Predicting the Admission Decision of a Participant to the School of Physical Education and Sports at Çukurova University by Using Different Machine Learning Methods Combined with Feature Selection

Gözde Özsert Yiğit

Mehmet Fatih Akay (advisor)

DISSERTATION.COM

Irvine, California
Boca Raton, Florida

Predicting the Admission Decision of a Participant to the School of Physical Education and Sports at Çukurova University by Using Different Machine Learning Methods Combined with Feature Selection

Dissertation.com
Boca Raton, Florida • Irvine, California
USA • 2017

ISBN-10: 1-61233-461-X (ebk.)
ISBN-13: 978-1-61233-461-5 (ebk.)

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Gözde ÖZSERT YİĞİT

PREDICTING THE ADMISSION DECISION OF A PARTICIPANT TO THE SCHOOL OF PHYSICAL EDUCATION AND SPORTS AT CUKUROVA UNIVERSITY BY USING DIFFERENT MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION

DEPARTMENT OF COMPUTER ENGINEERING

ADANA, 2016

ABSTRACT

The purpose of this thesis is to develop new hybrid admission decision prediction models by using different machine learning methods including Support Vector Machines (SVM), Multilayer Perceptron (MLP), Radial Basis Function (RBF) Network, TreeBoost (TB) and K-Means Clustering (KMC) combined with feature selection algorithms to investigate the effect of the predictor variables on the admission decision of a candidate to the School of Physical Education and Sports at Cukurova University. Three feature selection algorithms including Relief-F, F-Score and Correlation-based Feature Selection (CFS) have been considered. Experiments have been conducted on the datasets, which contain data of participants who applied to the School in 2006 and 2007. The datasets have been randomly split into training and test sets using 10-fold cross validation as well as different percentage ratios. The performance of the prediction models for the datasets has been assessed using classification accuracy, specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV). The results show that a decrease in the number of predictor variables in the prediction models usually leads to a parallel decrease in classification accuracy.

Key Words: Machine learning, feature selection, physical ability test

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Assoc. Prof. Dr. M. Fatih AKAY, for his supervision guidance, encouragements, patience, motivation, useful suggestions and his valuable time for this work.

I would like to thank members of the MSc thesis jury, Assoc. Prof. Dr. Ramazan ÇOBAN and Asst. Prof. Dr. Şule ÇOLAK, for their suggestions and corrections.

I would also like to thank Cukurova University Scientific Research Projects Center for supporting this work (Project no: FYL-2015-3845)**.**

Last but not the least, I would like to thank my husband Talat Yiğit and my family for their endless support and encouragements for my life and career.

Contents **PAGE**

1. Introduction

In order to admit a participant to the School of Physical Education and Sports at Cukurova University, the candidate has to be succesful in the physical test applied at the School. There are two parts in the physical ability test. Each of these parts contains two sub tests. The vertical jump test as well as the coordination and skill tests are applied in the first portion of the test. The second portion of the test comprises of the 30-meter dash test and 20-meter shuttle run test (Çukurova Üniversitesi Beden Eğitimi ve Spor Yüksekokulu, 2015). The details of each test is given Section 1.1 through Section 1.4.

1.1. The Vertical Jump Test

In this test, the participant waits for resetting of timing mat with weight equally balanced on both feet. After the mat is arranged, participant jumps vertically to reach the highest point he can and then he steps back on the mat. The score of the vertical jump test is calculated considering the time spent in the air. [Çukurova Üniversitesi Beden Eğitimi ve Spor Yüksekokulu, 2015). Figure 1.1 shows the vertical jump test.

