HW2: Speed Dating.
Part 2. Discrimination-aware classification

November 21, 2017

Assignment. In this assignment, you continue the exploration and evaluation of predictive models leant on the Speed Dating dataset. Your goal is to get deeper understanding of model performance and to study how to compare perfomance of different models and their internals/decision logic. Use visualization techniques appropriately (i.e. if this helps you to illustrate a finding/observation more clearly).

The dataset. In this assignment, we use the same Speed Dating dataset.

Submission. Submit your code and your report (with explicit answers to the questions 1.1, 1.2, 2.1 and 2.2 introduced below) via `canvas.tue.nl`, before **December 11**, 23:59 CET. A grade will be given for the group as a whole. Please include a peer-review paragraph stating the degree to which everyone in the group has contributed to the completion of the assignment. If someone contributed much more or much less than the others, and you think this should be reflected in grading, please state this explicitly as well. The report must be submitted as a **PDF**. If you work with Microsoft Word, please export your report to PDF before submitting.

1 Sensitive attributes in classification.

1.1 Modeling without sensitive attributes (30 points)

A concrete application of the predictive models trained in HW1, might be to use them to suggest which participants should be matched, to optimize for the number of successful matches w.r.t. the number of speed dates. Someone might prefer not to have an algorithm deciding with people of what race/income/... they should or should not be matched, in order to have a successful date. Hence, in this assignment, you are asked to build a predictive model that does not use these attributes, i.e. you exclude them from the model. It is up to you to decide which attributes you consider to be sensitive and exclude from the model. For example, one could suggest that the personalized model shouldn't take the attributes race and income into account. Evaluate the models you have built in a similar fashion as before (choosing an appropriate performance measure).

Additionaly, choose one of more discrimination measures from [2] and quantify whether/how much your models discriminate.

Extend your evaluation by also showing the performance of the (same) model that was exposed to previously excluded attributes, and compare these models in terms of accuracy and discrimination.

Summarize your results and findings.

Bonus: you can try to use/implements one of the approaches for discrimination-free classification and report your findings.

1.2 Qualitative comparison (30 points)

Two models that correspondingly include and exclude sensitive attributes are likely to have different behavior/decision logic. You are asked to find and comment on these differences. You can do so by comparing the differences in samples on which models disagree. First make an inventory of samples on which models disagree. After that you can plot – for various attributes – the distribution of this disagreement, and compare it to the global distribution of the attribute.

Come up with a computational approach to summarize the difference in model performance, e.g. by identifying frequent patterns describing a group of cases that are predicted correctly by one model, but not by the other. Compare those to frequent patterns describing cases that are predicted correctly by two models.

You can also try to "look inside" the models and analyse the differences.

Summarize your findings.

2 Patterns of discrimination

2.1 Is there still a bias towards gender? (30 points)

Revisit your unisex predictive model, omitting the gender of the subject and the partner. Having removed the direct reference to gender, have we eliminated completely discrimination? Does the model treat females and males the same? In the data, do women and men earn about the same? Do women and men occupy the same professions? Are women as attracted to younger/older partners?

Some may claim that the unisex model, still knows to discriminate based on gender, also without getting direct access to the genders. Find examples (subgroups related to gender), that illustrate that the unisex model still knows how to make predictions based on gender. One way to start your search, is to find an attribute that is strongly correlated to the gender attribute (gender correlated attribute), and see if gender discrimination happens indirectly through this attribute. For example, select all subjects with a particular feminine attribute for which the prediction is positive, now invert this one femine attribute, and observe the predictions.

After you have done your manual search for a bias towards gender, a more structured approach is as follows. Construct a set of association rules and report the ones that have an extended lift (or elift) [1] higher than a predefined threshold α (for example $\alpha = 2$, meaning this rule is twice as confident about the subgroup for this particular gender, than for all the persons).

Note that in order to measure the elift, you should have association rules of the form $A, B \Rightarrow C$, where A is the protected attribute (gender correlated attribute), B is some context, and C is the attribute you want to predict (dec_predicted). In order to construct these association rules, you need to mine all itemsets (only with a minimum support) that contain both A (gender correlated attribute) and C (dec_predicted), you can use Apriori to do this, but a brute force search where you choose $|B| \leq 2$ should also work.

2.2 Biclustering (10 points)

The idea of biclustering is to cluster a matrix (rows and columns) in the two dimensions at the same time (taking into account the distance between columns and the distance between rows). Note that once biclusters are found, it often makes sense to group columns of the same bicluster together and the same for rows, as to expose the biclustering in the visualizations. Remember that we look for interesting patterns and discriminations or biases that make a group of subjects, somehow coherent given their traits, prefer a specific group of partners, also with somehow specific characteristics. If each row of a matrix is a case from the dataset, and the columns are the features of the case (details about the subject, details about the partner, etc.). Experiment with various settings of the biclustering, and try to name few interesting preferences. For each execution, include the matrix, clearly showing the biclustering pattern and highlighting the points of interest in the patterns.

Note that it is often a good idea to scale the attributes that are used for biclustering to a fixed range.

Summarize you findings.

References

[1] D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 581–592. SIAM, 2009.

[2] I. Zliobaite. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4):1060–1089, 2017.