Project 1 : Classification of particles

Leonardo Aoun, Romain Artru, Thiercelin Marin
IC Faculty, EPFL, Switzerland

***Abstract*—In this report we will explain how we used the methods seen in class to do classification on a given dataset.**

I. Analysing and cleaning the dataset

At first, we looked at the data and made several visualizations. We realized that the data needed to be cleaned before being used for classification. A lot of data points had features that were unknown (with value -999), each features were represented in very different scales and most features presented outliers.

A. Treating missing values

We tried several approaches with respects of the missing values. We deleted the features that had too many -999, with a threshold of 60% of the train data set. We also deleted the training data points that had more than 20% of unknown values.
With the remaining -999s, we tried different methods: setting them to the mean or the median. For the training data, we computed two means/medians depending on the label, whereas for the test data, we computed a single mean/median.

B. Avoiding outliers

For all features in the train data, we saw that few points were having really strange values, and this points could pose problems to the convergence of the learning algorithm. We first classified the data points as outliers, for feature i if:

$$X_i \in [Median - 1.5 * IQR,\ Median + 1.5 * IQR] \quad (1)$$

where IQR is the interquatile range. Once we marked the points as outliers we considered two approaches: deleting the outliers or setting them to different values. The first approach revealed to be costly as a great part of the data points was an outlier for at least one feature. For this reason we decided to set the outlier to a different value, we chose to set them to the median as it is more resistant to the outliers.

C. Normalizing

The features had very different scales, so we decided to normalize them in order to better compare them. We used the formula :

$$X' = \frac{(X - \overline{X})}{\sigma(X)} \quad (2)$$

For the test data, we choosed to normalize using the mean and standard deviation of the train data, to avoid having difference in the data treatment and having models that do not fit.

Figure 1. Signal compression and denoising using the Fourier basis.

D. Divising the set in 4

As one of the features was integer valued with value in $\{0,1,2,3\}$ we decided to divide the data in 4 based on this value and train 4 different models. When dividing, some features were equal and some had variance 0 in some subsets, allowing to further reduce the number of features.

II. Feature extraction

III. Tips for Good Writing

The ideas for good writing have come from [1], [2], [3].

A. Getting Help

One should try to get a draft read by as many friendly people as possible. And remember to treat your test readers with respect. If they are unable to understand something in your paper, then it is highly likely that your reviewers will not understand it either. Therefore, do not be defensive about the criticisms you get, but use it as an opportunity to improve the paper. Before your submit your friends to the pain of reading your draft, please *use a spell checker*.

B. Abstract

The abstract should really be written last, along with the title of the paper. The four points that should be covered [2]:

1) State the problem.
2) Say why it is an interesting problem.
3) Say what your solution achieves.
4) Say what follows from your solution.

C. Figures and Tables

Use examples and illustrations to clarify ideas and results. For example, by comparing Figure 1 and Figure 2, we can see the two different situations where Fourier and wavelet basis perform well.

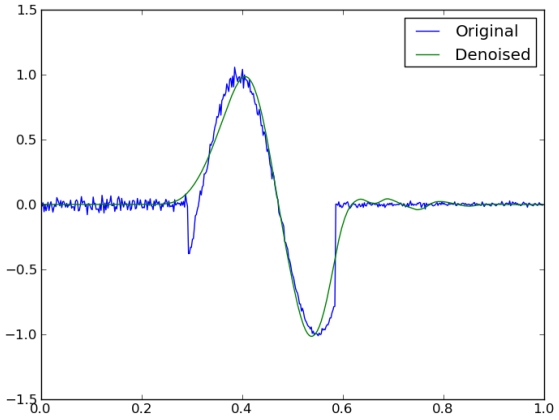

Figure 2. Signal compression and denoising using the Daubechies wavelet basis.

D. Models and Methods

The models and methods section should describe what was done to answer the research question, describe how it was done, justify the experimental design, and explain how the results were analyzed.

The model refers to the underlying mathematical model or structure which you use to describe your problem, or that your solution is based on. The methods on the other hand, are the algorithms used to solve the problem. In some cases, the suggested method directly solves the problem, without having it stated in terms of an underlying model. Generally though it is a better practice to have the model figured out and stated clearly, rather than presenting a method without specifying the model. In this case, the method can be more easily evaluated in the task of fitting the given data to the underlying model.

The methods part of this section, is not a step-by-step, directive, protocol as you might see in your lab manual, but detailed enough such that an interested reader can reproduce your work [3], [4].

The methods section of a research paper provides the information by which a study's validity is judged. Therefore, it requires a clear and precise description of how an experiment was done, and the rationale for why specific experimental procedures were chosen. It is usually helpful to structure the methods section by [5]:

1) Layout the model you used to describe the problem or the solution.
2) Describing the algorithms used in the study, briefly including details such as hyperparameter values (e.g. thresholds), and preprocessing steps (e.g. normalizing the data to have mean value of zero).
3) Explaining how the materials were prepared, for example the images used and their resolution.
4) Describing the research protocol, for example which examples were used for estimating the parameters (training) and which were used for computing performance.
5) Explaining how measurements were made and what calculations were performed. Do not reproduce the full source code in the paper, but explain the key steps.

