Machine Learning for Incomplete Data

Diego P. P. Mesquita, João P. P. Gomes[1]

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)
Campus do Pici s/n – Fortaleza – CE – Brazil

{jpaulo, diegoparente}@lia.ufc.br

Abstract. *Methods based on basis functions and similarity measures are widely used in machine learning and related fields. These methods often take for granted that data is fully observed and are not equipped to handle incomplete data in an organic manner. This assumption is often flawed, as incomplete data is a fact in various domains such as medical diagnosis and sensor analytics. Therefore, one might find it useful to be able to estimate the value of these functions in the presence of partially observed data. In this work, we present methodologies to estimate the Gaussian Kernel, the Euclidean Distance, the Epanechnikov kernel and arbitrary basis functions in the presence of possibly incomplete feature vectors.*

1. Introduction

Data completeness is a major assumption of most Machine Learning methods. In real-world problems, however, several data instances may suffer from unobserved/missing attributes. This issue, referred to as missing/incomplete data problem, may happen due to a variety of reasons such as sensor problems, device malfunction and operator mistakes [Eirola et al. 2014]. The simplest way to deal with missing data consists of removing the instances with missing attributes (listwise deletion) from the dataset. Even though this approach may work in some cases, discarding data samples usually leads to loss of important information to build a learning model [Eirola et al. 2013]. Another widely used approach is to perform a pre-processing step of missing data imputation. After filling the missing entries, any conventional learning method can be used. Examples of such an approach can be found in [Kang 2013], [Lobato et al. 2015], [Aste et al. 2015] and [Gheyas and Smith 2010].

According to Acuña and Rodrigues in [Acuña and Rodriguez 2004], problems with more than 5% of missing samples may require sophisticated handling methods. In such situations, good results can be achieved by not considering the imputation as a separate step. Instead, it is possible to design a learning method that can handle incomplete data in its formulation. By doing so, the inherent uncertainty of the imputation process is taken into account and it has shown to be beneficial in many cases [Sovilj et al. 2016]. On the other hand, direct imputation omits this uncertainty, which might be prejudicial depending on the context. To illustrate such concept, let us consider the work of [Eirola et al. 2013].

In [Eirola et al. 2013], the authors show a method to estimate the squared Euclidean distance between two vectors with missing components. A proper estimation of squared distances is fundamental in many machine learning algorithms such as distance-based methods and kernel methods. Let $X_i = (x_{i,1}, \ldots, x_{i,D})^T$ and $X_j =$

$(x_{j,1}, \ldots, x_{j,D})^T$ be two (independent) possibly incomplete feature vectors. In a statistical point-of-view, an imputation procedure can be seen as a method to fill the missing components with the most probable value. In such case, we can compute the squared Euclidean distance by using the expected values of each missing component. In this way, the squared distance may be given by:

$$\|\mathbb{E}[X_i] - \mathbb{E}[X_j]\|^2 = \sum_{d=1}^{D} (\mathbb{E}[x_{i,d}] - \mathbb{E}[x_{j,d}])^2, \tag{1}$$

An alternative approach is proposed in [Eirola et al. 2013]. Instead of estimating the expected value of the missing components, the authors propose a way to estimate the expected value of the squared Euclidean distance directly. After some straightforward mathematical developments, the authors find that the expected square Euclidean distance is given by:

$$\mathbb{E}[\|X_i - X_j\|^2] = \sum_{d=1}^{D} (\mathbb{E}[x_{i,d}] - \mathbb{E}[x_{j,d}])^2 + \mathrm{Var}[x_{i,d}] + \mathrm{Var}[x_{j,d}], \tag{2}$$

By observing Eqs. 1 and 2 we can notice that the imputation approach underestimates the value of the square distance. The difference between the two formulations is given by the variances of the missing components. The method of [Eirola et al. 2013] provides more precise estimates of the square distance and this fact also influences. Inspired by the results of [Eirola et al. 2013], we propose several strategies to estimate the expected values of other commonly used basis functions in the presence of missing data.

2. Main Contributions

The thesis in question has four major contributions, each addressing the use of important building blocks in Machine Learning in the presence of missing data. These developments mainly rely on the general assumption that data is missing-at-random (MAR) and that a model for the data can be estimated, such as Gaussian Mixture Models (GMMs).

> **[Contribution 1]** *Gaussian Kernel for incomplete data:* We present a method to estimate the expected value of the Gaussian kernel in the presence of incomplete data. For such, we model the square distance between two missing vectors as a sum of gamma-distributed random variables, whose governing parameters depend only on the non-central moments of the missing entries in the feature vectors. In this scenario, we show how that the expected value of the quantity of interest can be conveniently expressed in closed-form. The validity of the proposed method is empirically assessed under a range of conditions on simulated and real problems and the results compared to existing methods that indirectly estimate a Gaussian kernel function by either estimating squared distances or by imputing missing values and then calculating distances. Based on the experimental results, the proposed method consistently proved itself a more accurate technique and further extends the use of Gaussian kernels with incomplete data.