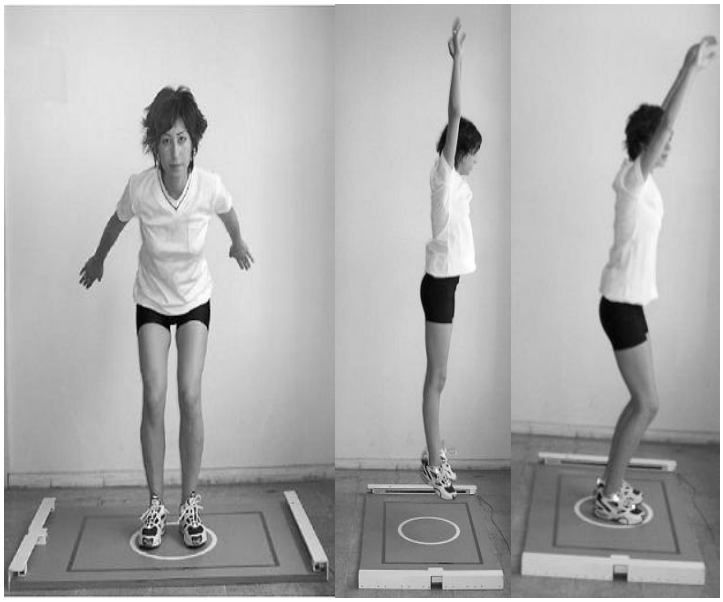

Figure 1.1.The vertical jump test

1.2. The Coordination and Skill Test

The coordination and skill test has several steps. This test starts with a front somersault. The participant gathers a ball, throws the ball up in the air and gets hold of the ball after passing across the horizontal barrier. After this step, he jumps over the balance tool and tries to keep stability while moving. Back somersault follows these steps. The other two steps are related with the participant's movement capabilities. The last part of the this test consists of sliding down and jumping over the obstacles. The setup of the coordination and skill test is shown in Figure 1.2.

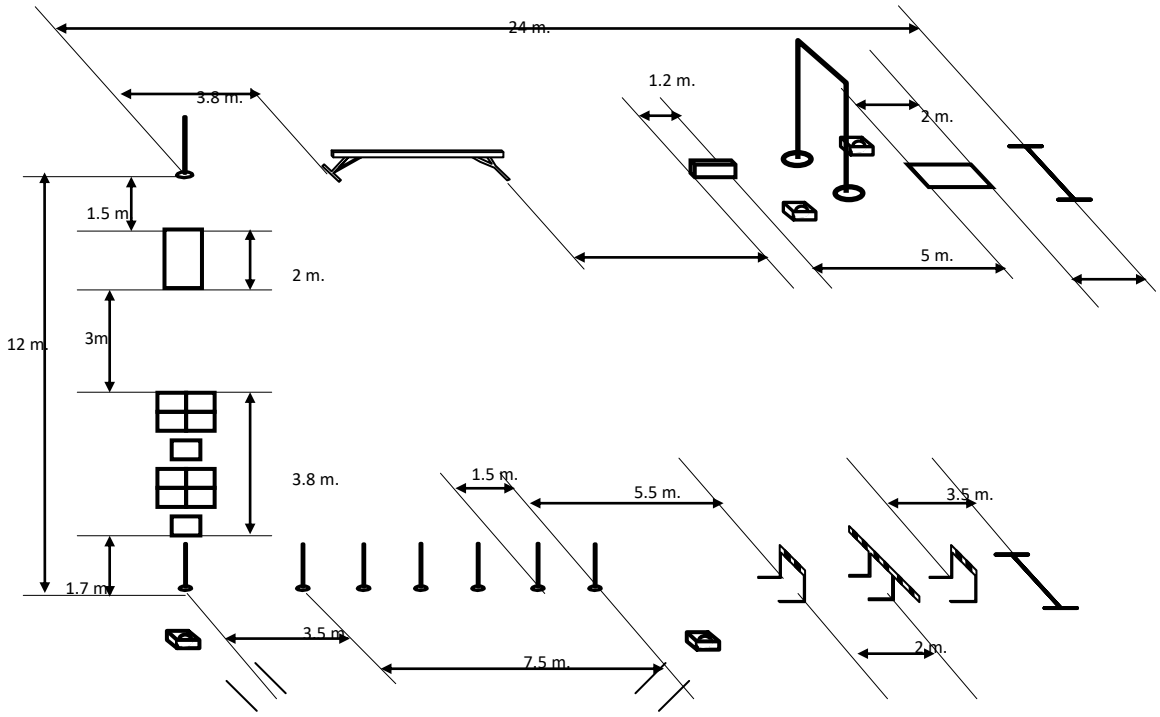

Figure 1.2.Structure of the coordination and skill test.

