Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2024:2**
http://www.math.su.se

Feature Construction of Multi-sets for Machine Learning Application

Sheila Farrahi*

February 2024

Abstract

Many machine learning algorithms require inputs in the form of fixed length vectors and cannot directly process data in the form of multi-sets. In reality, raw data does not always exist in the form of fixed length vectors and can exist in the form of unordered multi-sets of scalar values, where the number of elements in each multi-set can often vary across multiple data instances. Feature construction is a technique to find a better data representation for the machine learning algorithms in case the original representation of data is not in the form of fixed length vectors.

Statistical measures and density estimation provide insights into the data, therefore they can be used as tools in constructing fixed length vectors of features from unordered multi-sets. In this report, several methods based on statistical measures and density estimation are explored for feature constructions from unordered multi-sets.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: sh.farrahi@gmail.com. Supervisor: Chun-Biu Li.