MINISTRY OF EDUCATION

TECHNICAL UNIVERSITY
OF CLUJ-NAPOCA, ROMANIA

Electronics, Telecommunications and Information Technology

PhD THESIS
- ABSTRACT -

Convex optimization and learning theory methods for prediction and text summarization

PhD Student:
Marius-Claudiu Popescu, M.Sc.

PhD Supervisor:
Prof. Corneliu RUSU, PhD

Examination committee:
Chair:
Prof. Eng. **Sorin Hintea**, PhD - Technical University of Cluj-Napoca
PhD supervisor:
Prof. Eng. **Corneliu Rusu**, PhD - Technical University of Cluj-Napoca
Members:
- Prof. Eng. **Corneliu Burileanu**, PhD - POLITEHNICA University of Bucharest
- CS-I Eng. **Raul Cristian Muresan**, PhD - Transylvanian Institute of Neuroscience
- Assoc.Prof. Eng. **Lacrimioara Grama**, PhD - Technical University of Cluj-Napoca

- Cluj-Napoca -
2022

str. Memorandumului nr. 28, 400114 Cluj-Napoca, România
tel. +40-264-401200, fax +40-264-592055, secretariat tel. +40-264-202209, fax +40-264-202280
www.utcluj.ro

1 Introduction

Over the last decades, machine learning and natural language processing experienced a rapid growth. Their societal impact was also significant, changing many industries and aspects of everyday life. Such developments inevitably put forward new theoretical and practical challenges.

This thesis addresses some of these challenges. It presents four contributions to the fields of natural language processing and machine learning. The problems considered in the thesis are diverse, but we adopt a unified approach to them, based on two main tools: convex optimization and learning theory. The datasets used in the experiments are from different domains, but in the first two parts special emphasis is put on text data.

The first part contains a convex optimization formulation of the extractive text summarization problem, and a simple and scalable algorithm to solve it. The optimization program is constructed as a convex relaxation of an intuitive, but computationally hard integer programming problem. The key idea is to replace the constraint on the number of sentences in the summary with a convex surrogate. For approximately solving the program we have designed a specific projected gradient descent algorithm and analyzed its performance in terms of execution time and quality of the approximation.

Using the datasets DUC 2005 and Cornell Newsroom Summarization Dataset, we have shown empirically that the algorithm can provide competitive results for single document summarization and multi-document query-based summarization. On the Cornell Newsroom Summarization Dataset, which is used for single document sumamrization, it ranked second among the unsupervised methods tested. For the more challenging task of multi-document query-based summarization, the method was tested on the DUC 2005 Dataset. Our algorithm surpassed the other reported methods with respect to the ROUGE-SU4 metric, and it was at less than 0.01 from the top performing algorithms with respect to ROUGE-1 and ROUGE-2 metrics.

The second part of the thesis is dedicated to a geometrically motivated classification algorithm. The algorithm is a modification of the classical k-NN method, and it is based on the idea of conformal metric transformation. More concretely, our approach relies on replacing the constant metric with a variable and conformally equivalent one that is data dependent, and therefore it is more informative. We define a family of conformal transformations that, under some assumptions, induces distance functions that are efficiently computable. Using the intuition that the distances between points near a class boundary should be larger, a simple method for selecting a transformation is proposed.

We performed experiments on two datasets. The first set of experiments are with a sentiment prediction dataset, and in this case our method offers some improvements over the standard k-NN algorithm. In the second empirical analysis, we apply the method to a news taxonomy problem. In this case the results are mixed. We end the chapter with a discussion of the advantages and weaknesses of the method, and propose a number of possible improvements.

In the third part, we propose a new neural network architecture and training algorithm for generating interpretable models. The algorithm is derived using a learning bound for predictors that are convex combinations of functions from simpler classes. More explicitly, the hypothesis are polynomials over the input features, and are interpreted as convex combinations of homogeneous polynomials. Training is done by minimizing a surrogate of

the learning bound, using an iterative two phases algorithm. Basically, in the first phase the algorithm decides which monomials of higher degree should be added, and in the second phase the coefficients are recomputed by solving a convex program.

The interpretability is achieved by transforming the input features such that they can be viewed as reflecting the degree of truth of some proposition about the instance that is being classified. In this paradigm, the output of the trained neural network can be viewed as the truth value of compound proposition, and the network can be understood by humans.

We performed several experiments on binary classification datasets from different domains. The experiments show that the algorithm compares favorable in terms of accuracy and speed with other classification methods, including some new interpretable methods like Neural Additive Models and CORELS. In addition, the resulting predictor can often be understood and validated by a domain expert. The code is publicly available.

In the last part, we investigate the learnability of some hypothesis sets for regression and binary classification defined by quantum circuits. The analysis is based on concepts and results from quantum computing (Solovay–Kitaev theorem) and statistical learning theory (Rademacher complexity and covering numbers). The obtained learning bounds depend polynomially on the parameters defining the circuits set, namely, the number of qubits and the number of 1 and 2 qubits gates used for their implementation. Our setting is quite general: no realizability assumptions are made, and any 1 and 2 qubits gates are allowed. Finally, we compare the current bounds with others found in the literature, and discuss their implications for classification and regression on quantum data.

2 A Highly Scalable Method For Extractive Text Summarization Using Convex Optimization

The task of creating a short, accurate and fluent summary starting from a text document or a group of documents is called text summarization [31]. It is not too difficult for a human to perform such work, but designing and implementing an artificial system to achieve this task turned out to be challenging [26]. [1]

One method to generate a summary is by extracting and recombining the most relevant parts from the original text or texts. This process is known as extractive summarization and our work is focused on this problem. More concretely, the method described in this chapter is based on minimizing a convex function subject to some constraints, and on the properties of the l_1 norm [28]. The properties of this norm are well-known and it has many applications in signal processing (compressed sensing [7]) and statistics/machine learning (LASSO regression [37]). Among the algorithms based on the l_1 norm, the basis pursuit is a mechanism for sparse signal reconstruction from incomplete measurements, the LASSO is an automatic sparse feature selection method, while our algorithm can be interpreted as a sparse relevant parts extraction mechanism.

The summarization procedure has 3 important steps (see Figure 1). In the first one (preprocessing) a numerical representation of the text is obtained using the TF-IDF method. The second step (processing) is the most important, and it consists in solving an optimiza-

[1] The chapter is based on the paper [29]

tion problem. The step will be described in the next paragraphs. The last step (postprocessing) consist in extracting and concatenating the most relevant sentences.