Each participant has two opportunities for the two tests of the first part. A participant's final score is calculated using the highest scores he achieves. Participants who are successful in these two tests are allowed to take part in the second section of the physical ability test (Çukurova Üniversitesi Beden Eğitimi ve Spor Yüksekokulu, 2015).

1.3. 30-meter Dash Test

30-meter dash test evaluates the ability of participant's to quickly gain speed on 30 meters. The test contains running one maximum sprint over 30 meters. At the beginning of the test, one foot should be one step ahead. The front foot has to be on the starting line. The participant should hold on to this position for two seconds and moving isnot permitted. The instructor should provide information for the acceleration of the participant and motivate to continue running hard through the finish line.

The time starts with the first movement of participant or if there exists time system when it is triggered, and finishes when the participant reaches the finish line or the finishing system is triggered (TopendSports, 2015). Each participant has two trials and the lowest completion time of the participant is recorded. Illustration of the 30-meter dash is given in Figure 1.3.

Figure 1.3. 30-meter dash test

1.4. 20-meter Shuttle Run Test

The shuttle run test is performed at a pitch which has a distance of 20 meters between the two end points. This testisalso called the 'beep' test. The participant stands behind the starting line and starts running when instructed. In the beginnig, the speed of participant is comparatively slow. Running between the two lines should be continued in accordance with the recorded beeps. As the test progresses, speed of the participant is increased with every beep sound and the frequency of beeps is gradually increased. If the participant arrives the line before the beep sounds, he must hold up until the beep sounds before continuing. If the participant can not arrive before the beep sounds, he is given a warning and must keep on running to the line, then turns and attempts to get up to speed with the pace within two more ‘beeps’. The test is completed when participant gets two failures in consecutive (TopendSports, 2015).

Illustration of 20-meter shuttle run test is shown in Figure 1.4.

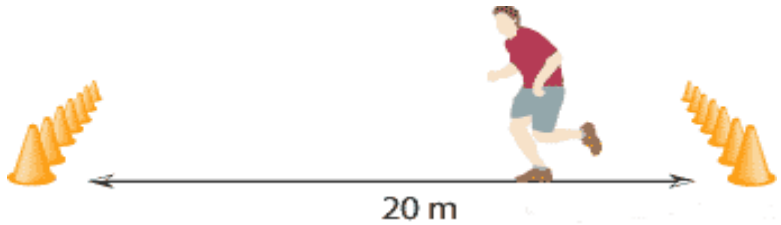

Figure 1.4. 20-meter shuttle run test

1.5. Determining a Participant's Admission Decision

A participant’s admission decision is related with the participant’s total scores from the physical ability test together with his National Student Selection Examination (NSSE) and National Student Placement Examination (NSPE) scores and Grade Point Average (GPA) at high school. The overall score of a participant who graduated from a sports branch at high school is calculated by Equation (1.1).

$$OVERALL\ SCORE = (PATS) + (0.52\ X\ GPA) + (0.36\ X\ NSPE) \quad (1.1.)$$

The overall score of a participant who graduated from another area at high school is calculated by Equation (2).

$$OVERALL\ SCORE = (PATS) + (0.16\ X\ GPA) + (0.47\ X\ NSPE) \quad (1.2.)$$

In Equation (1) and Equation (2), PATS is the physical ability test score. After the overall scores are calculated for all participants, the scores are sorted in descending order for each participant and a pre-defined number of participants are accepted to the School.

1.6. Previous Work

Developing admission decision prediction models has been an active reserach area for several years. In this regard, there exist a few studies in literature which have attempted to predict the admission decision of a candidate to the School of Physical Education and Sports of Cukurova University by using different machine learning methods.

