Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Challenges in biomedical data science: data-driven solutions to clinical questions

by

Samuele Fiorini

Theses Series **DIBRIS-TH-2017-XX**

DIBRIS, Università di Genova
Via Opera Pia, 13 16145 Genova, Italy http://www.dibris.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,

Robotica ed Ingegneria dei Sistemi

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum

Challenges in biomedical data science: data-driven solutions to clinical questions

by

Samuele Fiorini

September, 2017

Dottorato di Ricerca in Informatica ed Ingegneria dei Sistemi
Indirizzo Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy
http://www.dibris.unige.it/

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum
(S.S.D. INF/01)

Submitted by Samuele Fiorini
DIBRIS, Univ. di Genova
. . . .

Date of submission: September 12, 2017

Title: Machine Learning 4 healthcare.

Advisor: Annalisa Barla
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova
. . .

Ext. Reviewers:
Lo Scopriremo
Lo Scopriremo
Lo Scopriremo

Abstract

Abstract

Contents

1 Introduction

Part I

2 Background

2.1 What is data science and why we should care

{

- **Data engineering**
- **Data exploration**
- **Machine learning and data understanding**
- **Data visualization**

}

2.2 Challenges in biomedical data science

2.3 From clinical questions to learning task

In applied life science, the biological question at hand usually drives the data collection and therefore the statistical challenge to be solved. In order to achieve meaningful results, thorough data analysis protocol must be followed (see Section 3.2.3). In this section, my goal is to illustrate some of the most recurrent biological questions and how they can be translated into machine learning tasks.

2.3.1 How to predict phenotypes from observed data?

Starting from a collection of input measures that are likely to be related with some known target phenotype, the final goal here is to learn a model that represents the relationship between input and output. Several researches fall in this class, for instance in molecular (e.g. lab tests, gene expression, proteomics, sequencing) [Angermueller et al., 2016, Okser et al., 2014, Abraham et al., 2013] or radiomics/imaging studies (e.g. MRI, PET/SPECT, microscopy) [Min et al., 2016,

Helmstaedter et al., 2013]. Biological questions of this class are usually tackled by *supervised learning* models. In particular, when the observed clinical outcome is expressed as a one-dimensional continuous value, as in survival analysis, a *single-output regression* problem is posed. Moreover, if the outcome is vector-valued, as in the case of multiple genetic trait prediction [He et al., 2016], the problem can be cast in a *multiple-output regression* framework [Argyriou et al., 2008, Baldassarre et al., 2012]. Biological studies involving categorical outcomes translate into *classification* problems. In particular, if the clinical outcome assumes only two values, as in the *case-control* scenario, the classification problem is said to be *binary*, whilst, if multiple classes are observed, the classification task becomes *multi-class*.

2.3.2 Which variables are the most significant?

In the above case, a complementary question revolves around the interpretability of the predictive model. In particular, if dealing with high-throughput data, the main goal is to identify a relevant subset of meaningful variables for the observed phenomenon. This problem can be cast into a variable/feature selection problem [Guyon et al., 2002].

A machine learning model is said to be *sparse* when it only contains a small number of non-zero parameters, with respect to the number of features that can be measured on the objects this model represents [Hastie et al., 2015, Meier et al., 2008]. This is closely related to feature selection: if these parameters are weights on the features of the model, then only the features with non-zero weights actually enter the model and can be considered *selected*.

2.3.3 How to stratify the data?

Collecting measures from several samples, the final goal here is to divide them in homogeneous groups, according to some *similarity* criterion. In machine learning, this is usually referred to as *clustering* [Hastie et al., 2009].

2.3.4 How to represent the samples?

In order to formulate a model of some natural phenomenon, it is necessary to design and follow a suitable data collection protocol. A natural question that may arise here is whether the raw collected measures are intrinsically representative of the target phenomenon or if some transformation must be applied in order to achieve a data representation that can be successfully exploited by a learning machine. For instance, it may be plausible to assume that the data lie in a low-dimensional embedding or that they can be better represented by a richer polynomial or Gaussian expansion. A common solution, in this case, is to take advantage of *feature engineering* techniques to obtain hand crafted features. However, this process can be very

time-consuming and it may require the help of domain experts. The process of automatically identify suitable representations from the data itself is usually referred to as *(un)supervised feature learning* [Angermueller et al., 2016, Mamoshina et al., 2016].

2.3.5 Are there recurring patterns in the data?

Analyzing data coming from complex domains, one may be interested in understanding whether complex observations can be represented by some combination of simpler events. In machine learning this typically translates into *adaptive sparse coding* or *dictionary learning* problems [Masecchia et al., 2015, Alexandrov et al., 2013].