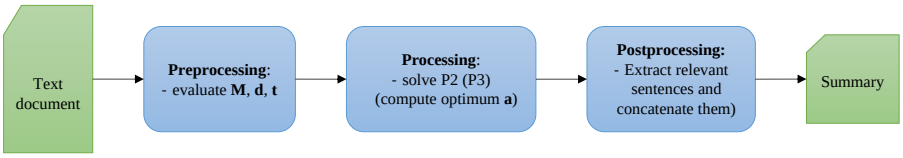

Figure 1: Diagram of the proposed method [29].

In principle, we try to capture the essential insight that a proper summary should have a numerical representation from which the vector associated with the entire text can be accurately reconstructed. More specifically, the intuition behind our approach is to select a maximum number of k sentences that best approximate the document, when some numerical representation of the text is used (k is some arbitrary positive integer). This leads as to the following integer programming problem:

$$\text{P1:} \quad \begin{cases} \min\limits_{\mathbf{a}\in\mathbb{R}^n} \left\{ \left\| \mathbf{M}^T\mathbf{a} - \mathbf{d} \right\|_2^2 - \lambda \mathbf{b}^T\mathbf{a} \right\} \\ \quad a_i \in \{0,1\},\ \forall i = 1,2,\ldots,n \\ \quad \|\mathbf{a}\|_0 \leq k. \end{cases} \tag{1}$$

The mathematical objects appearing in the equations have the following meaning. $\mathbf{M} \in \mathbb{R}^{n\times m}$ is a real valued matrix, where n represents the number of sentences and m represents the total number of distinct words. Each line of the matrix represents a sentence, and each column is associated to a word. $\mathbf{d}$ is a vector with each element d_j being the TF-IDF value for the word j. Each element a_i of the vector $\mathbf{a}$ indicates if the corresponding sentence (the i-th sentence) should be integrated in the summary ($a_i = 1$) or not ($a_i = 0$). The parameter k is the maximum number of sentences we want in the summary. The l_0 pseudo-norm represents the number of non-zero elements of a vector (this is not a true norm since it does not satisfy the triangle inequality). The real value λ is non-negative, and together with vector $\mathbf{b}$ are user provided parameters. They encode the influence of the "side information" we have, namely the *a priori* knowledge about the importance of different sentences.

By a reduction from the "subset-sum" [36] problem, we can show that P1 is an NP-hard problem. Therefore, in order to have a practical summarization technique, we need to rely on approximations. To get an approximation, we make use of the convex relaxation method, and we arrive at the following convex program:

$$\text{P2:} \quad \begin{cases} \min\limits_{\mathbf{a}\in\mathbb{R}^n} \left\{ \left\| \mathbf{M}^T\mathbf{a} - \mathbf{d} \right\|_2^2 - \lambda \mathbf{b}^T\mathbf{a} \right\} \\ \quad 0 \leq a_i \leq 1,\ \forall i = 1,2,\ldots,n \\ \quad \|\mathbf{a}\|_1 \leq k, \end{cases} \tag{2}$$

The program P2 can be solved efficiently. In order to have a highly scalable solution, we further substitute the program by a simpler one (P_{aux}, see Equation 3, in which C is a positive constant), and for the new optimization problem we design a projected gradient descent algorithm.

$$\mathrm{P}_{\text{aux}}: \quad \begin{cases} \min\limits_{\mathbf{a}\in\mathbb{R}^n} \left\{ \left\| \mathbf{M}^T\mathbf{a} - \mathbf{d} \right\|_2^2 - \lambda \mathbf{b}^T\mathbf{a} + C \cdot \max(0, \|\mathbf{a}\|_1 - k) \right\} \\ \quad 0 \le a_i \le 1, \ \forall i = 1, 2, \ldots, n \end{cases} \tag{3}$$

In order to test our method, we employed a series of experiments on artificial and real data. We first evaluated the quality of the approximation and the execution time for the proposed relaxation method. We then proceed to evaluate the method on two different tasks: single document summarization and query-based multi-document summarization. The evaluation was based on the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) tool, which became the *de facto* standard in this field [22].

For single document summarization, the most important experiments were done with the Cornell Newsroom dataset [16]. The main result is that our system has a better performance than most methods with a comparable complexity, but it is outperformed by some methods that are much more complex and make use of supervised learning.

In the case of query-based multi-document summarization, we have employed the DUC 2005 dataset [12]. Our method compares favorable with the other methods developed initially for DUC 2005, and with some newer algorithms, like the one introduced in [23]. Note that the best results were obtained with rather computationally intensive algorithms that relies on linguistic resources (like WordNet [24]).

In conclusion, the chapter presents a new algorithm for extractive text summarization based on some simple and intuitive ideas. Among the advantages of the presented method, we can underline the extensibility and the possibility to use side information and additional constraints. The method is fast enough to scale to large datasets and can be used in multiple contexts, ranging from simple, single document summarization to multi-document query-based summarization. The scalability was achieved by convex relaxation, and by designing a specific optimization algorithm for the problem at hand. Overall, the method provides a good trade off between speed and accuracy in many contexts.

2.1 Conformal transformation of the metric for k-nearest neighbors classification

The k-nearest neighbors (k-NN) algorithm is one of the most popular non-parametric classification algorithms. The main reasons are its simplicity, its ability to handle multi-class problems without any extra effort, and the fact that it does not require training [15] [11] [35]. However, in terms of accuracy it is usually surpassed by other methods, such as support vector machines (SVMs) , random forests and neural networks [35] [25] [18]. This state of affairs triggered a lot of research on how to improve the k-NN algorithm. [2]

In this chapter we investigate some potential improvements by analysing the geometry of the space of data. More concretely, we try to adjust locally the underling metric[3], such that different classes become better separated, while keeping some of the structure of the initial metric (e.g. the angles between vectors).

[2] The chapter is based on the paper [32]

[3] In this work we use the term "metric" in the sens of differential geometry, therefore an equivalent for "metric tensor". The term is in general used also to denote a distance, but we avoid this use. The terms "distance" or "distance function" are preferred.

While the idea is straight forward, obtaining an efficient algorithm that implements it is not an easy task. To keep efficiency, we make some rather crude approximations, that leads us to a new distance function. For each example in the training set, we compute a local dilatation factor. The change in the geometry is a particular kind of conformal transformation. We refer the reader to Figure 2 for a visual illustration of the idea.

