COMP-598: Applied Machine Learning

Mini-project #1: Building a new ML dataset

Due on September 30, 11:59pm.

Background:

You have been tasked by a local newspaper to build an ML system to predict the popularity of their online news articles. Their goal is to use this information to select how to present articles and sell advertisement.

We found one dataset that tackles a similar problem:
https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

Here is the associated paper describing analysis so far of this dataset (you may need to be on a McGill computer to access witout paying):
http://link.springer.com/chapter/10.1007%2F978-3-319-23485-4_53

Instructions: **READ CAREFULLY!**

The project has two parts.

Part 1: Using the dataset listed above, implement linear regression using the features as provided to accurately predict the *Number of shares* (feature #60).

- The code for linear regression can make use of standard algebra libraries (e.g. for matrix multiplication), but cannot use a machine learning library.
- You must have a working implementation for two types of solutions: closed-form (with matrix inversion) and gradient descent. You can optionally incorporate regularization (ridge, lasso) for either case.
- You should use the features as provided (no feature re-coding).
- You must incorporate good practices for training and testing split.

Part 2: Create a new complementary dataset for this task and apply linear regression.

- The dataset should be acquired from the web. You may need to build a web crawler to do this. You can use existing libraries or code to help you, as long as you reference in your report.
- The dataset must be presented in a CSV file, with each row corresponding to an example and each column corresponding to a feature. The last column should contain the target output variable.
- You must select a set of features to characterize each example. You must include a description of the features in your report in an Appendix called "Attribute Information". See the dataset listed above for an example. You can use features of different types (numerical, categorical text), but each example should have the same number of features, and all entries in the table must be provided (no missing data).
- The dataset must be labeled. You must identify a target variable (it does not have to be the number of shares), and ensure that each example is labeled.

There are many tools to help you collect data, clean it up, annotate, and label, which this project will give you an opportunity to explore. We will discuss some of these in class; the discussion board can be used for peer suggestions.

YOU CANNOT USE A DATASET THAT ALREADY EXISTS (e.g. downloaded from the web in table format with the features already designed). You must have the rights to publicly share the dataset.

Team organization:
The project must be completed in a group of (exactly) 3 students. You will be required to work with different team members on each mini-project. Please plan accordingly: If you want to do the final project with your best friend, don't work with them for this first project! You can use the class discussion board on *myCourses* to find team members. Anyone auditing the course is welcome to participate in the submission and/or review process. However you should not work with people who are taking the course for credit, to avoid mis-matched expectations.

Submisson requirements for part 1:

- You must **submit the code** developed to implement linear regression. The code can be in a language of your choice. The code must be well-documented. The code should include a README file containing instructions on how to run the code. Submit the code as an attachment (see below).
- You must **submit a written report** describing your experimental methodology (e.g. any data pre-processing, training/test split, optimization trick, set of hyper-parameters, etc.) and your results (e.g. training and testing results, error curves, etc.)

Submission requirements for part 2:

- You must **submit the dataset**, in CSV format. Do not submit online. Add a URL to your report.
- The dataset should include a large number of examples (min. hundreds, ideally tens of thousands).
- The dataset should include a rich feature set.
- The dataset should include the output labels for each data example.
- The code for linear regression applied to your new dataset should be the same as for part 1.
- You must **submit a written report** containing:
    - Problem description (motivate your choice of dataset and prediction question, explain the context)
    - Methodology:
        - how and where the dataset was collected
        - how the data can be used for the target task
        - what constitutes an example
        - how features were selected and how they are represented in the file
        - the specific prediction task (output variable)
        - the experimental methodology for application of linear regression (as for part 1)
    - Results of your linear regression analysis (including training and testing error)
    - Final discussion (how useful is this dataset, advantages/limitations, other potential uses, future ideas for dataset acquisition)
    - Appendix: the data dictionary ("Attribute Information").

General guidelines for written reports:

- A single report per team. Include both Part 1 and Part 2 in the same report.
- A project title.
- A list of team members (enter them as "authors" on the submission website).

- A URL where your dataset can be accessed (do not upload the data on the submission website). Put this URL visibly, somewhere on the 1st page (e.g. near the title, in a footnote, or in the abstract). Check carefully that the link works!
- Whenever possible, use figures, tables and graphs to illustrate your work. Always include captions, axes labels, etc.
- Use appropriate referencing style throughout your report (with references listed in a separate section near the end, usually after the appendix).
- Spell-check and proof-read carefully.
- The main text of the report should not exceed 6 pages. The appendix and references can be in excess of the 6 pages. The format should be double-column, 10pt font, min. 1"margins. You can use the standard IEEE conference format, e.g. *ewh.ieee.org/soc/dei/ceidp/docs/CEIDPFormat.doc*. Only acceptable file format is *.pdf*.
- Before the references, add the following statement: "We hereby state that all the work presented in this report is that of the authors." Make sure this statement is truthful!
- Before the references, also add a section with a Statement of Contributions. Briefly describe the contributions of each team member towards each of the components of the project (e.g. defining the problem, developing the methodology, coding the solution, performing the data analysis, writing the report, etc.)

Evaluation criteria:

- Quality and clarity of your report, including descrption of methodology and analysis of results.
- Quality and clarify of your code, ease of running.
- Overall novelty, impact, quality and usefulness of new dataset, clarity of dataset description. Ease of dataset download.
- Overall organization, presentation and writing.

The same evaluation criteria will be used for peer-reviews and evaluation by TAs and instructor. The final grade will be attributed 90% based on the assessment of TAs and instructor. The additional 10% is attributed following your participation in the peer-review process (i.e. for assessing reports of other groups).

Submission instructions:

We will be using an online conference management system to coordinate submission of reports and peer-reviews: https://easychair.org/conferences/?conf=comp598

You should create an account (one per person) on this site before Sept.30. You will use your account both as an "author" and as "PC members" (=peer reviewer) throughout the course. Make sure to use the same account for all your activities and submissions.

When submitting your report, create a "New Submission" (one per group, any of the authors can do this), and link your team members as co-authors. You can revise your submission anytime up to the deadline; do not create more than one submission. The code should be submitted as an "Attachment". Acceptable file formats for this are *.zip, .gz, .tar, .tgz*. Make sure that the code is set up so that we can run it (e.g. include a README file).

Final remarks:

As specified in the syllabus, you are expected to display **initiative, creativity, scientific rigour, critical thinking, and good communication skills**. You don't need to restrict yourself to the requirements listed above – feel free to go beyond, and explore further. It is not expected that all team members will contribute equally to all components. However every team member should make integral contributions to the project.