[Contribution 2] *Euclidean distance estimation in incomplete datasets:* We present a method to estimate the expected value of the Euclidean distance between two possibly incomplete feature vectors. Under the MAR assumption, we show that the Euclidean distance can be modeled by a Nakagami distribution, for which the parameters we express as a function of the moments of the unknown data distribution. The proposed method, named Expected Euclidean Distance (EED), was validated through a series of experiments using synthetic and real-world data. Additionally, we show and application of EED to the Minimal Learning Machine (MLM), a distance-based supervised learning method. Experimental results show that EED outperforms existing methods that estimate Euclidean distances in an indirect manner. We also observe that the application of EED to the MLM provides promising results.

[Contribution 3] *Epanechnikov kernel for incomplete data:* The Epanechnikov Kernel (EK) is a popular kernel function that has achieved promising results in many machine learning applications. We propose a method to estimate the EK when input vectors are only partially observed, *i.e.*, some of its features are missing. In the proposed method, named Expected Epanechnikov Kernel (EEK), the expected value of the kernel function is estimated given the distribution of the data and the observed values of the feature vectors.

[Contribution 4] *Arbitrary basis functions for incomplete data:* We presented a methodology to estimate the value of basis functions from incomplete feature vectors. The proposed strategies used the unscented transform to compute the expected value of the transforms. It is important to highlight that our strategies require $O(1)$ samples, more specifically three, independent of the number of missing entries on the input vector. The proposed strategies were validated in artificial and real-world scenarios, outperforming other methods in the literature.

3. Summary of publications

The publications that resulted from this thesis can be categorized into four groups according to the aforementioned contributions they are related to. A brief explanation of each paper is presented below.

[Group 1] An initial effort towards estimating the Gaussian kernel for incomplete data is presented in [Mesquita and Gomes 2017]. In this paper, we use the expected Euclidean distance formulation presented in [Eirola et al. 2013] as a building block to estimate the Gaussian kernel. Additionally, we present and application of the method in a RBF neural network. In [Mesquita et al. 2016d], we also developed a k-means algorithm for incomplete data that was used to select the centroids of the RBF network. The final formulation of the Gaussian kernel estimation, as presented in the third chapter of the thesis, is under evaluation in a well-reputed peer-reviewed journal.

[Group 2] Our first developments in estimating square distances began with an application of the method presented in [Eirola et al. 2013] to the Minimal Learning Machine [de Souza et al. 2015], which resulted in a MLM for incomplete data, presented in [Mesquita et al. 2015c]. During the making of this thesis, we proposed improvements [Mesquita et al. 2017c] for the MLM which served as building blocks for the missing data extensions. The final formulation of the Euclidean

distance estimation, as presented in the fourth chapter of the thesis, is presented in [Mesquita et al. 2017a] alongside a more elaborate version of the MLM for incomplete data.

[Group 3] Drawing inspiration from the methodologies as mentioned earlier we proposed to estimate the Euclidean distance and the Gaussian kernel, we presented a method to estimate the expected value of the Epanechnikov kernel [Mesquita et al. 2017b].

[Group 4] Our initial endeavor in this line, presented in [Mesquita et al. 2016c], consisted in an extension of the well-known Extreme Learning Machines to handle incomplete feature vectors using the unscented transform. A more general formulation to deal with generic basis functions, as presented in the sixth chapter of the thesis, is under evaluation in a well-reputed peer-reviewed journal.

Besides the contributions summarized above, other publications occurred as by-products of the work being conducted. These comprehend both contributions to Machine Learning and applications in other areas. In partnership with software engineering researchers, we developed a reject-option framework for Software Defect Prediction [Mesquita et al. 2016a]. We proposed the use of the Successive Projections Algorithm to prune Extreme Learning Machines [Mesquita et al. 2015a]. We also explored the construction of MLM ensembles [Mesquita et al. 2015b], which motivated further improvements in MLM methodology [Mesquita et al. 2017c]. Working on an application to the textile industry, we developed an ELM variant to deal with uncertainty explicitly stated with respect to input patterns [Mesquita et al. 2016b]

It worth noting that most contributions presented were already published in well-established vehicles, which indicates the relevance of the work developed in this thesis. Table 1 shows the distribution of these papers according to the QUALIS of the vehicle in which these were published.

	A1	**A2**	**B1**	**B2**	**B3**	**B4**
Journals	3	1				1
Conferences			4	2		

Tabela 1. Distribution of publications with respect to the classification of the publication vehicle.

