T-61.6010: Non-discriminatory ML: Report

Due on Wednesday, Nov 18, 2015

Indré Žliobaité

Jonathan Strahl

Fairness-aware regulariser, T. Kamishima et al.

Software and data for the classifier with fairness-aware regulariser [2] are publicly available online, reproducing the experiments only requires running the files and maybe some knowledge of shell scripting and python to make some corrections.
To better understand the code and to see how reproducible the results are I made a new implementation in MATLAB.
The Python code with the shell scripts used to run the code is quite hard to follow in my opinion. It is very well commented though, and the paper describes the regularizer implementation well, so following the parts of the code for the loss and gradient was well supported.

Implementation in MATLAB

All source files are available at Kamishima.net[1]. The first set of source files (data-adultd.tgz) pre process the Adult[2] data set [3] as described in [1]: categorical data is dummy coded and any integer predictors are equally binned into four bins (split over the interquartile ranges). The end result is a numerical discretized matrix and a binary matrix. The matrices are copied into the data folder for the code that runs the cross validation and produces the results. A summary script produces the results with the performance measures described in the paper.
To explore these results I used the same discrete and binary output matrices. The MATLAB implementation uses a non-linear programming solver of unconstrained multivariable functions (fminunc[3]) to optimise the model parameters. The standard logistic regression models use the same optimisation method.

Results

Each performance metric in the Kamishima's paper [2] is implemented in order to make a direct comparison with the results from the original paper. The MATLAB results in table 1 show similar values to the original scores 2 for linear regression (LR) with all predictors (LRs), with the sensitive feature removed (LRns) and with the prejudice aware regularizer (PR) for η=1 being similar to the original for η=5. For larger values of η the results degrade. It would be interesting to look into why this is: is it that the training data is reused for scoring, would it do better generalizing; or is it something technical in the tools used?

Acknowledgements

Thanks to 'quinnliu' on GitHub for the shared logistic regression implementation code in MATLAB as a starting platform for this implementation.

References

[1] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[2] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In PeterA. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer Berlin Heidelberg, 2012.

[3] M. Lichman. UCI machine learning repository, 2013.

[1] http://www.kamishima.net/fac/
[2] https://archive.ics.uci.edu/ml/datasets/Census+Income
[3] http://se.mathworks.com/help/optim/ug/fminunc.html

Table 1: MATLAB results

model	accuracy	NMI	NPI	UEI	CVS	PI/MI
LRs	0.8524	0.2721	0.0518	0.0402	0.1885	0.2055
LRns	0.8517	0.2698	0.0489	0.0398	0.1832	0.1956
PR $\eta = 1$	0.8329	0.2109	0.0133	0.0822	0.0863	0.0679
PR $\eta = 5$	0.7804	0.0898	NaN	0.2675	0.0315	NaN
PR $\eta = 15$	0.8350	0.2186	0.0611	0.1056	0.1581	0.3013
PR $\eta = 30$	0.7638	NaN	NaN	0.3571	0	NaN
PR $\eta = 100$	0.7638	NaN	NaN	0.3571	0	NaN

Table 2: Original results

model	accuracy	NMI	NPI	UEI	CVS	PI/MI
LRs	0.851	0.267	5.21E-02	0.040	0.189	2.10E-01
LRns	0.850	0.266	4.91E-02	0.039	0.184	1.99E-01
PR $\eta = 1$	NA	NA	NA	NA	NA	NA
PR $\eta = 5$	0.842	0.240	4.24E-02	0.088	0.143	1.91E-01
PR $\eta = 15$	0.801	0.158	2.38E-02	0.212	0.050	1.62E-01
PR $\eta = 30$	0.769	0.046	1.68E-02	0.191	0.010	3.94E-01
PR $\eta = 100$	NA	NA	NA	NA	NA	NA