2.3.6 How to deal with missing values?

Applied life science studies must often deal with the issue of missing data. For instance, peaks can be missed in mass-spectrometry [Jung et al., 2014] or gene expression levels can be impossible to measure due to insufficient array resolution or image corruption [Stekhoven and Bühlmann, 2011, Troyanskaya et al., 2001]. Common strategies, such as discarding the samples with missing entries, or replacing the holes with the mean, median or most represented value, fall short when the missing value rate is high or the number of collected samples is relatively small. In machine learning this task usually translates into a *matrix completion* problem [Candès and Recht, 2009].

3 State of the art

3.1 Basic notation and definitions

In this thesis, the data are described as input-output pairs, $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times k}$, respectively. The i-th row of X is a d-dimensional data point $\boldsymbol{x}_i$ belonging to the input space $\mathcal{X} \subseteq \mathbb{R}^d$. The corresponding outputs $\boldsymbol{y}_i$ belong to the output space $\mathcal{Y}$.

The nature of the output space defines the problem as *binary classification* if $\mathcal{Y} = \{-1, +1\}$, *multi-category classification* if $\mathcal{Y} = \{1, 2, \ldots, k\}$, *regression* if $\mathcal{Y} \subseteq \mathbb{R}$ and *vector-valued regression* if $\mathcal{Y} \subseteq \mathbb{R}^k$.

Predictive models are functions $f : \mathcal{X} \to \mathcal{Y}$. The number of relevant variables is d^*. In feature selection tasks, the number of selected features is $\tilde{d}$.

A kernel function acting on the elements of the input space is defined as $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$, where $\phi(\boldsymbol{x})$ is a *feature map* from $\mathbb{R}^d \to \mathbb{R}^{d'}$. Feature learning algorithms project the data into a p-dimensional space.

3.2 Machine learning

3.2.1 Supervised learning

3.2.1.1 Regularization methods

3.2.1.2 Ensemble methods

3.2.1.3 Deep learning

3.2.2 Unsupervised learning

3.2.2.1 Manifold learning

3.2.2.2 Clustering

3.2.3 Model selection and evalutation

3.2.3.1 Model selection strategies

3.2.3.2 Feature selection stability

3.2.3.3 Performance metrics

Part II

4 ADENINE: a Data exploration tool

5 Model for biological age prediction [temp. title]

6 Temporal model for multiple sclerosis evolution

7 Temporal model for glucose predictions

8 Conclusion

Bibliography

[Abraham et al., 2013] Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, 37(2):184–195. [Cited on page 9.]

[Alexandrov et al., 2013] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421. [Cited on page 11.]

[Angermueller et al., 2016] Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, 12(7):878. [Cited on pages 9 and 11.]

[Argyriou et al., 2008] Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3):243–272. [Cited on page 10.]

[Baldassarre et al., 2012] Baldassarre, L., Rosasco, L., Barla, A., and Verri, A. (2012). Multi-output learning via spectral filtering. *Machine learning*, 87(3):259–301. [Cited on page 10.]

[Candès and Recht, 2009] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717. [Cited on page 11.]

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422. [Cited on page 10.]

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*, volume 2. Springer. [Cited on page 10.]

[Hastie et al., 2015] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press. [Cited on page 10.]

[He et al., 2016] He, D., Kuhn, D., and Parida, L. (2016). Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics*, 32(12):i37–i43. [Cited on page 10.]

[Helmstaedter et al., 2013] Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S., and Denk, W. (2013). Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174. [Cited on page 10.]

[Jung et al., 2014] Jung, K., Dihazi, H., Bibi, A., Dihazi, G. H., and Beißbarth, T. (2014). Adaption of the global test idea to proteomics data with missing values. *Bioinformatics*, 30(10):1424–1430. [Cited on page 11.]

[Mamoshina et al., 2016] Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Molecular pharmaceutics*, 13(5):1445–1454. [Cited on page 11.]

[Masecchia et al., 2015] Masecchia, S., Coco, S., Barla, A., Verri, A., and Tonini, G. P. (2015). Genome iny model of metastatic neuroblastoma tumorigenesis by a dictionary learning algorithm. *BMC medical genomics*, 8(1):57. [Cited on page 11.]

[Meier et al., 2008] Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71. [Cited on page 10.]

[Min et al., 2016] Min, S., Lee, B., and Yoon, S. (2016). Deep learning in bioinformatics. *arXiv preprint arXiv:1603.06430*. [Cited on page 10.]

[Okser et al., 2014] Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet*, 10(11):e1004754. [Cited on page 9.]

[Stekhoven and Bühlmann, 2011] Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118. [Cited on page 11.]

[Troyanskaya et al., 2001] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525. [Cited on page 11.]