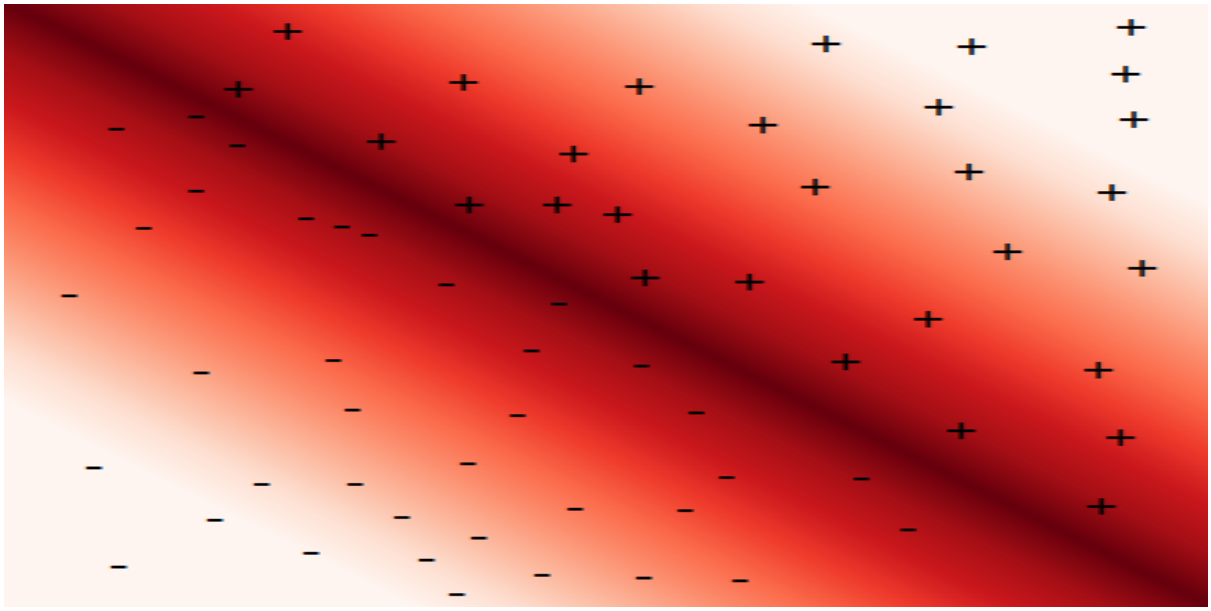

Figure 2: A depiction of a data dependent conformal transformation of $\mathbb{R}^2$. The points belong to two classes: "+" and "-". The boundary between the two classes is the principal diagonal, and the original metric is the Euclidean metric. The modified metric is generated by applying a local dilatation factor and the color intensity indicates the strength of the dilatation: near the boundary the space is strongly dilated,while far away the effect is weaker. In the white region the Euclidean metric is preserved. ([32] Copyright © 2020 IEEE)

The distance function induced by the new metric is used to replace the original distance (e.g. Euclidean distance) in the k-NN algorithm. The new method was tested on 2 datasets consisting of labeled text documents. Our experiments indicate that the method can be used to gain some performance improvement, at least in some cases.

Our approach is illustrated in Figure 3. The family of functions Ψ is chosen such that the induced distance function is easy to compute. More concretely, we select $\Psi_r : \mathbb{R}^d \to [1, \infty)$ defined by:

$$\Psi_r(x) = \begin{cases} c_i^2 & , \text{ if } x \in B_r^i, \\ 1 & , \text{ otherwise,} \end{cases} \tag{4}$$

where $r \in \mathbb{R}_+$ is some positive parameter, $c_i \in [1, \infty), i = [n]$ are a collection of values which will be computed from data, and B_r^i is the ball (with respect to the original metric g) of radius r centered on some data point x_i.

Using some geometric arguments, we can show that, under some technical conditions, the new distance is given by:

$$d_{g^c}(x, x_i) = \sqrt{(x - x_i)^T G (x - x_i)} + r(c_i - 1), \tag{5}$$

where G is the matrix associated to g.

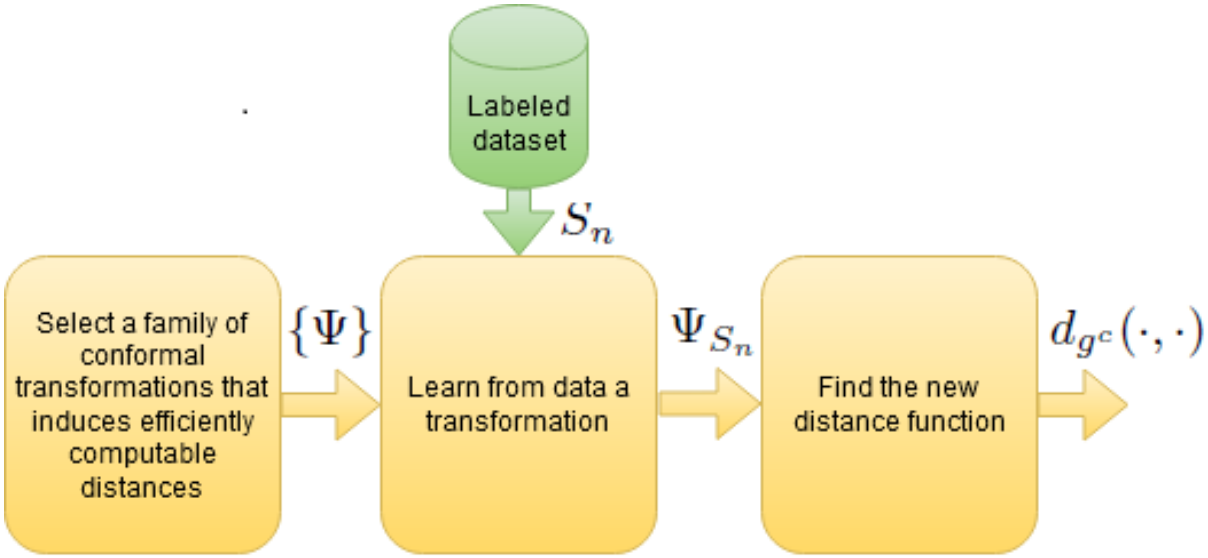

Figure 3: An overview of the proposed method. ([32] Copyright © 2020 IEEE)

Let us now fix some parameter $\epsilon > 0$, representing the radius of the ball around each training point in which we search for neighbors while computing Ψ_r. In other words, for a point x_k to be considered a neighbor of x_i, it must be true that $d_g(x_i, x_k) \leq \epsilon$. For a training point $x_i \in S_n$, $N_t^i \in \mathbb{N}$ represents the number of neighbors found around x_i (note that it is not related to k, the number of neighbors used for prediction) and N_c^i is the number of neighbors that have the same label as x_i. The associated dilatation coefficient c_i is given by:

$$c_i = \frac{10^{\frac{N_t^i+1}{N_c^i+1}}}{10}. \tag{6}$$

Therefore we get a training algorithm that will provide the new distance (see Algorithm 1).