The first study in this field was carried out by (Acikkar and Akay, 2008). Multilayer Perceptron (MLP) has been used to develop admission decision prediction models. Two datasets consisting of the real test results of participants in the years of 2006 and 2007 have been used. Several performance metrics including classification accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) have been reported for the developed prediction models. The authors concluded that MLP was a feasible tool in this application domain.

In a follow-up work by (Acikkar and Akay, 2009), Support Vector Machines (SVM) based admission decision prediction models have been developed. The same datasets were used to develop the admission decision prediction models. It was shown that SVM-based prediction models perform slightly better than MLP-based

prediction models. It was concluded that the SVM classification can be a useful tool for this application area.

Detailed results of the SVM and MLP prediction models on the 2006 and 2007 datasets for (Açıkkar and Akay, 2009; Açıkkar and Akay, 2008) are given in Table 1.1 and Table 1.2, respectively.

Table 1.1 Results of the SVM and MLP prediction models for the 2006 dataset

Study	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV(%)**
Açıkkar and Akay, 2009	97.94	97.50	98.00	96.00	99.05
Açıkkar and Akay, 2008	97.17	92.50	99.00	97.50	97.14

Table 1.2 Results of the SVM and MLP prediction models for the 2007 dataset

Study	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**PPV (%)**	**NPV (%)**
Açıkkar and Akay, 2009	93.12	85.00	97.42	94.92	92.99
Açıkkar and Akay, 2008	90.51	85.00	93.46	89.33	93.19

In (Akay et al., 2014) the authors used MLP on the same datasets to develop admission decision prediction models. Each dataset has been split into train and test sets using different ratios. More specifically, the datasets are proportionally grouped in 90-10%, 80-20%, 70-30%, 60-40% and 50-50% portions. For the 2006 dataset, the highest classification accuracy (i.e. 96.20%) has been obtained for the case of 80-20% split whereas for the 2007 dataset, the highest classification accuracy (i.e. 86.34%) has been obtained for the case of 80-20% split.

There are also other studies in literature which developed admission decision prediction models by using different machine learning methods. In (Abut et al., 2015), the authors used four classification methods including SVM, Logistic Regression, RBF Network and K-Means Clustering (KMC) on the 2006 dataset by

employing different validation techniques. The results have shown that the performance of SVM employing 10-fold cross validation is superior compared to the performance of other methods. The reported value for the highest classification accuracy was 97.90%.

In (Turhan, 2015), the author used different machine learning algorithms including SVM, MLP, Logistic Regression, RBF Network, Single Decision Tree and KMC on the 2006 and 2007 datasets. Classification accuracy and several other performance metrics have been used to assess the performance of the machine learning methods on the datasets. SVM using 10-fold cross validation achieves the highest accuracy with 97.90% and 91.45% for the 2006 and 2007 datasets, respectively. The ranking among the six classifiers in terms of achieved classification accuracy has been determined as SVM, Logistic Regression, MLP, RBF Network, Single Decision Tree and KMC.

1.7. Feature Selection

Feature selection is helpful in locating the discriminative features that are the most appropriate to predict the class. Feature selection is used in data mining and statistics. The basic approach of feature selection is to choose a subset of input variables by removing non-relevant features.

In theory, if one knew the full statistical distribution, the use of more features can yield better results. On the other hand, since using a large number of features consume memory and time, doing so may lead the algorithms become wasteful. Therefore, in the preprocessing step, it may be advantageous to pick the relevant and necessary features. Obviously, the favorable circumstances of utilizing feature selection may be enhancing comprehensibility and bringing down cost of data acquisition and handling. As a result of all the advantages, feature selection has

interested much consideration inside the machine learning, artificial intelligent and data mining communities (Sun et al., 2011).

