10-07-22

CSC 859 AI Explainability and Ethics Fall 2022 **** DO NOT POST ONLINE except on iLearn ***

HW 3 – A simple ML experiment with Random Forest ML (RF) using SciKit SW Toolkit – 20 points

DUE: TBD in class (to be posted on iLearn)

D. Petkovic

HW 3 objective

Learn how to use RF with basic SW tools (SciKit or R) for typical ML pipeline:

- Train RF and choose the optimal one
- Run trained RF on test data
- Compute RF feature ranking
- Present in easy to read way, check, analyze and evaluate results – this is also graded and very imprtant

- **Total 20 points**

HW 3 outline

- Implement basic ML pipeline:
  - Import and check given training data from spreadsheet (to be posted on iLearn)
  - Train RF using recommended ranges for RF hyper-parameters ntree, mtry and cutoff (if possible) range of parameters.
  - Choose best RF (min OOB) and save its optimal hyper-parameters
  - Rebuild best RF using optimal hyper-parameters➔ RF run time
  - Extract MDA feature ranking for top 10 features from best trained RF above
  - Use RF run-time to predict a class of 2 samples (run time operation)
  - Present, check and discuss results so they can be understood by others
  - Submit code as PDF document

Training data

E1 cluster set from J. Craig Venter data.Paper describing biology of it is below:

  – Aevermann B., Novotny M., Bakken T., Miller J., Diehl A., Osumi-Sutherland D., Lasken R., Lein E., Scheuermann R.: “Cell type discovery using single cell transcriptomics: implications for ontological representation”, Human Molecular Genetics 27(R1): R40-R47 · March 2018

- All of these training databases have only numerical features (columns), no missing data, and class label (1 to 0) in last column – 1 is E1 sample, 0 is NOT-e1
- Posted on iLearn as e1 cluster data: *e1 positive.xls*
- XLS file has 609 columns with 608 features e.g. values of gene expressions, and last column called “label” as class label, with value of 1 indicating + class sample (e1), and value of 0 indicating – (non e1) class sample.

Tools and resources

- Scikit Python ML tool

OR

- R tool

- Check Ilearn section on “class reading material” Section 2 for good tutorials
- Check posted tutorial slides on SciKit and R class presentations on iLearn

- **Ask help from class expert (TBD) and your team mates. They can help you but YOU must do all the coding, writing and analysis**
- **Must list who helped you in the HW write-up**

HW 3 report – each point below in separate section of PDF HW 3 report – total 20 points

1. Title page (class, school, semester, your name, date, HW 3)
2. Training database used: describe source of data, and audit the DB following guidance in class slides. Provide basic data stats 1-1.5 pages – **2 points**
3. SW tools: explain what tools you used, 1/2 page
4. Experimental Methods and Setup: what ntree, mtry, cutoff ranges you used for RF training (e.g. for grid search). NOTE: SciKit does now allow CUTOFF changes – it uses default of 0.5 and say so; for R tool vary cutoff too., Consult class slides on RF. ½ page or so – **2 points**
5. Results of RF Training and Accuracy Estimates: Train RF for chosen range of parameters and show results for best trained RF, namely: best *ntree, mtry, cutoff*; confusion matrix; OOB;F1; **Discuss results and say what measure you used to optimize RF (OOB or F1 score) (see next slides)**. Pages: as needed – **10 points**
6. Feature Ranking: For best trained (run time) RF show feature ranking using MDA measure, for top 10 ranged features. How does it help you in explaining how RF worked? What did you learn doing this step? Discuss and analyze **– 3 points**
7. RF Run Time test: Take 1 positive and 1 negative sample from training data and run them through best trained RF from 5. to predict its class. Show tool output and classification results and compare for accurate prediction. **Discuss results (e.g. is the prediction correct).** 1 page or so – **2 points**
8. Organization, formatting, results presented in an easy to read way – **1 points**
9. Appendix I: Show key pieces of code YOU developed to run the tool (copy the code into PDF format – do not send executable) .
10. Resources: List resources used and names of students who helped you (if any)

Presenting results – one PDF

HW 3 consists of:

- Title page
- Main body:
  - 10 separate sections as explained before (use same titles), one for each HW 3 item. Each section to be easy to read text documenting what you did with main results shown in graphs, tables – as in a good technical report. (Graphs and table can be cut and paste from tools)
  - Resources and references used, and who helped you
- Appendix: PDF with your code OR pointer to code in your github or something appropriate which does not look like executable to mail filters

Presenting results (also part of the grading)

- Your report has to be self contained and easy to read for average ML professional
- Have separate sections for each task outlined above
- Each chart/result must be fully explained by describing what it is about, listing all data parameters used, explaining the meaning of X and Y axes etc.
- Show ALL outputs and ALL parameters e.g.: ntree, mtry, cutoff (for scikit it is 0.5 fixed), misclassification matrix (label matrix fields); OOB, accuracy; F1 score. Clearly mark and present these parameters.
- For MDA provide list of top 10 ranked features and their MDA values in the table. Plot it for better visualization (optional)
- Must be formatted and presented well for easy reading.
- Please imbed output of tools (e.g. results of classification or training) into a PDF text but make sure you annotate it fully with text
- In Appendix have copy of the code in PDF format (so it does not cause security issues if in executable format). Document the code with header (purpose, HW 3, your name) and in-line for major sections of code

On presentation and provenance

- Provenance of ML pipeline (data, settings, algorithms, training, methods, results) is the key in auditable and transparent ML
- Your report must be understandable to general ML expert who is NOT familiar with your work

HW 3 format and submission

- HW 3 is open book, open internet
- Can ask help from others BUT has to be your own work – no code copy from other students “as is” is allowed and will be checked.
- Do not just cut and paste from class slides or resources, use your language (modify a bit at least)
- Reference all resources you used and list names of those who helped you
- **DUE TBD**
- **If extension needed must ask via e-mail before the deadline**

- **Submission:** One PDF file with name “CSC 859 Fall 2022 HW 3 <your last name>” send as attachment to email to petkovic@sfsu.edu with subject line “CSC 859 Fall 2022 HW 3 <your last name>”

Please you must not post or share HW 3 or the data provided

This HW might be challenging BUT….

- It will improve your understanding of ML and build confidence
- Will add valuable skills: update your CV skills section with tools and methods you used
- You will also practice writing of professional ML reports (and will be graded on that)

- Save HW 3 and make it part of your portfolio for job search (but do not distribute it – show it only if asked)
- Have fun!!!!!!!!!!!!