E. Results

Organize the results section based on the sequence of table and figures you include. Prepare the tables and figures as soon as all the data are analyzed and arrange them in the sequence that best presents your findings in a logical way. A good strategy is to note, on a draft of each table or figure, the one or two key results you want to address in the text portion of the results. The information from the figures is summarized in Table I.

When reporting computational or measurement results, always report the mean (average value) along with a measure of variability (standard deviation(s) or standard error of the mean).

IV. Tips for Good Software

There is a lot of literature (for example [6] and [7]) on how to write software. It is not the intention of this section to replace software engineering courses. However, in the interests of reproducible research [8], there are a few guidelines to make your reader happy:

- Have a `README` file that (at least) describes what your software does, and which commands to run to obtain results. Also mention anything special that needs to be set up, such as toolboxes[1].
- A list of authors and contributors can be included in a file called `AUTHORS`, acknowledging any help that you may have obtained. For small projects, this information is often also included in the `README`.
- Use meaningful filenames, and not `temp1.py`, `temp2.py`.
- Document your code. Each file should at least have a short description about its reason for existence. Non obvious steps in the code should be commented. Functions arguments and return values should be described.
- Describe how the results presented in your paper can be reproduced.

A. LaTeX Primer

LaTeX is one of the most commonly used document preparation systems for scientific journals and conferences. It is based on the idea that authors should be able to focus on the content of what they are writing without being distracted by its visual presentation. The source of this file can be used

[1]For those who are particularly interested, other common structures can be found at http://en.wikipedia.org/wiki/README and http://www.gnu.org/software/womb/gnits/.

Basis	Support	Suitable signals	Unsuitable signals
Fourier	global	sine like	localized
wavelet	local	localized	sine like

Table I
CHARACTERISTICS OF FOURIER AND WAVELET BASIS.

as a starting point for how to use the different commands in LaTeX. We are using an IEEE style for this course.

1) Installation: There are various different packages available for processing LaTeX documents. On OSX use MacTeX (http://www.tug.org/mactex/). On Windows, use for example MikTeX (http://miktex.org/).

2) Compiling LaTeX: Your directory should contain at least 4 files, in addition to image files. Images should be in `.png`, `.jpg` or `.pdf` format.

- IEEEtran.cls
- IEEEtran.bst
- groupXX-submission.tex
- groupXX-literature.bib

Note that you should replace groupXX with your chosen group name. Then, from the command line, type:

```
$ pdflatex groupXX-submission
$ bibtex groupXX-literature
$ pdflatex groupXX-submission
$ pdflatex groupXX-submission
```

This should give you a PDF document `groupXX-submission.pdf`.

3) Equations: There are three types of equations available: inline equations, for example $y = mx+c$, which appear in the text, unnumbered equations

$$y = mx + c,$$

which are presented on a line on its own, and numbered equations

$$y = mx + c \tag{3}$$

which you can refer to at a later point (Equation (3)).

4) Tables and Figures: Tables and figures are "floating" objects, which means that the text can flow around it. Note that `figure*` and `table*` cause the corresponding figure or table to span both columns.

V. SUMMARY

The aim of a scientific paper is to convey the idea or discovery of the researcher to the minds of the readers. The associated software package provides the relevant details, which are often only briefly explained in the paper, such that the research can be reproduced. To write good papers, identify your key idea, make your contributions explicit, and use examples and illustrations to describe the problems and solutions.

ACKNOWLEDGEMENTS

The author thanks Christian Sigg for his careful reading and helpful suggestions.

REFERENCES

[1] Editorial, "Scientific writing 101," *Nature Structural & Molecular Biology*, vol. 17, p. 139, 2010.

[2] S. P. Jones, "How to write a great research paper," 2008, microsoft Research Cambridge.

[3] G. Anderson, "How to write a paper in scientific journal style and format," 2004, http://abacus.bates.edu/ ganderso/biology/resources/writing/HTWtoc.html.

[4] J. B. Buckheit and D. L. Donoho, "Wavelab and reproducible research," Stanford University, Tech. Rep., 2009.

[5] R. H. Kallet, "How to write the methods section of a research paper," *Respiratory Care*, vol. 49, no. 10, pp. 1229–1232, 2004.

[6] A. Hunt and D. Thomas, *The Pragmatic Programmer*. Addison Wesley, 1999.

[7] J. Spolsky, *Joel on Software: And on Diverse & Occasionally Related Matters That Will Prove of Interest etc..: And on Diverse and Occasionally Related Matters ... or Ill-Luck, Work with Them in Some Capacity*. APRESS, 2004.

[8] M. Schwab, M. Karrenbach, and J. Claerbout, "Making scientific computations reproducible," *Computing in Science and Engg.*, vol. 2, no. 6, pp. 61–67, 2000.