We tested the algorithm on two datasets: a sentiment analysis dataset (Large Movie Review Dataset [14] [2]), and a news categorisation dataset (the data consist of news headlines and short descriptions from the year 2012 to 2018, obtained from HuffPost [1]).

In the case of sentiment analysis, the new algorithm gives a significant advantage over the standard k-NN (the error decreases by about 0.04-0.07, for all values of k). Taking into account that, in contrast to the distance metric learning techniques [3], our modification requires a relatively small computational overhead, the results provide a reason to further pursue this research direction. However, in the case of news categorisation, the results were mixed: for some values of k a small advantage is visible, but for other values of k, the performance actually decreases slightly.

We can conclude that the proposed algorithm is very easy to implement, and it does not have a long running time, but in terms of accuracy, the empirical findings are mixed, inviting further investigations.

Algorithm 1: k-CNN training algorithm ([32] Copyright © 2020 IEEE)

input : *Data:* S_n; *Parameters:* g, ϵ, r

for $(x_i, y_i) \in S_n$ **do**
  NN_{x_i} = FindNearestNeighbors(g, S_n, x_i, ϵ);
  $N_t = |NN_{x_i}|$;
  $N_c = 0$;
  for $(x_k, y_k) \in NN_{x_i}$ **do**
    if y_i == y_k **then**
      N_c++;
    end
  end
  $c_i = \frac{10^{\frac{N_t+1}{N_c+1}}}{10}$;
end
d_{g^c} = GenerateTheNewDistance$(g, \{c_i\}, S_n, r)$;

output: $d_{g^c}(\cdot, \cdot)$ *(the new distance)*

3 Deep interpretable polynomial networks

Machine learning models are often treated as "black boxes", meaning that only the performance of the predictor is relevant, while understanding the predictor is not an objective. However, in some practical circumstances it is important to have a model that humans can understand [34, 13]. Some classification methods, like decision trees and Boolean formulas learned from data are inherently interpretable, but in many cases they do not perform very well [25, 35]. Another issue is that they have a performance-interpretability trade-off that is hard to control: in order to gain accuracy, an increase in complexity is required, but this in turn will increase the difficulty of exploring the learned model. This shows that it is worth exploring new approaches in this direction. [4]

In the present chapter we develop an algorithm that generates polynomial neural networks for binary classification that display a certain level of interpretability. The interpretability is achieved by restricting our attention to the cases in which the features and the labels can be understood to express the degree of truth of a proposition about the instance that is being classified. With this assumption, the classifiers are interpreted as compound propositions in a loosely defined "real-valued logic". The atomic propositions, each one associated to a feature, are connected by "logical connectives", expressed by simple arithmetic operations.

The proposed models are polynomials over a subset of the real numbers, and the learning process is equivalent to finding the appropriate monomials and coefficients. Therefore, we consider the following hypothesis set (d_{max} is the maximum degree):

[4] The content of the chapter was submitted to be published under the title "Deep interpretable polynomial networks"

Name	Boolean operation	Arithmetic operation(s)
Negation	$\neg x$	$a(1-x)$
Conjunction	$x \wedge y$	axy
Disjunction	$x \vee y$	$ax + by$

Table 1: The correspondence between Boolean logic operators and arithmetic operators used in this chapter (x and y are either Boolean logical values or real values from the interval $[0,1]$, while $a \in (0,1]$ and $b \in (0,1]$ are arbitrary parameters ($a + b \leq 1$ for disjunction)).

$$\begin{aligned}\mathcal{H} = \{h : [0,1]^{2n} \to [0,1] \mid h(x) = \textstyle\sum_{d=1}^{d_{\max}} \sum_{\|\alpha\|_1=d} w_\alpha x^\alpha, \\ \textstyle\sum_{\|\alpha\|_1=d} w_\alpha = w_d, w \geq 0\}.\end{aligned} \tag{7}$$

For some technical reasons, it is more convenient to work with the following slightly transformed set of functions:

$$\mathcal{G} = \{g : [0,1]^{2n} \to [-1,1] | g(x) = 2h(x) - 1, h \in \mathcal{H}\}. \tag{8}$$

The main problem that needs to be solved is that the hypothesis space grows exponentially with the degree of the polynomials. In order to efficiently explore the hypothesis set, we use the method of bounding the generalization error developed in [10]. In principle, the technique consists of finding an upper bound on the error by taking advantage of some special structure of the predictors, and then minimizing this bound. In our case the "special structure" is the fact that a polynomial of degree $d_{\max}$ can be written as a convex combination of other polynomials, each one having only terms of degree exactly d, with d ranging between 0 and $d_{\max}$. Using this approach, we can prove the following learning bound on the true risk ($R[g]$) in terms of the margin empirical risk ($\hat{R}_\rho[g]$):

Theorem 1. Let $\mathcal{G}$ be the class of functions defined by Equation (8), and fix $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample S of size m drawn *i.i.d.* according to some probability distribution P, the following inequality holds for all $g \in \mathcal{G}$:

$$\begin{aligned}R[g] \leq & \hat{R}_\rho[g] + \frac{8}{\rho} \sum_{d=1}^{d_{\max}} w_d \sqrt{\frac{2d}{m} \log \frac{e(2n+d-1)}{d}} + \frac{2}{\rho} \sqrt{\frac{\log(d_{\max})}{m}} + \\ & + \sqrt{\left\{ \frac{4}{\rho^2} \log \left[\frac{\rho^2 m}{\log(d_{max})} \right] \right\} \frac{\log(d_{\max})}{m} + \frac{\log\left(\frac{2}{\delta}\right)}{2m}}.\end{aligned}$$

The bound is the starting point for the algorithm design. In principle, we will like to construct a procedure to minimize the bound. Since this is a relatively difficult task, we relay on some simplifications. We remove some terms that are not very important, replace the empirical risk by a convex surrogate, and introduce a new regularization/calibration constant $\lambda \geq 0$ and the new variables β, obtained by dividing the original weights (w's) by the margin (ρ).

