Data Science Homework 3

Andrés Ponce, **彭思安** P76107116

April 4, 2022

1 Introduction

Current machine learning and deep learning models have shown impressive abilities to learn patterns from data. From object recognition [1] to text generation [2], many different fields have been influenced by machine learning. These models will often take a data sample as well as its label, and constantly adjust its weights to minimize the loss function. Machine learning models often require large amounts of data, which they use to learn the patterns that will be used when they are tested. A constant issue with current models is obtaining a large enough amount of data which accurately represents the data the model will encounter after training.

When trying to train a model, an important first step is ensuring the quality of the input data. Due to the messy and often chaotic nature of data collection in real applications, a preprocessing step is necessary to ensure data quality. Properties of the data such as the mean and standard deviation can tell us a lot about the nature of the data. One issue that is not so straightforward to solve is missing data. Sometimes data is missing completely at random, with no relation between the missing pieces of data, while other times some factor has an influence on the missing attributes.

To address missing data several *imputation* methods have been developed, where missing data attributes are filled in using some other properties of the data. Some methods rely on statistical properties such as the mean, while others use clustering techniques to fill in the missing data. Still others use an entire neural network to estimate the missing values.

The present assignment investigates different data imputation methods given different datasets and different amounts of missing data in each to determine the effectiveness of these methods.

2 Methods

2.1 Mean

Perhaps the simplest imputation method is using the mean. In this method, the mean along the columns of the non-missing data is used to fill in all the missing values of that column.

3 Experimental Analysis

4 Conclusions

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.