1.8. Motivation, Purpose and Contributions of This Thesis

There is only one study (Açıkkar et al., 2014) in literature that uses machine learning methods combined with a feature selection algorithm to develop hybrid admission decision prediction models for the School of Physical Education and Sports at Cukurova University. In (Açıkkar et al., 2014), MLP combined with a feature selection algorithm has been used to develop a prediction model to predict the admission decision. As a feature selection algorithm, Relief-F has been selected. Predictor variables are gender, NSSE and NSPE scores, GPA, area at high school and scores from coordination and skill test, vertical jump test, 30-meter dash test and 20-meter shuttle run test. The results have shown that the model including all the predictor variables yields the best classification accuracies, independent of which activation function has been used at the output layer. Among the results obtained by using different activation functions, the double layered MLP model using the linear activation function has specialization yielded the best classification accuracy (i.e. 96.50%).

Apparently, more research is required with the help of several different machine learning methods combined with different feature selection algorithms in order to identify the effect of the predictor variables on the admission decision. The purpose of this thesis is to develop new hybrid admission decision prediction models by using different machine learning methods including SVM, MLP, RBF Network, TB and KMC combined with feature selection algorithms to investigate the effect of the predictor variables on the admission decision of a candidate to the School of Physical Education and Sports at Cukurova University. With the help of the prediction models that have been developed in this thesis, the participant can have an

idea about the importance of each predictor variable and hence prepare for the physical test using an approprite training program.

In this thesis, two datasets, namely 2006 dataset and 2007 dataset, have been utilized. By using the Relief-F and F-score feature selection algorithms, ranking of the attributes has been calculated. Then, based on these ranking scores, several models have been developed by removing the attribute with the lowest score at a time. In contrast to Relief-F and F-score algorithms, the CFS algorithm gives a set of selected variables to develop a single model. The models have been evaluated using different machine learning methods including SVM, MLP, RBF, TB and KMC. For model testing, 10-fold cross validation and percentage splits of data have been used. The performances of the machine learning methods for two datasets have been assessed utilizing classification accuracy, specificity, sensitivity, PPV and NPV.

This thesis has two main contributions when compared to the studies in literature. First of all, this is the first thesis in literature that develops hybrid admission decision prediction models to the School of Physical Education and Sports at Cukurova University using several machine learning methods combined with different feature selection algorithms. Secondly, by integrating feature selection algorithms into machine learning methods, this thesis yields discriminating the useful and redundant features for admission decision prediction.

1.9. Overview of Datasets

Two different datasetshave been used in this thesis. The datasets were provided by the School of Physical Education and Sports of Cukurova University. These datasets include data of participants who performed the physical ability tests in the years of 2006 and 2007. They contain nine attributes including gender, the scores from the NSSE and NSPE, GPA, the specialization area at high school, the scores from the vertical jump test, coordination and skill test, 30-meter dash test and 20-

meter shuttle run test. The datasets also include two classes. These classes are represented with the values 0 and 1, where 0 means “reject” and 1 means “admit”. There are 143 subjects (87 males and 56 females) and 117 subjects (73 males and 44 females) in the datasets of 2006 and 2007, respectively.

The predictor variables and their statistics for each dataset are given in Table 1.3. and Table 1.4., respectively.

Table 1.3. Statistical analysis of each predictor variable for the 2006 dataset

Predictor Variable	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Gender	0	1	0.61	0.49
NSSE	165.34	259.89	205.88	19.48
NSPE	186.78	262.93	226.20	14.90
GPA	37.88	93.46	75.98	10.89
Specialization area	0	1	0.14	0.348
Vertical jump test score	23	59	39.87	8.076
Coordination and skill test score	25.37	46.13	30.67	3.36
30-meter dash test score	3.74	5.32	4.30	0.39

Table 1.4. Statistical analysis of each predictor variable for the 2007 dataset

Predictor Variable	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Gender	0	1	0.62	0.49
NSSE	184.21	274.59	217.07	20.69
NSPE	190.46	283.21	237.68	14.69
GPA	37.03	99.38	77.27	10.38
Specialization area	0	1	0.094	0.29
Vertical jump test score	24	65	50	9.39
Coordination and skill test score	25.47	39.73	29.86	3.06
30-meter dash test score	3.74	5.94	4.26	0.40

2. Overview of Methods

2.1. Support Vector Machines

SVM is related to statistical learning theory (Vapnik, 1999), which was introduced in 1992 (Boser, 1992). The SVM solves classification problems by utilizing an adaptable representation of the class limits by executing automatic complexity control. With the help of this feature, the problem of overfitting is reduced.