In all of the publications mentioned so far, Mr. Mesquita is listed as first author. Nonetheless, he contributed to other 4 conference (split equally in B1 and B2 qualified venues) publications and one Journal paper (QUALIS B2). For further information, the reader can refer to Mr. Mesquita's home page [1].

4. Concluding Remarks

This paper outlined the contributions in Mr. Mesquita's Master's thesis. The developments therein presented can be seen as building blocks for adapting several Machine Learning methods to cope with incomplete data, making room for a range of possible unfoldings. The products of the work developed during the making of the thesis comprise seven conference papers and five journals articles, besides two papers still under review.

[1] `http://lia.ufc.br/~diegoparente/`

Referências

Acuña, E. and Rodriguez, C. (2004). *The Treatment of Missing Values and its Effect on Classifier Accuracy*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Aste, M., Boninsegna, M., Freno, A., and Trentin, E. (2015). Techniques for dealing with incomplete data: a tutorial and survey. *Pattern Analysis and Applications*, 18(1):1–29.

de Souza, A. H., Corona, F., Barreto, G. A., Miche, Y., and Lendasse, A. (2015). Minimal learning machine. *Neurocomput.*, 164(C):34–44.

Eirola, E., Doquire, G., Verleysen, M., and Lendasse, A. (2013). Distance estimation in numerical data sets with missing values. *Information Sciences*, 240:115 – 128.

Eirola, E., Lendasse, A., Vandewalle, V., and Biernacki, C. (2014). Mixture of gaussians for distance estimation with missing data. *Neurocomputing*, 131:32 – 42.

Gheyas, I. A. and Smith, L. S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neurocomputing*, 73(16–18):3039 – 3065.

Kang, P. (2013). Locally linear reconstruction based missing value imputation for supervised learning. *Neurocomputing*, 118:65 – 78.

Lobato, F., Sales, C., Araujo, I., Tadaiesky, V., Dias, L., Ramos, L., and Santana, A. (2015). Multi-objective genetic algorithm for missing data imputation. *Pattern Recognition Letters*, 68, Part 1:126 – 131.

Mesquita, D. P., Gomes, J., Rodrigues, L. R., and Galvao, R. K. (2015a). Pruning extreme learning machines using the successive projections algorithm. *IEEE Latin America Transactions*, 13(12):3974–3979.

Mesquita, D. P., Gomes, J. P., Junior, A. H. S., and Nobre, J. S. (2017a). Euclidean distance estimation in incomplete datasets. *Neurocomputing*, 248:11 – 18. Neural Networks : Learning Algorithms and Classification Systems.

Mesquita, D. P., Rocha, L. S., Gomes, J. P. P., and Neto, A. R. R. (2016a). Classification with reject option for software defect prediction. *Applied Soft Computing*, 49:1085 – 1093.

Mesquita, D. P. P., Gomes, Antônio Nilo Araújo Neto, J. F. Q. n. J. P. P., and Rodrigues, L. R. (2016b). Using robust extreme learning machines to predict cotton yarn strength and hairiness. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN*, pages 1–6.

Mesquita, D. P. P. and Gomes, J. P. P. (2017). Radial basis function neural networks for datasets with missing values. In Madureira, A. M., Abraham, A., Gamboa, D., and Novais, P., editors, *Intelligent Systems Design and Applications*, pages 108–115, Cham. Springer International Publishing.

Mesquita, D. P. P., Gomes, J. P. P., and Junior, A. H. S. (2015b). Ensemble of minimal learning machines for pattern classification. In Rojas, I., Joya, G., and Catala, A., editors, *IWANN - Advances in Computational Intelligence*, pages 142–152, Cham. Springer International Publishing.

Mesquita, D. P. P., Gomes, J. P. P., and Junior, A. H. S. (2017b). Epanechnikov kernel for incomplete data. *Electronics Letters*, 53(21):1408–1410.

Mesquita, D. P. P., Gomes, J. P. P., and Rodrigues, L. R. (2016c). Extreme learning machines for datasets with missing values using the unscented transform. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 85–90.

Mesquita, D. P. P., Gomes, J. P. P., and Rodrigues, L. R. (2016d). K-means for datasets with missing attributes: Building soft constraints with observed and imputed values. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN*, pages 599–604.

Mesquita, D. P. P., Gomes, J. P. P., and Souza Jr, A. H. (2015c). A minimal learning machine for datasets with missing values. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part I*, pages 565–572. Springer International Publishing.

Mesquita, D. P. P., Gomes, J. P. P., and Souza Junior, A. H. (2017c). Ensemble of efficient minimal learning machines for classification and regression. *Neural Processing Letters*, 46(3):751–766.

Sovilj, D., Eirola, E., Miche, Y., Björk, K.-M., Nian, R., Akusok, A., and Lendasse, A. (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing*, 174, Part A:220 – 231.