UNIVERSIDADE DE AVEIRO
DEPARTAMENTO DE ELECTRÓNICA TELECOMUNICAÇÕES E INFORMÀTICA

ML - PROJECT 1 2020/2021

Each group of two students is supposed to work on one project topic.
You are strongly encouraged to propose a machine learning problem you would prefer to work, not listed below, that may reflect better your interests. Please, discuss your idea with the instructor.

I. PROJECT GOALS

The goal of this project is to apply suitable machine learning algorithms learned in class or self-learned to solve a specific data science problem (classification or regression). Represent the results in graphical/table formats and make analysis and conclusions.

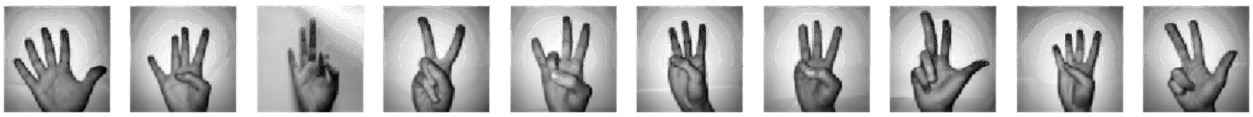
II. PROJECT PROPOSALS

Sign language understanding

Hand gestures and sign language are the most commonly used methods by deaf and non-speaking people to communicate among themselves or with speech-able people. However, understanding sign language is not a universal skill. For this reason, building a system that recognizes hand gestures and sign language can be very useful to facilitate the communication gap between speech-able and speech-impaired people.

Project proposal 1: Identification of digits from sign language images

Data source: https://www.kaggle.com/ardamavi/sign-language-digits-dataset



Project proposal 2: American sign language understanding

Data source: https://www.kaggle.com/datamunge/sign-language-mnist/

Project proposal 3: Mammographic Mass Data Set
This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The aim is to discriminate benign from malignant cases assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases are benign.

Data source: http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Project proposal 4: Heart Disease Data Set
This dataset contains 4 heart disease related datasets. For the present project you will use the Cleveland database and the referred subset of 14 features form a total of 76 attributes. The goal is to distinguish presence (values 1,2,3,4) from absence (value 0) of heart disease in the patient.

Data source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Project proposal 5: Bank Marketing Data Set
The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe or not a term deposit.

Data source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Project proposal 6: The German Traffic Sign Benchmark (GTSB)
GTSB is a multi-class, single-image classification challengeheld at IJCNN 2011. Automatic recognition of traffic signs is required in advanced driver assistance systems and constitutes a challenging real-world computer vision and pattern recognition problem. A comprehensive, lifelike dataset of more than 50,000 traffic sign images has been collected. It reflects the strong variations in visual appearance of signs due to distance, illumination, weather conditions, partial occlusions, and rotations. The dataset comprises 43 classes with unbalanced class frequencies.

Data source: https://www.kaggle.com/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign

Machine Learning in optical telecommunications
Data will be provided by the instructor. See more details in the Appendix.

Project proposal 7 (64-QAM classification): Classification of the 64 QAM transmitted symbols from the noisy signals obtained at the receiver with NN classification models.

Project proposal 8 (16-QAM classification). Classification of the 16-QAM transmitted symbols from the noisy signals obtained at the receiver. Comparison between different classifiers (e.g. NN, SVM, Logistic regression).

Project proposal 9 (64-QAM regression): Recover the 64 QAM transmitted symbols from the noisy signals obtained at the receiver with NN regression models.

Project proposal 10 (16-QAM regression): Recover the 16-QAM transmitted symbols from the noisy signals obtained at the receiver with NN regression models.

III. PROJECT ASSESMENT (25 % of the final grade)

1. **Report.** The project is evaluated based on a submitted report (IEEE Latex format). The work done by each student has to be explicitly specified. All project's files (pdf and Latex files of the report, the presentation slides and the code implementing the algorithms) are sent to the course instructor (petia@ua.pt) in a compressed format having the following name: P1_ML2019_XXXXX_YYYYY (where XXXXX and YYYYY are substituted by the academic (mechanographic) number of each student. If the file is too big to email as an attached document, feel free to use any big file transfer option you may know (we transfer, dropbox, link in a cloud. etc.)

2. **Oral presentation** of the report in class (about 10-15 min.).

IV. Evaluation criteria (total score 20)

1. Report content (10):

- Data description and preprocessing (if necessary normalization, feature selection, transformation, etc.). Motivation for choosing the particular problem.
- Data visualization (histograms, box plots, other plots).
- Short description of the implemented ML models.
- Model training (data splitting – train, validate, test, k-fold Cross validation). Visualize graphically the cost function trajectory over iterations. Training with regularized and non-regularized cost function.
- Model hyper-parameter selection - regularization parameter *lambda*, number of NN hidden layer units, number of hidden layers (if necessary), *sigma*, *C, k*, etc.. Systematic approach instead of just one or several randomly chosen values (see slide 27 in lecture 5).
- For a classification problem, you need to present the confusion Matrix (accuracy, precision, recall, F1 score).
- Performance comparison between the models.
- Results in graphical or table formats.
- Conclusions.

