Assignment 3: Interaction

The **deadline** is July 5th at 14:00 Saarbrücken standard-time. You are free to hand in earlier. You will have to choose **one** topic from the list below, read the articles, and hand in a **report** that critically discusses this material and answers the assignment questions. Reports should summarise the key aspects, but more importantly, should include original and critical thought that show you have acquired a meta level understanding of the topic – plain summaries will *not* suffice. All sources you use should be appropriately referenced, any text you quote should be clearly identified as such. The expected length of a report is between 3 to 5 pages, but there is no limit.

For the topic of your assignment, choose one of the following:

1. *Ulterior*

Discovering non-trivial interactions or correlations requires a powerful score. As we have seen, entropy has many likeable properties, especially that it can detect any type of interaction. Using entropy to define a good score for multivariate continuous-valued random variables is a bit tricky though. Nguyen et al. [1] showed with CMI that we can succesfully use cumulative entropy to this end. Soon after, they improved over CMI with MAC [2] – loosely speaking, a multivariate extension of MIC, that uses cumulative entropy for maximal correlation analysis. Recently, they proposed UDS, which is *universal*. Each of these measures are shown to improve over each other, but one has to remain critical; your task is to carefully read these three papers, connect the dots on how they relate, identifying their respective strengths and weaknesses, identify which corners they cut, and reason about an *even more* ideal score that you could imagine building upon the main ideas of these papers.

2. *Surprise!*

Say the only things you know about your data are that a) its binary, b) its dimensions, and c) its row and column margins. Although this may not seem like much, you can now express an expectation for results mined on the actual data. First read [3], and then [4]. The latter show some impressive plots! Investigate closely what the differences between the different proposed ways to sample random data are. Explain these in your own words, and discuss in particular their assumptions, their strengths, and weaknesses.

But wait, entry-wise PDFs... Haven't we seen this before? Didn't De Bie [5] write a paper on this a few years back? Didn't he use very nice theoretical backing, namely the MaxEnt principle? Discuss how [4] and [5] relate. The model De Bie proposes reduces to a Rasch model, which is a model Golshan et al. compare to. Investigate their results once again, and discuss what it means that they find that the MaxEnt (Rasch) model based on row and column margins obtains higher error than their own methods. The MaxEnt model is a maximum likelihood estimator, so isn't it impossible to get lower errors than the MaxEnt model? Where does this come from? Is this good, or bad? or, shouldnt we care? and so on.

3. *Caveat Lector*

Determining causal relations from data is perhaps the holiest of grails in data mining. Not surprisingly, it is not an easy task. We recently proposed a new approach to inferring whether X causes Y, or vice versa. The main principle is simple; if describing X is easier when given Y, than vice versa, we say that Y is more likely to be caused X than vice versa. Our first attempt to instantiate this principle was by defining a practical score based on cumulative entropy. Experiments show it works rather well. At the same time, it's a first attempt. And first attempts always mean there are some points of improvement.

In this assignment your task is to carefully read [6] and critically discuss it. That is, identify the most important choices the author makes, and evaluate whether you agree with him, or whether alternate solutions (also) make (more) sense. In other words, find as many points of improvement of the principle, as well as of the practical instantiation, as possible. That is, you should consider every aspect critically. Does considering $\Delta_{X' \to Y}$ instead of

$\Delta_{X \to Y}$ really make more sense? Further, instead of calculating it directly, we *estimate* conditional and multivariate cumulative entropy. What are the effects of the way we estimate it on the causality score? Would it be possible to improve in theory? Would it be possible to improve in practice? How?

(Bonus) Recall from the lecture that Heikinheimo et al. [7] show how to mine low entropy sets and trees, and that these trees may seem to represent some kind of causality. Investigate the connection to [6].

4. *Independence* (Hard)

For a long time, practical causal inference depended on Reichenbach's principle of common cause, which allows us to identify causality between X, Y, and Z up to Markov equivalence classes via conditional independence tests. Modern approaches focus on identifying the most likely causal direction between pairs of variables X and Y. One approach, based on the notion of the Additive Noise Models (ANMs), is while rather restricted in theory, relatively easy to instantiate and quite succesful in practice. Another approach, based on the algorithmic Markov condition, is very powerful in theory, but much more difficult to instantiate directly.

Read both [8] and [9] and besides paraphrasing the main ideas, such as why and how can we use ANMs resp. Kolmogorov Complexity to identify causal directions, most importantly, investigate how the two are related. Can the one be seen as an instance of the other? What are the implications? How can we instantiate a score based on the algorithmic Markov condition? Can we plug in any compressor, or does it have special requirements? How would you go about to instantiate it? Would Krimp [?], or Pack [10] do, and if so how? and so on.

Return the assignment by email to tada-staff@mpi-inf.mpg.de by **5 July, 1400 hours**. The subject of the email must start with [TADA]. The assignment must be returned as a PDF and it must contain your name, matriculation number, and e-mail address together with the exact topic of the assignment.

Grading will take into account both Hardness of questions, as well as whether you answer the Bonus questions.

References

You will need a username and password to access the papers outside the MPI network. Contact the lecturer if you don't know the username or password.

[1]   Nguyen, H.V., Müller, E., Vreeken, J., Keller, F. & Böhm, K. CMI: An Information-Theoretic Contrast Measure for Enhancing Subspace Cluster and Outlier Detection (../papers/a3/cmi-vu,mueller,vreeken,boehm.pdf). In Proceedings of the SIAM International Conference on Data Mining (SDM), Austin, TX, pages 198-206, SIAM, 2013.

[2]   Nguyen, H.V., Müller, E., Vreeken, J. & Böhm, K. Multivariate Maximal Correlation Analysis (../papers/a3/mac-nguyen,muller,vreeken,bohm.pdf). In Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, China, pages 775-783, JMLR, 2014.

[3]   Gionis, A., Mannila, H., Mielikäinen, T. & Tsaparas, P. Assessing Data Mining Results Via Swap Randomization (../papers/a3/gionis06swaprnd.pdf). In Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, pages 167-176, ACM, 2006.

[4]   Golshan, B., Byers, J.W. & Terzi, E. What do row and column marginals reveal about your dataset? (../papers/a3/terzi13rowcols.pdf). In Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, 2013.

[5]   De Bie, T. Maximum entropy models and subjective interestingness: an application to tiles in binary databases (../papers/a3/tijl11maxent.pdf). Data Mining and Knowledge Discovery, 23(3):407-446, Springer, 2011.

[6]   Vreeken, J. Causal Inference by Direction of Information (../papers/a3/vreeken15ergowapx.pdf). In Proceedings of the SIAM International Conference on Data Mining (SDM'15), SIAM, 2015.

[7]   Heikinheimo, H., Seppänen, J.K., Hinkkanen, E., Mannila, H. & Mielikäinen, T. Finding low-entropy sets and trees from binary data (../papers/a3/heikinheimo07lesets.pdf). In Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Jose, CA, pages 350-359, 2007.

[8]   Peters, J., Janzing, D. & Schölkopf, B. Identifying Cause and Effect on Discrete Data using Additive Noise Models (../papers/a3/peters10discreteanm.pdf). , pages 597-604, 2010.

[9]   Janzing, D. & Schölkopf, B. Causal Inference Using the Algorithmic Markov Condition (../papers/a3/janzing10algomarkov.pdf). IEEE Transactions on Information Technology, 56(10):5168-5194, 2010.

[10]   Tatti, N. & Vreeken, J. Finding Good Itemsets by Packing Data (../papers/a3/tatti08pack.pdf). In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), Pisa, Italy, pages 588-597, 2008.

 (http://mmci.uni-saarland.de)