Mathematical Statistics
Stockholm University
Bachelor Thesis **2019:4**
http://www.math.su.se

A Comparative Study of Linear Discriminant Analysis and K-Nearest Neighbors for Statistical Classification

Nik Tavakolian*

January 2019

Abstract

We study the statistical classification methods: *Linear Discriminant Analysis* and *K-Nearest Neighbors*. We also consider extensions of these methods. The purpose of the study is to obtain a better understanding of when these methods are suitable for use. The performance of classification methods is considerably affected by the characteristics of the data. Understanding how different data attributes impact the performance of these methods is therefore important for applying them effectively to real-world classification problems. In this study we examine the theoretical properties of the classifiers and evaluate their performance statistically for three classification problems, using four evaluation metrics. The classification problems were obtained from the UCI Machine learning repository [10] and we used Accuracy, Log-Loss, Precision and Recall as evaluation metrics. We found that Linear Discriminant Analysis and its extensions are suitable when the sample size is small, and that its assumption of normality is a condition for optimality and not a prerequisite for it to perform well. K-Nearest Neighbors was found to suffer from diminished performance when the class distribution of the data was skewed, or when the number of independent variables was large. A weighted extension of K-Nearest Neighbors was found to effectively alleviate the former problem.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: nik.tavakolian@gmail.com. Supervisor: Chun-Biu Li and Disa Hansson.