In the end, we get the following optimization problem:

$$\begin{aligned} \min_{\beta} \quad & c(\beta) \\ \text{s.t.} \quad & \beta \geq 0, \end{aligned} \tag{9}$$

where the objective function is

$$c(\beta) = \frac{1}{m} \sum_{i=1}^{m} e^{1-2y_i \sum_{d=0}^{d_{\max}} \sum_{j=1}^{n_d} \beta_{dj} x_{dji}} + \lambda \sum_{d=1}^{d_{\max}} \beta_d \sqrt{\frac{d}{m} \log \frac{e(2n+d-1)}{d}}, \tag{10}$$

which is a convex function in $\beta = (\beta_{01}\ \beta_{11}\ \beta_{12}\ ...\ \beta_{d_{\max} n_{d_{\max}}})$. In the above equation, $\beta_d = \sum_{j=1}^{n_d} \beta_{dj}$, y_i is the label of the data point i (with the value -1 or 1), and x_{dji} is the value of the j-th monomial of degree d calculated with the features of the point i

This is a convex program, therefore for a fixed number of variables it can be solved efficiently. Unfortunately in our case the dimension of β grows exponentially with $d_{\max}$. In order to solve this problem we adopt the strategy of solving progressively bigger programs.

The proposed training algorithm is iterative, and each iteration consists of two phases. In the first phase we try to decide which new monomials to include, by computing some partial derivatives of the current objective function. In the second phase, the coefficients are recomputed by solving a simple convex program. The process continues until no further benefits are obtained by adding new terms, or some resources threshold is reached (see Algorithm 2).

We evaluated the algorithm on many datasets. The most relevant experiments were performed on the COMPAS dataset [33], which contains the criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County from 2013 and 2014. The performance of the algorithm on this dataset was comparable, but slightly worse than that of a new interpretable neural network, Neural Additive Models (NAM) [4]. On the other hand, NAM has a much larger number of parameters and hyperparameters (see Table 2). On the same dataset, our algorithm outperformed Learning Certifiably Optimal Rule Lists (CORELS) [5], a recently proposed rules-based machine learning method (see Table 3).

In conclusion, we presented a new algorithm for binary classification that can retain some of the features of deep neural networks, such as expressiveness and scalability, while generating models that can be understood. Our theoretical and empirical studies indicate that the algorithm has good generalization capabilities, and in many practical situations provides interpretable models.

3.1 Learning bounds for quantum circuits in the agnostic setting

Quantum circuit learning consist of learning a function from a class defined based on a family of circuits with finite resources (number of qubits, number of 2-qubits gates, etc.). In this chapter, the class of functions investigated is composed of the output probabilities associated to a fixed computational basis state. This is a natural choice for doing classification or regression on quantum data. [5]

[5] The chapter is based on the paper [30]

Algorithm 2: DIPNN training algorithm

input : *Data:* S; *Parameters:*λ,d_{max},b

d = 1;
%Initialize the weights vector; β^*_{1i} *are variables initialised*
%with 0
$\beta^{(1)} = (\beta^*_{01}, \beta^*_{11}, \beta^*_{12}, ..., \beta^*_{1(2n)}, 0, ..., 0)$;
$new\_monomials = True$;
while $d \leq d_{\max}$ *AND* $new\_monomials$ **do**
*%***Phase1**
%Update $\beta^{(d)}$ *by solving the program*
%with the current list of variables:

$$\beta^{(d)} = \min_{\beta^{(d)}} \quad c(\beta^{(d)})$$
$$\text{s.t.} \quad \beta^{(d)} \geq 0$$

*%***Phase2**
d++;
$new\_monomials = False$;
%Update the list of variables:
for $k = 1, ..., n_d$ **do**

$$\frac{\partial c(\beta^{(d)})}{\partial \beta_{dk}} = -\frac{2}{m}\sum_{i=1}^{m} y_i x_{dki} e^{1-2y_i \sum_{d=0}^{d_{\max}} \sum_{j=1}^{n_d} \beta_{dj} x_{dji}} + \lambda\sqrt{\frac{d}{m}\log\frac{e(2n+d-1)}{d}};$$

if $\frac{\partial c(\beta^{(d)})}{\partial \beta_{dk}} < 0$ **then**

$$\beta^{(d)}_{d,k} = \beta^*_{d,k};$$

$$new\_monomials = True;$$

end
else

$$\beta^{(d)}_{d,k} = 0;$$

end
end
end
output: $\beta^{(d)}$

Neural network	AUC	Standard deviation of AUC	Running time (s)	No. of learned parameters	No. of hyperparameters
Single Task NAM	0.737	0.010	65	>100000	8
Multitask NAM	**0.739**	0.010	>65	>100000	8
DIPNN ($d_{max} = 1$)	0.727	0.013	5	26	3
DIPNN ($d_{max} = 2$)	0.732	0.010	40	<76	3
DIPNN ($d_{max} = 3$)	0.732	0.010	91	<126	3
DIPNN ($d_{max} = 4$)	0.731	0.010	138	<176	3
DIPNN ($d_{max} = 5$)	0.731	0.010	188	<226	3
DIPNN ($d_{max} = 6$)	0.731	0.010	261	<276	3

Table 2: Comparison between NAM and DIPNN.

The learning bounds are quite simple and easy to interpret. They are expressed in terms of basic quantum circuits parameters, namely, the number of qubits and the number of 2-qubits gates. The dependence on these parameters is polynomial, which implies efficient statistical learnability [35].

The strategy for deriving the bounds is straightforward. The covering number is bounded using tools from quantum computing (e.g. Solovay–Kitaev Theorem [19]), then the result is used to bound the Rademacher complexity. The basic set of functions is then used to define hypothesis sets for regression and classification on quantum data. For some general loss functions, bounds on the generalisation error are obtained in both cases using the Rademacher complexity.

We assume that the data points are drawn independently from a fixed, but unknown probability distribution. In other words, an *i.i.d.* ("independent and identically distributed") dataset

$$S = \{(|\phi_i\rangle, y_i) | \, |\phi_i\rangle \in \mathbb{C}^d, d \in \mathbb{N}^*, y_i \in [0,1], i \in [n]\} \tag{11}$$

is available. Each input vector describes a pure state on q qubits ($q \in \mathbb{N}^*$, $d = 2^q$), while the output is a positive unitary or subunitary real number.

