Chapter 4
Handling Missing Data

Abstract. In this chapter we address the problematic of dealing with missing data in Neural Networks (NNs). Missing data is an ubiquitous problem with numerous and diverse causes. Therefore, handling Missing Values (MVs) properly is a crucial issue. Usually, pre-processing techniques, such as imputation, are used for estimating the missing data values. However, they limit considerably the scope of application of NNs. To circumvent this limitation we introduce a novel Neural Selective Input Model (NSIM) that accommodates different transparent and bound models, while providing support for handling MVs directly. By embedding the mechanisms to support MVs we can obtain better models that reflect the uncertainty caused by unknown values. Experiments on several datasets with both different distributions and proportion of MVs show that the NSIM approach is very robust and yields good to excellent results. Furthermore, the NSIM performs better than the state-of-the-art imputation techniques either with higher prevalence of MVs in a large number of features or with a significant proportion of MVs, while delivering competitive performance in the remaining cases.

4.1 Missing Data Mechanisms

Incomplete data is an unavoidable problem for most real-world databases, which often contain missing data [105, 102]. In particular, in domains such as gene expression microarray experiments or clinical medicine, databases routinely miss pieces of information [228, 147]. Missing Values (MVs) can exist either by design (e.g. a survey questionnaire may allow people to leave unanswered questions) or by a combination of several other factors which prevent the data from being collected and/or stored. The reasons for the prevalence of MVs include among others, sensors failure, malfunction of the equipment used to record the data, data transmission problems, different patients performing different medical diagnosis tests according to their doctor and insurance coverage, merging two or more databases with a different set of attributes [68, 24, 160]. Independently of

N. Lopes and B. Ribeiro, *Machine Learning for Adaptive Many-Core Machines – A Practical Approach,* Studies in Big Data 7,
DOI: 10.1007/978-3-319-06938-8_4,

the causes associated to the existence of MVs, the fact is that most scientific data procedures are not designed to handle them [200]. In particular in the ML area, many of the most prominent algorithms (e.g. SVMs, NNs) fail to consider MVs at all. Nevertheless, handling them in a properly manner has become a fundamental requirement for building accurate models and failure to do so usually results in models with large errors [68]. To circumvent the Missing Values Problem (MVP), ML algorithms usually rely on data preprocessing techniques such as imputation for estimating the missing data. Hence, in this case, estimated data will have the same relevance and credibility of real-data. Thus, wrong estimates of crucial variables can substantially weaken the capacity of generalization of the resulting models and originate in unpredicted and potentially dramatic outcomes [130]. Moreover, estimation methods such as imputation were conventionally developed and validated under the assumption that MVs occur in a random manner. Nevertheless, this assumption does not always hold in practice [228].

The presence of MVs in data observations is one of the most frequent problems that must be faced when building ML systems [142]. Hence, given an input matrix $\mathbf{X}$, we can build a binary response indicator matrix, $\mathbf{K} \in \{0,1\}^{N \times D}$ such that:

$$K_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{ is observed} \\ 0 & \text{if } X_{ij} \text{ is missing} \end{cases} . \tag{4.1}$$

Assuming that we sort the rows (input vectors) of $\mathbf{X}$ by their number missing of variables (features). Then we can divide $\mathbf{X}$ into the observed input matrix, $\mathbf{X_{obs}}$, containing the samples for which all the variables (features) values are known, and into the unknown input matrix, $\mathbf{X_{miss}}$ containing the samples that have variables with MVs:

$$\mathbf{X} \equiv \begin{bmatrix} \mathbf{X_{obs}} \\ \mathbf{X_{miss}} \end{bmatrix} , \tag{4.2}$$

we can define the conditional distribution for the missing data as:

$$p(\mathbf{K} \mid \mathbf{X}, \xi) = p(\mathbf{K} \mid \mathbf{X_{obs}}, \mathbf{X_{miss}}, \xi) , \tag{4.3}$$

where ξ denotes the unknown parameters which define the missing data mechanism [68, 154].

Little and Rubin [119] define three types of missing data mechanisms according to their causes: MAR, MCAR and NMAR.

4.1.1 Missing At Random (MAR)

The data is said to be MAR if the causes for the *missingness* are independent of the missing variables, but traceable or predictable from other observed variables [68]. In such cases we can define the conditional distribution for the missing data as (4.4):