2. Report formatting (3) :

- IEEE Latex format, affiliation (Department, University, subject, course instructor), abstract, keywords, work load per student.
- Sufficiently detailed report.
- References, reference citation in the report.
- Clear figures (title, legends, axis labels) and tables referred in the text.

3. Oral presentation (4)

- Slide Organization, slide numbers, affiliation.
- Clear and convincing presentation by both students.

4. Novelty and contributions (3)

- Based on the references and what has been done previously by other authors, propose a better solution, e.g. improve the performance of the ML model in solving the problem you work with.

Deadline for project 1 submission (report + presentation slides): 6/May, 2021.

Project presentation: 7/May, 2021, in class.

Appendix: Machine Learning in optical telecommunications

Equalization of Fiber Optic Channels - decode Quadrature Amplitude Modulated (QAM) signals transported over an optic link.

1. Introduction

Quadrature Amplitude Modulation (**QAM**) is the name of a family of digital modulation methods widely used in modern telecommunications to transmit information. Like all modulation schemes, QAM conveys data by changing some aspect of a carrier signal, or carrier wave, (usually a sinusoid) in response to a data signal. In the case of QAM, the carrier wave is the sum of two sinusoidal waves of the same frequency, 90° out of phase with each other (in quadrature). These are often called the "I" or in-phase component, and the "Q" or quadrature component. Each component wave is amplitude modulated, i.e. its amplitude is varied to represent the data to be carried, before the two are added together. Amplitude modulating two carriers in quadrature can be equivalently viewed as both amplitude modulating and phase modulating a single carrier.

A **constellation diagram** is a representation of a signal modulated by a digital modulation scheme such as QAM. It displays the signal as a two-dimensional xy-plane scatter diagram in the complex plane. The number of constellation points in a diagram gives the size of the "alphabet" of symbols that can be transmitted by each sample, and so determines the number of bits transmitted per sample. It is usually a power of 2. A diagram with 16 points (see Fig.1), for example, represents a modulation scheme that can separately encode all 16 combinations of four bits and therefore can transmit 4 bits per sample.
After passing through the communication channel the signal is decoded by a demodulator. The function of the demodulator is to classify each sample as a symbol. The performance of the fiber optic communication channels is often limited by a phenomenon known as dispersion, which causes optical pulses to broaden as they propagate through the fiber, thus giving rise to inter-symbol interference (ISI). Migration towards greater speed and longer links in fiber optic systems augment problems with dispersion in optical fibers, and dispersion compensation is consequently an increasingly important issue. Due to the ISI (see Fig.2), the demodulator may misidentify that sample as other symbol, resulting in a symbol error.

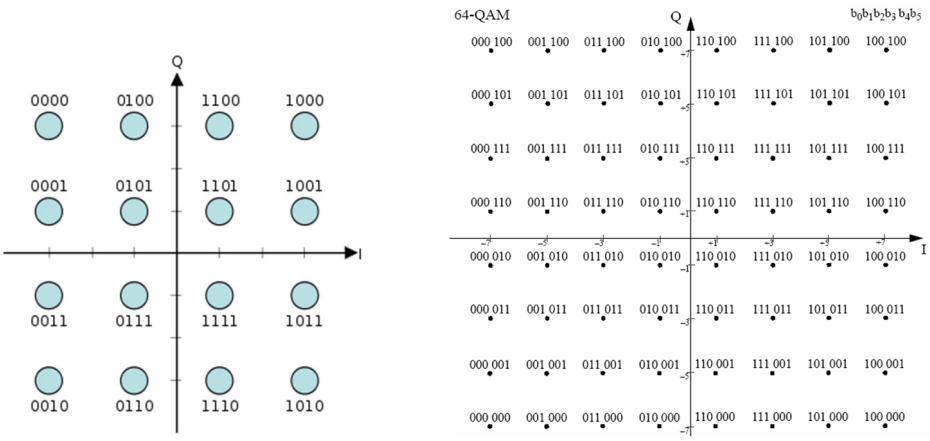


Fig.1 Constellation diagram for rectangular 16-QAM (left) and 64-QAM (right).