Let $\gamma \in \mathbb{N}^*$ be some fixed positive integer. We will assume that an input state undergoes a transformation described by a quantum circuit acting on q qubits that can be implemented using γ local 1-qubit and 2-qubit gates. Here by "local" we mean that the 2-qubit gates act on consecutive qubits. The parameter γ will be called the circuit size.

The restriction to local gates will simplify the analysis. Also, many quantum computing algorithms are based on such gates [27]. However, at the price of a polynomial factor, any non-local 2-qubits gate can be implemented using local gates.

We also assume that the gates are not applied in parallel, that on each layer of the circuit only one gate is present. Notice that this assumption is merely formal, it does not restrict the circuit class. Indeed, if in the layer i, we have k gates, we can design an equivalent

Algorithm and hyperparameters	Accuracy	Standard deviation of accuracy	Running time (s)
CORELS ($\lambda = 0.005, mc = 1$)	0.654	0.009	9.37
CORELS ($\lambda = 0.01, mc = 1$)	0.660	0.016	0.08
CORELS ($\lambda = 0.005, mc = 2$)	0.660	0.009	171
CORELS ($\lambda = 0.01, mc = 2$)	0.656	0.011	197
CORELS ($\lambda = 0.005, mc = 3$)	0.664	0.014	150
CORELS ($\lambda = 0.01, mc = 3$)	0.664	0.017	162
CORELS ($\lambda = 0.05, mc = 4$)	0.664	0.017	208
CORELS ($\lambda = 0.01, mc = 4$)	0.664	0.017	214
DIPNN ($d_{max} = 1$)	0.673	0.008	8
DIPNN ($d_{max} = 2$)	**0.676**	0.010	95
DIPNN ($d_{max} = 3$)	0.674	0.011	139
DIPNN ($d_{max} = 4$)	0.675	0.011	192
DIPNN ($d_{max} = 5$)	0.675	0.010	244
DIPNN ($d_{max} = 6$)	0.674	0.010	552

Table 3: Comparison between CORELS and DIPNN.

circuit with $k-1$ additional layers such that one gate remain on layer i and the other ones are moved to the next layers, $i+1, i+2, ..., i+k-1$ (the order is irrelevant) [27]. With this assumption, the depth of the circuit will also be equal to γ.

Since any gate acting on a single qubit can be substituted trivially by one acting on two qubits, we will consider for simplicity that all gates are 2-qubits gates (in our analysis having only "true" 2-qubits gates is the worst case).

The family of quantum circuits that we have just described will be parameterised by d and γ, and we will call it $\mathcal{C}_{d,\gamma}$. For an example of quantum circuit from this set, please see Figure 4.

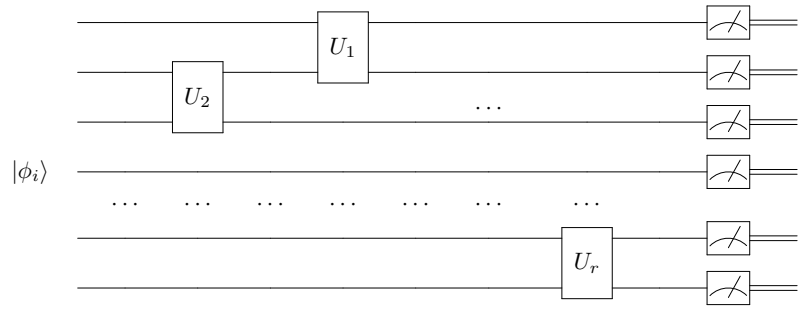

Figure 4: Example of quantum circuit from $\mathcal{C}_{d,\gamma}$.

The effect of the quantum circuit on the input will be described by the following family of unitary transformations:

$$\mathcal{U}_{d,\gamma} = \{U \in \mathbb{C}^{d\times d} | U^\dagger U = UU^\dagger = I_d, \\ \textit{U is implemented by a circuit from } \mathcal{C}_{d,\gamma}\}. \tag{12}$$

To a class $\mathcal{U}_{d,\gamma}$ we can assign the following set of real valued functions (which will also serve as a hypothesis set for regression [35]):

$$\mathcal{H}_{d,\gamma} = \{h : \mathbb{C}^d \to [0,1] | h(|\phi\rangle) = |\langle 0|^{\otimes q} U |\phi\rangle|^2, U \in \mathcal{U}_{d,\gamma}\}. \tag{13}$$

In other words, each function give us the probability of observing the all-zeros output $(00..0)$, when a measurement in the computational basis is performed on the output of a quantum circuit from the set $\mathcal{C}_{d,\gamma}$. This is arguably the simplest way to define a real function starting from a quantum circuit. Of course, there is nothing special about the basis element $|0\rangle^{\otimes q}$, any other element could be used.

The hypothesis class defined by Equation (13) can be used for regression on quantum data. In standard regression, the output takes values in an arbitrary interval of $\mathbb{R}$ [25], but since any interval $[a,b] \subset \mathbb{R}$ can be mapped by a bijection to $[0,1]$, no loss of generality occurs.

For binary classification a small modification is needed. To each function $h \in \mathcal{H}_{d,\gamma}$, a function $g_h : \mathbb{C}^d \to \{-1,1\}$, defined by

$$g_h(|\phi\rangle) = \begin{cases} 1, & \text{if } h(|\phi\rangle) - \frac{1}{2} \geq 0; \\ -1, & \text{otherwise,} \end{cases} \tag{14}$$

is assigned. The set of all functions g_h will be called $\mathcal{G}_{d,\gamma}$.

Let as also observe that in the case of binary classification the sample will have labels from the set $\{-1,1\}$.

To define the risk we need to use a loss function. For regression, a widely used loss function is the L_p loss ($p \in \mathbb{N}^*$), and the induced risk functionals are [25]:

$$R_p[h] = \mathop{\mathbb{E}}_{x\sim D}[|h(x) - f(x)|^p],$$
$$\hat{R}_p[h] = \frac{1}{n}\sum_{i=1}^{n} |h(x_i) - f(x_i)|^p,$$

for some fixed function f, a hypothesis h, a probability distribution D, and a sample $\{x_1, x_2, ..., x_n\}$.

