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 14, 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

This PhD thesis is divided in two parts. Part I presents a thorough description of the multi-disciplinary prerequisites that are relevant for the comprehension of Part II, which, in turn, presents the original contributions of my work.

Part I is organized as follows: Chapter 2 introduces the concept of *data science* (Section 2.1) and its declination toward life science studies. It also describes the major challenges of the field (Section 2.2) along with several examples of the most common clinical/biological questions and their translation to data analysis tasks (Section 2.3). Chapter 3 summarizes basic notation and definitions adopted throughout the thesis (Section 3.1) and presents an overview of the statistical and technological tools that are mostly relevant for this work. In particular, this chapter defines the concept of *machine learning* from a general perspective and provides rigorous description of a selection of supervised and unsupervised learning strategies (Section 3.2). At the end of this chapter, hints on the computational requirements and implementation strategies are also presented (Section 3.3).

Part II describes the main contributions of my PhD work which consisted in the process of translating into data analysis tasks a number of biological questions coming from real-world clinical environments. For each task, this second part shows how the previously introduced tools can be exploited in order to develop statistically sound models that are capable of providing insightful answers to different clinical questions. This part is organized as follows: Chapter 4 introduces ADENINE, an open-source Python framework for large-scale data exploration I developed during my PhD. **{Chapter 5 describes a work I developed in collaboration with Gaslini hospital on biological age estimation from blood samples}**. Chapter 6 describes the development of a temporal model that aims at following the evolution of multiple sclerosis patients exploiting the use of patient-friendly and inexpensive measures such as patient centered outcomes. Chapter 7 describes a machine learning time-series forecasting approach for glucose sensor data collected by type I and type II diabetic patients.

Part I

2 Background

2.1 What is data science and why should we care?

{

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

}

- cross-disciplinary field
- Drew Conway's Data Science Venn Diagram, first published on his blog in September 2010
- data-intensive applications (maybe)

2.2 Challenges in biomedical data science

The process of modeling complex systems often implies collecting large amount of data in the field of life science, where large, multivariate and noisy measurements are typically acquired with the aim of describing multifactorial diseases.

In the era of personalized medicine, biospecimen collection and biological data management is still a challenging and expensive task [Toga and Dinov, 2015]. Only few large-scale research enterprises, such as ENCODE [Consortium et al., 2004], ADNI [Jack et al., 2008], MOPED [Kolker et al., 2012] or TCGA [2] , have sufficient financial and human resources to manage, share and

[1] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
[2] https://cancergenome.nih.gov

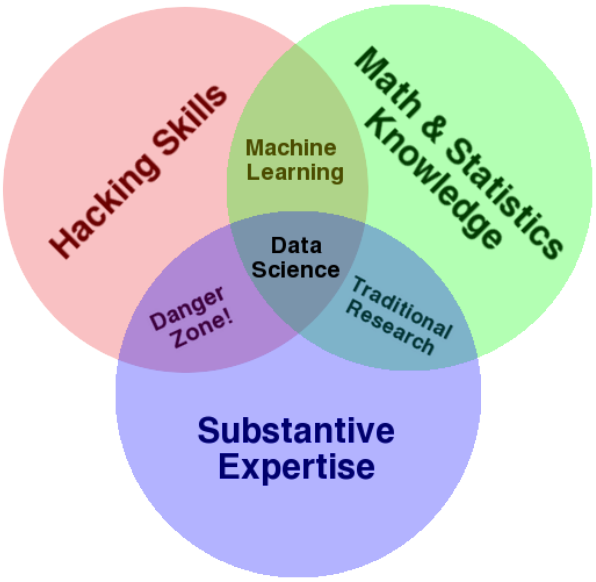

Figure 2.1: Drew Conway's Data Science Venn Diagram[1].

