Title of the Seminar Paper

Seminar Data Mining

Chengjie Zhou
Department of Informatics
Technische Universität München
Email: chengjie.zhou@in.tum.de

***Abstract*— Text of the abstract**
***Keywords*— Data Mining**

I. Introduction

One of the most important ascpects of machine learning is classification. Typical algorithms and models that are used for classification include logistic regression, naive bayes, decision trees and so on. In the early 90s, Vladimir Vapnik and his colleagues developed a new algorithm called Support Vector Machine (SVM), which is an algorithm that is optimized for classification and regression analysis. In this paper, we will focus on the main usages of SVM, including the generalization of linear decision boundaries for classification. We will also discuss the role of kernel in SVM, as well as the implementation, evalution methods, application and many different aspects of SVM.

In order to thoroughly understand the classification problem, we first need to look at a simple example in one dimension. Suppose there are two groups of data points that are distributed separately on the number line. The classification then becomes obvious: The threshold that separates both groups simply lies in the middle of the the most outward data points from both groups. This classification method is called the Maximum Margin Classifier, since the margin that separates both groups is maximized. TODO: GRAPH

Nevertheless, the Maximum Margin Classifier is not always applicable, since the data points are not always idealy distributed in a separated way. If an outlier happens to appear in the dataset, it would push the threshold to one side, since the data point is much closer to the other group. This would result in a severe misclassification, since the data that are close to one group now belongs to the other group because of the shift of the threshold. A solution to this problem is to allow some misclassification, so that the threshold has higher bias and is less sensitive to outliers and the classifier performs better when there is new data. This margin that allows some misclassification is called the soft margin. The determination of soft margin could be tricky, since there are limitless points of thresholds to choose from. One way to find the optimal threshold is to use Cross Validation. Cross Validation is a method that splits the dataset into several parts. For each repetition, one part of the dataset is used for testing and the rest is used for training. After training through all the repetition, the average position of the threshold represents the most ideal position of the soft margin. TODO: GRAPH This classifier is called soft margin classifier, also known as support vector classifier. The name "Support Vector" derives from the fact that the data points that are closest to the threshold are called support vectors.

Support Vector Classifier works for multi-dimensional data as well. However, the dimension of the threshold will increase as the dimension of the data increases. The multidimension threshold is called hyperplane, which is formally defined as a flat affine subspace. The mathematical definition of hyperplane is given by: (Page 418 elements learning)

$$\{x \in \mathbb{R}^p : x^T\beta + \beta_0 = 0\} \tag{1}$$

TODO: Rosenblatt (Page 130 elements) TODO: Optimal (Page 132 elements)

However, one problem is still not solved. Even though the support vector classifiers allows misclassifications and is less sensitive to outliers, it is still not ideal for data that is not linearly separable. In order to solve this problem, the idea of SVM is introduced. SVM is a machine learning model that is built based on the idea of support vector classifier, but in order to cope with the non-linearly separable data, SVM implements a kernel. The kernel is a function that maps the data in a way that the data becomes linearly separable, but not transforms them. The kernel function is defined as:

$$K(x_i, x_j) = \phi(x_i)^T\phi(x_j) \tag{2}$$

II. Chapter-1

blabla

A. Subchapter

blabla with three references [1]–[3]

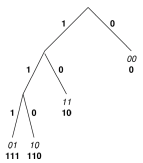

Fig. 1. Tree

III. Summary and Outlook

blabla

TABLE I

BEISPIELTABELLE

Spalte1	Spalte2
0	1

REFERENCES

[1] B. Claise, "IPFIX protocol specifications," Internet-Draft, draft-ietf-ipfix-protocol-07, December 2004.

[2] A. C. Snoeren, C. Partridge, L. A. Sanchez, C. E. Jones, F. Tchakountio, S. T. Kent, and W. T. Strayer, "Hash-based IP traceback," in *ACM SIGCOMM 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2001.

[3] A. Belenky and N. Ansari, "IP traceback with deterministic packet marking," *IEEE Communications Letters*, vol. 7, no. 4, pp. 162–164, 2003.