We can define the risks for binary classification in a similar manner. Let us introduce a cost function $\psi : \mathbb{R} \to \mathbb{R}_+$ such that $\psi(x) \geq \mathbf{1}_{x>0}$ ($\mathbf{1}_{P(x)}$ is the indicator function of the predicate $P(x)$: for some x, it takes value 1 if $P(x)$ is true, and 0 otherwise). Some popular choices are $\psi(x) = e^x$, $\psi(x) = log_2(1+e^x)$, and $\psi(x) = (1+x)_+$ [6]. Using this function we can define a surrogate empirical risk in the following way:

$$\hat{R}_\psi[h] = \frac{1}{n}\sum_{i=1}^{n} \psi(-y_i h(x_i)). \tag{15}$$

The true risk for classification is defined as the probability of predicting the wrong label, that is:

$$R_c[h] = P(h(x) \neq y). \tag{16}$$

Now we can state the main results of this chapter. For regression, we can prove the following theorem:

Theorem 2. Let $\mathcal{H}_{d,\gamma}$ be the hypothesis set defined in Equation (13), and $p \geq 1$ an integer. Then, for all $h \in \mathcal{H}_{d,\gamma}$, and any $\delta > 0$, with probability at least $1 - \delta$ over a sample S of size n, the following bound on the generalization error under the L_p loss is true:

$$R_p[h] - \hat{R}_p[h] = O(\sqrt{\frac{\gamma log^c(\gamma) log(q)}{n}} + \sqrt{\frac{log\frac{1}{\delta}}{n}}). \tag{17}$$

In the case of binary classification, we get the following bound for the generalisation error:

Theorem 3. Let $\psi : \mathbb{R} \to \mathbb{R}_+$ be a uniformly bounded and Lipschitz continuous cost function such that $\psi(x) \geq \mathbf{1}_{x>0}$, and S an *i.i.d.* sample. Then, with probability at least $1 - \delta$, the following bound is true for all $g \in \mathcal{G}_{d,\gamma}$:

$$R_c[g] - \hat{R}_\psi[g] = O(\sqrt{\frac{\gamma log^c(\gamma) log(q)}{n}} + \sqrt{\frac{log\frac{1}{\delta}}{n}}). \tag{18}$$

In both cases, the constant c can be taken to be 1.

The proofs are based on Rademacher complexity [21, 20, 35], covering number [35], [19] and the Solovay–Kitaev theorem [17]. They relay on upper bounds on the Rademacher complexity in terms of the logarithm of the covering number. The logarithm of the covering number is upper bounded in two broad steps. First, we reduce to a finite family of quantum circuits that approximate well enough, in a certain precise sense, the initial set of quantum circuits. Solovay–Kitaev theorem is a key component in this step. Second, we upper bound the size of this set using elementary combinatorics.

The main conclusion of the chapter is the learnability of quantum circuits in the agnostic case. As long as the circuit that needs to be learned can be represented by an equivalent circuit with a finite number of gates on 2 qubits, the true risk will be close to the empirical risk with high probability, for a sample of polynomial size (in the number of gates and qubits). In Table 4, a succinct comparison with some similar results is provided.

Source	Bound	Setting	Input	Output
[9]	$O(\frac{log\|\mathcal{H}\|+log(\frac{1}{\delta})}{\epsilon''^2})$	Finite class, Realizable	Quantum basis states	Pure states
[9]	$O(\frac{log^3\|\mathcal{H}\|(log\|\mathcal{H}\|+log(\frac{1}{\delta}))}{\epsilon''^2})$	Finite class, Realizable	Quantum basis states	Mixed states
[8]	$O(\frac{\Delta\gamma^2 log\gamma \log^2(\frac{\Delta\gamma^2 log\gamma}{(\beta-\alpha)\epsilon'})+log(\frac{1}{\delta})}{\epsilon'})$	Finite/infinite class, Realizable	Quantum states and/or measurement computational basis state	Outcome probability
This thesis	$O(\frac{q\gamma \log(q\gamma) log(q)+log(\frac{1}{\delta})}{\epsilon^2})$	Finite/infinite class, Agnostic	Quantum states	Outcome probability

Table 4: Sample complexity for quantum circuit learning. ([30]Copyright ©2021 Springer)

4 Conclusions

In this work we made a number of theoretical, algorithmic and empirical contributions to the fields of natural language processing and machine learning. The most significant contributions are: a new convex optimization-based extractive summarization algorithm, a geometrically -motivated version of the k-NN algorithm, a new interpretable polynomial neural network, along with an algorithm to train it, and finally, a learning bound for quantum circuits in the agnostic case.

In Chapter 2, we derived a new algorithm for extractive text summarization starting from some basic properties that a good summary should have. The main ingredients were the convex relaxation procedure and the projected gradient descent method for constructing optimization algorithms. Beside scalability, which was the main focus, the algorithm has some other advantages, like its versatility, and the possibility to use side information and additional constraints. Probably the main message of the chapter is that convex optimization, coupled with powerful numerical text representations, can be used to design very fast and high quality text summarization algorithms.

Chapter 3 contains a detailed description of an algorithm which makes use of conformal transformations in order to generate a better distance function for the k-NN classification method. The main achievement of the chapter was the extension of the research on distance learning in a new nonparametric direction. The algorithm was tested on some challenging datasets, with some positive outcomes. We hope the ideas from this chapter can serve as a starting point for further research in this fascinating sub-field of machine learning.

The next chapter (Chapter 4) describes a key contribution of the thesis, namely an interpretable polynomial neural network. Deep neural networks have become some of the most powerful machine learning methods, but they often lack interpretability. Therefore, any effort to address this shortcoming is valuable. Our approach brings forth in this area a number of new ideas, inspired or borrowed from theoretical breakthroughs (tight learning bounds for convex ensembles) and different sub-fields of artificial intelligence (e.g. real-valued logic-based methods). We were able to derive a new learning bound for some interesting classes of polynomial functions, and use it to design a new learning algorithm. The experiments that we performed on many datasets show that the new algorithm behaves very well, outperforming or being close to many classic learning algorithms, and also state-of-the-art interpretable models. The method can be extended and improved in many ways.

In Chapter 5 we addressed a theoretical problem from the very new research area of quantum circuit learning, namely the learnability of quantum circuits when the realizability assumption does not hold. Using tools from quantum computing and learning theory, we were able to derive meaningful bounds on the excess risk for binary classification and regression. Such results may have an impact on the design of machine learning algorithms for quantum data. We compare the bounds with those published in the previous papers, and we conclude that for the cases in which a comparison can be done, the new bounds are tighter, being of type $\gamma log(\gamma)$ [6] (but we must underline the fact that they were derived in different settings).