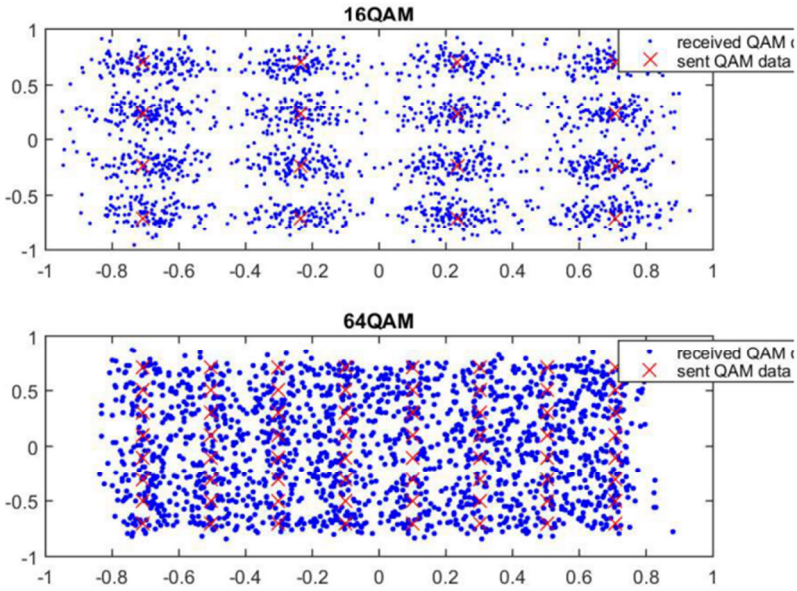


Fig.2. Constellation diagram for rectangular 16-QAM (upper) and 64-QAM (lower) transmitted (red crosses) and received (blue dots) symbols. Real data.

2. Work Assignment

The goal of this set of projects is to apply machine learning methods to compensate the inter-symbol interference and correctly decode the symbols that have being transmitted.

2.1 Classification

The first possible approach is to consider the channel equalization as a classification problem (64 or 16 classes) and build a reliable classifier. Neural Network (ANN) classifier (see Fig.3), seems to be a promising candidate for this task.

Two proposed projects:

Project proposal 1 (64-QAM classification): Classification of the 64 QAM transmitted symbols from the noisy signals obtained at the receiver with NN classification models.

Project proposal 2 (16-QAM classification). Classification of the 16-QAM transmitted symbols from the noisy signals obtained at the receiver. Comparison between different classifiers (e.g. NN, SVM, Logistic regression).

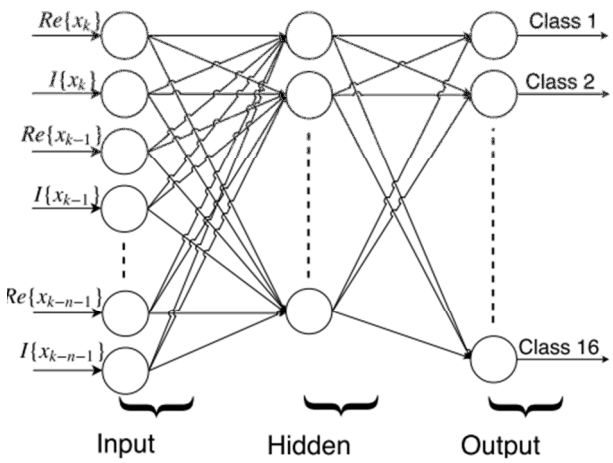


Fig. 3 NN for classification

The second possible approach is to consider the channel equalization as a regression problem (the inputs are past received samples, the output is the original transmitted sample) and apply a NN as a regression model (see Fig.4).

Two proposed projects:

Project proposal 3 (64-QAM regression): Recover the 64 QAM transmitted symbols from the noisy signals obtained at the receiver with NN regression models.

Project proposal 4 (16-QAM regression): Recover the 16-QAM transmitted symbols from the noisy signals obtained at the receiver with NN regression models.

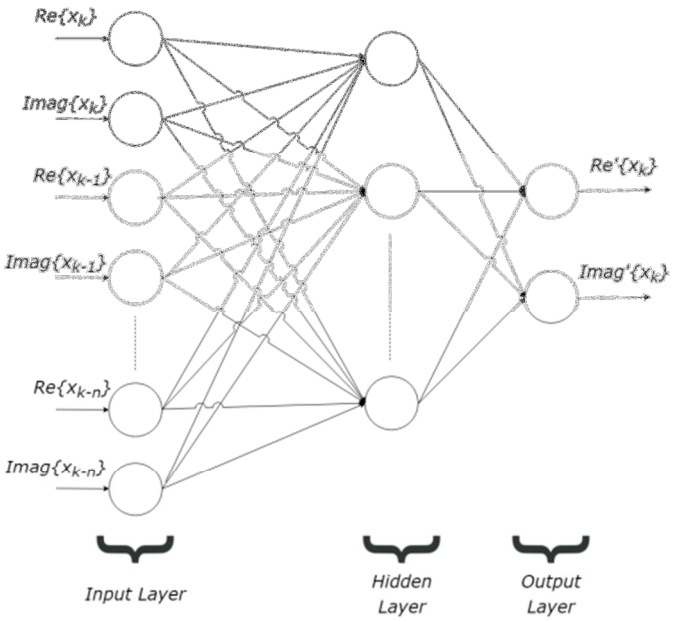


Fig. 4 NN for regression