatically recognised and treated accordingly in the further processing. If NeuroBayes is used for shape-reconstruction, the inclusive shape can be fixed by introducing direct connections between the input and output layer of the network. Thus the network training corresponds to learning deviations of the inclusive shape. Furthermore, the input variables can be transformed such that the first new variable contains all linear information about the mean, the second variable all linear information about the width of the distribution to be learned, etc. The statistical significance of each input variable is computed automatically at the end of the preprocessing. A further option can be chosen to reject all (transformed) variables with a significance lower than $n \cdot \sigma$ ($n = 1, \ldots, 9$). A very important option from the extensive list of *individual* variable preprocessing is to handle variables with a default value or δ-function. This can e.g. occur if not all parts of a detector can be read out for each candidate and the thus resulting set of input variables is incomplete. Discrete input variables can be interpreted as members of ordered classes (i.e. the value of the variable indicates a certain order, e.g. a lose, medium or tight cut) or unordered classes (i.e. class A is different from class B but no further information can be obtained from the order the classes). A Bayesian regularisation scheme is applied to efficiently treat outliers far away from the bulk of the values. Instead of using the input variable directly, a new variable can be defined containing the one-dimensional correlation to the training target by performing a regularised spline-fit. The fit can either be done for a general continuous variable or can be forced to be monotonous. Furthermore, the influence on the correlation to the training target of other input variables on a given variable can be removed. All preprocessing options are very robust and work completely automatic without the need for further user interaction.

2.5. Regularisation

The use of regularisation techniques is of vital importance during network training. Employing techniques based on Bayesian statistics the NeuroBayes Teacher practically eliminates the risk of overtraining and enhances the generalisation abilities of the network. Key aspects include the constant determination of the statistical relevance of individual connections and entire network nodes, separate regularisation constants for at least three groups of weights, the automatic relevance determination of the input variables and (in the case of shape-reconstruction) the automated shape regularisation of the output nodes. Statistically insignificant network connections and even entire nodes are removed during the network training to ensure that the network learns only real features of the data. The thus obtained trained network represents the minimal topology needed to correctly reproduce the characteristics of the data while being insensitive to statistical fluctuations.

3. Examples From High Energy Physics

NeuroBayes is successfully used in many physics analyses. Significant improvement has been obtained in the identification of jets contain-

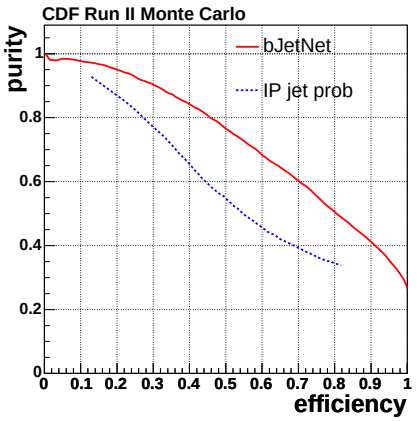

Figure 2. Improvement achieved by selecting jets containing B-hadrons with NeuroBayes compared to the previously used method. The upper curve (in red) has been obtained using the NeuroBayes Expert, the lower curve (in blue) represents the best result achieved with conventional methods. A significant improvement of up to 20% in the purity of the selected jets at the same efficiency is obtained.

ing B mesons, the determination of the b flavour (i.e. b or $\bar{b}$) and the identification of the particle type ($e^{\pm}$, $\mu^{\pm}$) in the CDF experiment. These improvements are of crucial importance in the observation of the B_s mixing and the measurement of the mixing frequency Δm_s. Figure 2 illustrates the improvement obtained using the NeuroBayes Expert to select events containing B mesons [3]. The performance is evaluated by determining purity and efficiency for multiple cuts on the prediction from the NeuroBayes Expert (or the final discriminating variable in case of other methods). The efficiency is defined as the number of signal candidates past a given network cut divided by the number of all signal candidates, purity is defined as the number of signal candidates pas a given network cut divided by the number of all particles past the cut. This results in a characteristic curve in the purity-efficiency plane. The ideal working point is the point with the smallest distance to the upper right corner representing 100% purity at 100% efficiency, though any other working point may be chosen depending on the specific needs of the respective analysis. A further application is the automated cut optimisation in case of resonance signals: Instead of optimising e.g. the ratio of signal yield divided by background by successively optimising cuts on various variables, these variables can be used as input to the NeuroBayes Teacher. The Teacher is then trained to distinguish between resonant and background events. This method can be applied also on data only in case simulated events are not available by using the side-bands of the resonance as estimates for the background. NeuroBayes can also be used to determine the quantum numbers J^{PC} of an unknown particle as discussed in [4]. Multiple networks are trained using dedicated simulated events each generated according to a specific assumption of the quantum numbers J^{PC}. Applying the NeuroBayes Expert to the data, the resonance will be either enriched or strongly suppressed depending on whether the correct J^{PC} assignment is found.