distribute access of heterogeneous types of biological data. To date, many biomedical studies still rely on a small number of collected samples [McNeish and Stapleton, 2016; Button et al., 2013; Yu et al., 2013]. This effect is even worse in case of rare diseases [Garg et al., 2016] or in high-throughput molecular data (e.g. genomics and proteomics) where the dimensionality of the problem can be in the order of hundreds of thousands. The setting where the number of measured variables heavily outnumbers the amount of collected samples is usually referred to as *large p small n* scenario, or simply $n \ll p$. In this case, the main goal of the learning step is often to identify a meaningful subset of relevant variables that are the most representative of the observed phenomenon. In machine learning, this is known as variable/feature selection and several techniques addressing this task were presented so far [Guyon et al., 2002]. Variable selection not only increases the prediction power of the learning machine, but it also promotes model interpretability, that is crucial in biology [Altmann et al., 2010]. Regardless [Okser et al., 2014] of the learning machine, regularization can be introduced in several ways and it is of fundamental use in order to achieve the following desired properties:

- identify models with good generalization properties, even with a limited amount of collected samples;
- achieve solutions that are robust to noise;
- learn the data structure when unknown;
- exploit prior knowledge on the data structure;

- promote interpretability performing variable selection;
- reduce the feasible set in order to help solving inverse problems.

In this paper we illustrate how regularization impacts in finding robust and meaningful models and we clarify how to choose the most suitable regularization scheme according to the biological context. The remainder of the paper is organized as follows: ....

2.3 From clinical questions to data analysis

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

In its most classical definition, the aim of modeling is to infer some unknown structure underlying the data. This process can be very hard as many unwanted and concurrent factors may mislead it resulting in a model with poor predictive power. For instance, the acquisition devices may introduce random fluctuations in the measures or the amount of collected samples may be small with respect to the number of observed variables which, in turn, may not even be representative of the target phenomenon. From a modeling standpoint, every combination of the factors above can be seen as *noise* affecting the data. Precautions in the model formulation process must be taken in order to achieve solutions that are *robust* to the noise effect. **{maybe we should not introduce noise in such a broad way (see next)}**

In the field of machine learning, a common strategy to build predictive models out of noisy data is the *regularization*. In its broader definition this refers to the process of introducing additional information in order to solve a possibly ill-posed problem [Tikhonov, 1963; Evgeniou et al., 2000]. The obtained result is a function that fits the training data while having good generalization properties, i.e. accurate predictions on previously *unseen* test data [Hastie et al., 2009]. In machine

learning, a model that fits well the training data but performs poorly on new samples is said to be *overfitting* the training set.

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

3.3 Computational requirements and implementations

- MPI
- GPU and accelerators

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, 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 11.]

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 12.]

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347. [Cited on page 10.]

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

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3):243–272. [Cited on page 11.]

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 11.]

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376. [Cited on page 10.]

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717. [Cited on page 13.]

Consortium, E. P. et al. (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640. [Cited on page 9.]

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in computational mathematics*, 13(1):1–50. [Cited on page 14.]

Garg, R. P., Dong, S., Shah, S. J., and Jonnalagadda, S. R. (2016). A bootstrap machine learning approach to identify rare disease patients from electronic health records. *CoRR*, abs/1609.01586. [Cited on page 10.]

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 pages 10 and 12.]

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*, volume 2. Springer. [Cited on pages 12 and 14.]

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press. [Cited on page 12.]

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 11.]

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 11.]

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L Whitwell, J., Ward, C., et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of magnetic resonance imaging*, 27(4):685–691. [Cited on page 9.]

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 13.]

Kolker, E., Higdon, R., Haynes, W., Welch, D., Broomall, W., Lancet, D., Stanberry, L., and Kolker, N. (2012). Moped: model organism protein expression database. *Nucleic acids research*, 40(D1):D1093–D1099. [Cited on page 9.]

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 12.]

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 12.]

McNeish, D. M. and Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28(2):295–314. [Cited on page 10.]

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 12.]

Min, S., Lee, B., and Yoon, S. (2016). Deep learning in bioinformatics. *arXiv preprint arXiv:1603.06430*. [Cited on page 11.]

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 pages 10 and 11.]

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118. [Cited on page 13.]

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. In *Soviet Math. Dokl.*, volume 5, pages 1035–1038. [Cited on page 14.]

Toga, A. W. and Dinov, I. D. (2015). Sharing big biomedical data. *Journal of big data*, 2(1):7. [Cited on page 9.]

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 13.]

Yu, D., Huber, W., and Vitek, O. (2013). Shrinkage estimation of dispersion in negative binomial models for rna-seq experiments with small sample size. *Bioinformatics*, 29(10):1275–1282. [Cited on page 10.]