A. Linear SVM

Assume a training set of N data points, $S = \{x_k, y_k\}$ where $x_k \in R^n$ is an input vector and $y_k \in R$ is an output vector. The SVM problems are related with hyperplanes which separate the data. The equation of the hyperplanes are determined by a vector w and a bias b. Decision function for the hyperplanesis $w^t.x + b = 0$. The margin of separation ρ can be made maximum by constructing the optimal hyperplane. The support vector method technique aims at building a classifier

$$f(x) = sign(w^t.x + b). \qquad (2.1.)$$

The $\mathbf{w}$ and b parameters are limited with

$$\min_i |w.x_i + b| \geq 1. \qquad (2.2.)$$

After the vectors are divided without any problem, and as a result of this dividing process, if the space between the nearest vector and the hyperplane is

maximum, it is called to be divided by a hyperplane. Consequently, a dividing hyperplane in standard form has to fulfill the constraints given in (2.3),

$$y_i\left(w.x_i+b\right)\geq 1,\ i=1,2,...,n. \qquad (2.3.)$$

A point $\mathbf{x}_i$ having the distance d from the hyperplane $(\mathbf{w},b)$ is,

$$d((w,b),xi)=\frac{y_i(x_i.w+b)}{\|w\|}\geq\frac{1}{\|w\|} \qquad (2.4.)$$

ρ can be calculated as

$$\rho=\frac{2}{\|w\|}\ . \qquad (2.5.)$$

Hence, SVM searches for a separating hyperplane by minimizing

$$\Phi\left(w\right)=\frac{1}{2}\left(w.w\right). \qquad (2.6.)$$

$\Phi\left(w\right)$ in (2.6) can be minimized by performing the structural risk minimization principle,

$$\|w\|^2\leq c. \qquad (2.7.)$$

h is the series of standard hyperplanes in space that has *n*-dimension and is limited by,

$$h \le \min\left[\left(R^2 c\right), d\right] + 1, \qquad (2.8.)$$

in which R is a hypersphere's radius surrounding all training vectors. As a result of this, minimizing (2.6) is equal to minimization of the upper bound.

The limitations of (2.3) can be reduced by presenting slack variables $\xi_i \ge 0,\ i = 1, 2, ..., n$, therefore (2.3) can be rewritten as

$$y_i.\left(w.x_i + b\right) \ge 1 - \xi_i,\ i = 1, 2, ..., n. \qquad (2.9.)$$

Under these circumstances, the problem of optimization becomes

$$\Phi\left(w, \xi\right) = \frac{1}{2}\left(w.w\right) + C\sum_{i=1}^{n} \xi_i\ . \qquad (2.10.)$$

In (2.10.)C is a user specified positive fixed constant. The saddle point of Lagrangian function is utilized in the solution of the problem given in (2.10).

$$L\left(w, b, \alpha, \xi, \gamma\right) = \frac{1}{2}\left(w.w\right) + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left| y_i\left(w.x_i + b\right) - 1 + \xi_i \right| - \sum_{i=1}^{n} \gamma_i \xi_i$$

(2.11.)