The shape reconstruction mode of NeuroBayes has been integrated into the BSAURUS [2] package used in many DELPHI analyses to inclusively determine the energy of a B hadron and provide a measure of the error. Using this method the core resolution of the pull $(E_{rec} - E_{true})/E_{true}$ has been improved from $\approx 40\%$ to $\approx 10\%$ as estimated by simulation for events measured in the LEPII phase. A further example which has been of crucial importance in the observation of B_s^{**} at DELPHI [5] is the determination of the azimuthal angle on the inclusively reconstructed B hadron. Compared to the previously best classical approach in BSAURUS [2], the resolution improved significantly. Using the measure of the uncertainty provided by NeuroBayes (e.g. the width of the predicted probability density distribution), only the high resolution candidates can be selected for the subsequent analysis.

4. Technology Transfer And Applications In Industry

The technology provided by the NeuroBayes neural network package does not only play a vital role in many physics analyses in the DELPHI and CDF collaboration but also have a wide range of applications in industry. The spin-off company *phi-t* (`www.phi-t.de`) has been founded with financial support by the exist-seed programme of the German Federal Ministry for Education and Research (BMBF) in 2002. The shape reconstruction feature has been successfully used to optimise the buy-trade strategies in investment banking.

A successful project with the Badische Gemeinde Versicherungen has made it possible to offer radically new policies for car insurances to young drivers identifying low-risk customers. Identifying customers likely to cancel the contract in the near future enables the company to get involved at an early stage of this decision and thus helps to prevent losing the customer. Further applications in industry range from the identification of (side-) effects of drugs in medicine and pharmacy, to credit scoring (Basel II), financial time series prediction, risk minimisation in trading strategies to fraud detection in insurance claims.

The power of the NeuroBayes technology has been demonstrated recently at the Data Mining Cup 2005 [6] with $\approx$ 500 participants from the whole world. Using NeuroBayes, six students from the University of Karlsruhe were able to win the positions $2, \ldots, 7$ in the final score.

5. Conclusion

The NeuroBayes neural network package provides a sophisticated tool for the analysis of highly correlated data. NeuroBayes is based on Bayesian statistics and the network output can be directly interpreted as the Bayesian *a posteriori* probability. Taking *a priori* knowledge into account, NeuroBayes will never return unphysical results. The use of advanced regularisation and pruning techniques practically eliminate the risk of overtraining and lead to an enhanced generalisation ability of the trained network. An automated and completely robust preprocessing prepares the input optimally for the network training. The various preprocessing options act either on all input variables (e.g. linear decorrelation of the variables, expansion in orthogonal polynomials,. . .) or individual variables (e.g. treatment of variables with δ-functions or default values, ordered and unordered classes, . . .) and cover a wide range of cases. NeuroBayes can be used both for classification and for the prediction of complete probability density distributions on a *per candidate* basis. Significant improvements in physics analyses were obtained in the DELPHI and CDF collaborations using NeuroBayes. The foundation of the company *phi-t* transfers this technology to industry and has led to e.g. the development of new car insurance policies.

REFERENCES

1. M. Feindt, A Neural Bayesian Estimator for Conditional Probability Densities, arXiv:physics/0402093
2. M. Feindt et al., BSAURUS - A Package For Inclusive B-Reconstruction, hep-ex/0102001
3. C. Lecci, THESIS TITLE, PhD thesis (2005), Institut für Exp. Kernphysik, University of Karlsruhe
4. C. Marino, THESIS TITLE Diploma thesis (2005, in preparation), Institut für Exp. Kernphysik, University of Karlsruhe
5. Z. Albrecht, Analyse von orbital angeregten B-Mesonen PhD thesis (2003), Institut für Exp. Kernphysik, University of Karlsruhe
6. Data Mining Cup 2005, `www.data-mining-cup.de`