Overall, the thesis shows the potential of old and new ideas and methods from convex

[6] γ is the number of gates with a fixed number of inputs.

optimization and learning theory to constitute the foundation for principled algorithms design and analysis, in order to address exciting problems in natural language processing and data-driven prediction.

5 List of Publications

P1 M.C. Popescu, "Learning bounds for quantum circuits in the agnostic setting," *Quantum Information Processing*, vol. 20, no. 9, 2021. DOI:10.1007/s11128-021-03225-7, WOS:000692406000004, Q1.

P2 M.C. Popescu, L. Grama, and C. Rusu, "A highly scalable method for extractive text summarization using convex optimization," *Symmetry*, vol. 13, no. 10, 2021. DOI:10.3390/sym1310 WOS: 000717213500001, Q2.

P3 M.C. Popescu, C. Rusu, and L. Grama, "Word embeddings for romanian language and their use for synonyms detection," in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2021, pp. 151–155. DOI: 10.1109/SpeD53181.2021.9587432, WOS: 000786794700027

P4 M.C. Popescu, L. Grama, and C. Rusu, "Conformal transformation of the metric for k-nearest neighbors classification," in *2020 IEEE 16th Int. Conf. on Intelligent Computer Communication and Processing (ICCP)*, 2020, pp. 229–234. DOI: 10.1109/ICCP51029.2020.926624 WOS: 000646618600029; © 2020 IEEE

P5 M.C. Popescu, L. Grama, and C. Rusu, "On the use of positive definite symmetric kernels for summary extraction," in *2020 13th International Conference on Communications (COMM)*, 2020, pp. 335–340. DOI: 10.1109/COMM48946.2020.9142041, WOS: 00061272390005

P6 M.C. Popescu, L. Grama, and C. Rusu, "Automatic text summarization by mean-absolute constrained convex optimization," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, July 4-6 2018, pp. 1–5. DOI: 10.1109/TSP.2018.8441416, WOS: 000454845100158

P7 M.C. Popescu, L. Grama, C. Rusu and M. Sirbu, "Communication protocol for wireless sensor networks for energy consumption optimization," in *Proc. WCSIT 2014*, July 3-4, 2014, Iasi, Romania, ISBN: 978-4-88552-286-4 C3855 (IEICE).

References

[1] News category dataset. https://www.kaggle.com/rmisra/news-category-dataset. Accessed: 2020-02-01.

[2] Large movie review dataset v1.0. https://ai.stanford.edu/~amaas/data/sentiment/. Accessed: 2019-01-30.

[3] A. H. A. Bellet and M. Sebban. A survey on metric learning for feature vectors and structured data. Technical report, University of Saint-Etienne, 2014.

[4] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets, 2020.

[5] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(1): 8753–8830, Jan. 2017. ISSN 1532-4435.

[6] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[7] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008. doi: 10.1109/MSP.2007.914731.

[8] M. C. Caro and I. Datta. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence*, 2(2):1–14, 2020.

[9] K.-M. Chung and H.-H. Lin. Sample efficient algorithms for learning quantum channels in pac model and the approximate state discrimination problem. *arXiv preprint arXiv:1810.10938*, 2018.

[10] C. Cortes, M. Mohri, and U. Syed. Deep boosting. In *Proceedings of the Thirty-First International Conference on Machine Learning (ICML 2014)*, 2014.

[11] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[12] DUC.2005. Document understanding conference 2005. http://www-nlpir.nist.gov/projects/duc/, 2005.

[13] H. J. Escalante, S. Escalera, I. Guyon, X. Baro, Y. Gucluturk, U. Guclu, and M. van Gerven. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer Publishing Company, Incorporated, 1st edition, 2018. ISBN 3319981307.

[14] A. L. M. et al. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT 2011, pages 142–150. Association for Computational Linguistics, 2011. ISBN 978-1-932432-87-9.

[15] E. Fix and J. L. Hodges. Discriminatory analysis- nonparametric discrimination: Consistency properties(technical report 4, project no. 21-29-004). Technical report, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[16] M. Grusky, M. Naaman, and Y. Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065.

[17] A. W. Harrow, B. Recht, and I. L. Chuang. Efficient discrete approximations of quantum gates. *Journal of Mathematical Physics*, 43(9):4445–4451, 2002.

[18] N. C. J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521813972.

[19] A. Y. Kitaev. Quantum computations: algorithms and error correction. *Russian Mathematical Surveys*, 52(6):1191–1249, Dec. 1997. doi: 10.1070/rm1997v052n06abeh002155. URL https://doi.org/10.1070/rm1997v052n06abeh002155.

[20] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[21] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

[22] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

[23] M. Litvak and N. Vanetik. Query-based summarization using MDL principle. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, page 22–31, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1004.

[24] G. A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38: 39–41, 1995.

[25] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[26] A. Nenkova and K. McKeown. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2-3):103–233, 2011. doi: 10.1561/1500000015.

[27] M. A. Nielsen and I. Chuang. Quantum computation and quantum information, 2010.

[28] C. Popescu, L. Grama, and C. Rusu. Automatic text summarization by mean-absolute constrained convex optimization. In *Proceedings of 41st Int. Conf. on Telecommunications and Signal Processing*, pages 706–709, Athens, Greece, July 2018. IEEE. doi: 10.1109/TSP.2018.8441416.

[29] C. Popescu, L. Grama, and C. Rusu. A highly scalable method for extractive text summarization using convex optimization. *Symmetry*, 13(10):1824, 2021.

[30] C. M. Popescu. Learning bounds for quantum circuits in the agnostic setting. *Quantum Information Processing*, 20(9):1–24, 2021.

[31] M. C. Popescu, L. Grama, and C. Rusu. On the use of positive definite symmetric kernels for summary extraction. In *2020 13th International Conference on Communications (COMM)*, pages 335–340, 2020. doi: 10.1109/COMM48946.2020.9142041.

[32] M. C. Popescu, L. Grama, and C. Rusu. Conformal transformation of the metric for k-nearest neighbors classification. In *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 229–234. IEEE, 2020.

[33] ProPublica. Compas data and analysis for 'machine bias', 2016. URL https://github.com/fchollet/keras.

[34] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, May 2019.

[35] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[36] R. L. R. Thomas H. Cormen, Charles E. Leiserson and C. Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, 3rd edition edition, 2009. ISBN 978-0262033848.

[37] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. URL http://www.jstor.org/stable/2346178.