In (2.11), $\alpha_i \ge 0,\ \xi_i \ge 0, i = 1, 2, ..., n$ are Lagrange multipliers. (2.11) must be solved in terms of $\mathbf{w},\ b,$ and ξ_i. Classical Lagrangian duality empowers the first issue, turning (2.11) into a dual problem of it and this makes the solution easier. (2.12) shows the dual problem to be solved

$$\max_{\alpha}\left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i.x_j\right)\right] \qquad (2.12.)$$

with constraints

$$\sum_{i=1}^{n} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i = 1, 2, ..., n. \qquad (2.13.)$$

There is a unique solution for this classic quadratic optimization problem. In respect of the Kuhn-Tucker theorem of optimization theory (2.16.),

$$\alpha_i \left[y_i \left(w.x_i + b \right) - 1 \right] = 0,\ i = 1, 2, ..., n. \qquad (2.14.)$$

If $\mathbf{x}_i$ satisfy (2.15), then (2.14) will have non-zero Lagrange multipliers

$$y_i \left(\mathbf{w}.\mathbf{x}_i + b \right) = 1. \qquad (2.15.)$$

The points in (2.15.) are called support vectors (SV's). Small subset of the training vectors of the SV's determines the hyperplane. Therefore, if the optimal solution α_i^* does not take a value of zero, the classifier function can be represented as

$$f\left(x\right) = \operatorname{sgn}\left\{ \sum_{i=1}^{n} \alpha_i^* y_i \left(x_i.x \right) + b^* \right\} . \qquad (2.16.)$$

In (2.16.) b^* is the solution of (14) for any non-zero α_i^*.

B. Non-linear SVM

The majority of the datasets can not be decently divided by a linear separating hyperplane. However, they can be linearly divided if mapped into a higher dimensional field by utilizing a nonlinear mapping. Therefore, $z = \phi(x)$ that converts

the input vector x having a dimension d into a vector z having a dimension d' is defined and $\phi()$ is selected so that $\{\phi(x_i, y_i)\}$ (new training data) is divisible with a hyperplane.

The data points from the input space into some space of higher dimension are mapped by using the function

$$\varphi(.) : R^n \rightarrow R^{nh}. \qquad (2.17.)$$

Optimal function (2.9) transforms (2.18) using the same constraints,

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right)\right] \qquad (2.18.)$$

where

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\{\varphi\left(\mathbf{x}_i\right).\varphi\left(\mathbf{x}_j\right)\right\} \qquad (2.19.)$$

is the kernel function.

The exponential kernel function is given by (2.20)

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left\{-\frac{\left|\mathbf{x}_i - \mathbf{x}_j\right|^2}{2\sigma^2}\right\}, \qquad (2.20.)$$

whereas the polynomial kernel function is given by (2.21)

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left(\mathbf{x}_i.\mathbf{x}_j+1\right)^q,\ q=1,2,... \qquad (2.21.)$$

where the parameters σ and q in (2.20) and (2.21) have to be prearranged.

Then, the classifier function is specified as

$$f\left(\mathbf{x}\right)=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i K\left(\mathbf{x}_i,\mathbf{x}\right)+b^*\right\} \qquad (2.22.)$$

and b^* is the solution of (2.22) for any non-zero α_i^*.

2.2. Multi-Layer Perceptron

An MLP is a type of artificial neural network model and also feed-forward process. This model maps a set of input data over a set of convenient outputs. An MLP includes multiple layers of nodes in a coordinated graph and each layer is completely connected to the following one. Each node is a neuron with a nonlinear activation function except the input nodes. This method uses a handled learning technique. This technique is named as back propagation in order to train the network. MLP is an alteration of the standard linear perceptron and can recognize inseparable data (Imrie and Durucan, 2000). An MLP has a form as given Figure 2.1.

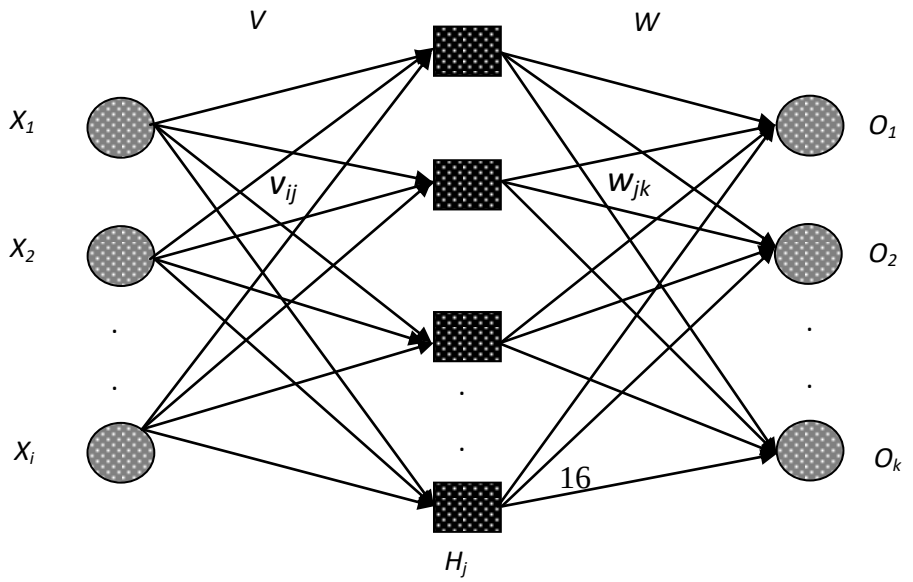

Figure 2.1. A typical MLP Structure

A data model consisting of the x_i values used in the input layer i^{th} is transmitted through the network towards j in the first hidden layer. The weighted outputs $w_{ji}x_i$ are received of the previous layer's units per hidden unit. The outputs from this process are summed. After that, these values are to be turned into an output value using an activation function.

Activation function is $f^{(n)}(x)$ at layer n. The output unit of has two-layer.

$$out_k^{(2)} = f^{(2)}(\sum_j out_j^{(1)}.w_{jk}^{(2)}) = f^{(2)}(\sum_j f^{(1)}(\sum_i in_i w_{ij}^{(1)}).w_{jk}^{(2)}) \qquad (2.23.)$$

If the activations at hidden layer are linear, (2.23.) is reduced to

$$out_k^2 = f^{(2)}(\sum_j out_j^{(1)}.w_{jk}^{(2)}) = f^{(2)}(\sum_j f^{(1)}(\sum_i in_i w_{ij}^{(2)}) \,. \qquad (2.24.)$$

Nevertheless, (2.24) is equal to a network having one layer with weight $w_{ik} = \sum_j w_{ij}^{(1)} w_{jk}^{(2)}$. This network can not be used on non-linearly separable problems.

A. Non-Linear Activation/Transfer Function

The values of the logistic sigmoid function range from 0 to 1. The standard sigmoid is basically used with the hyperbolic tangent. It has the feature,

$$f(x) = \tanh(x) = 2Sigmoid(2x) - 1 \;\; f(-x) = -f(x)\,, \qquad (2.25.)$$

and its derivative is given by (2.26.)

$$f'(x) = 1 - f(x)^2 \,. \qquad (2.26.)$$

B. Learning

The same steps are used for training $N-layer$ neural networks as the networks having a single layer. The network weights $w_{ij}^{(n)}$ are set to make the output cost function given in (2.27) minimum,

$$E_{SSE} = \frac{1}{2}\sum_{p}\sum_{j}(t\arg_j^p - out_j^{(N)p})^2 \,, \qquad (2.27.)$$

or

$$E_{CE} = -\sum_{p}\sum_{j}[t\arg_j^p .\log(out_j^{(n)p}) + (1-t\arg_j^p).\log(1-out_j^{(N)p})] \qquad (2.28.)$$

and once again this can be done by a series of gradient descent weight changes

$$\Delta w_{kl}^{(m)} = -\eta\frac{\partial E(\{w_{ij}^{(n)}\})}{\partial w_{kl}^{(m)}} \quad . \qquad (2.29.)$$

$out_j^{(N)}$ is the only output of the last layer. This becomes apparent in the error function E of the output. However, outputs of the final layer are related with all the weights of the previous layers. This learning algorithm automatically sets $out_j^{(n)}$ of the previous layers.

C. Training

Training for multi-layer networks performs in same way with networks having a single layer: