Instructions for project work

TIF150, Information theory for complex systems

Contents

1 Overview

The course includes an optional but recommended project. Projects may be practical applications of information theory, or something more theoretical. We give a few suggestions/ideas for projects in this document, but you are more than welcome to come up with your own ideas. We recommend that you briefly discuss your project idea with one of us (Kristian or Susanne) before starting.

The work should be done in groups of 1–3 students and is awarded up to 10 points on the exam. We expect that you spend about 20–40 hours per person on the project.

1.1 Report requirements and recommendations

Describe your work in a report. You may structure the report however you like, but if you are in in doubt it is probably a good idea to structure it like a scientific paper with introduction, method, results, and discussion. The report may start with an abstract, but this is optional. Feel free to include figures and tables as you find appropriate, but do not add figures just because you can. They should all clearly contribute to the story you are telling.

Think of your classmates as the target audience. You do not have to explain basic information theoretical concepts discussed in the course. However, if you use concepts not covered in the course, make sure to describe them in a way that your classmates would understand.

If your group has more than one member, the report **must** contain a contribution report. In the contribution report, describe the contributions made by each group member to the

work you present. Also describe who wrote the different parts of your report. Submissions without contribution reports will be not be graded.

The main text of your report must be no more than 4000 words. The main text is everything except figure and table captions, abstract, appendices, and the contribution report. If you have extra figures, data, etc, which you want to include, you may put them in an appendix. Just remember that the report should be understandable without reading the appendices. Reports that clearly exceed the word limit (we appreciate that this may be counted somewhat differently by different softwares) will not be graded.

When grading your reports, we will consider the following:

(1) Does the report clearly describe what you have done? Is the text easy to follow? Are figures and tables relevant and clearly explained in figure captions, labels, etc?

(2) Is your work of sufficient technical quality, i.e., is it free from serious errors in calculations, etc?

(3) Do you describe and discuss your results in an interesting way, i.e., do you interpret the results, explain why they are relevant or interesting, etc?

(4) Is the amount and quality of the work reasonable considering the group size? Groups of 2–3 people are expected to do a bit more than a single person.

Submit your report by email to Susanne (susannep[at]chalmers.se) **no later than Friday 21 March 2020, 18:00**. Late reports will not be graded.

Reports that fail to comply with the requirements concerning the contribution report, the report length or the deadline will not be graded. You do not get a second chance.

1.2 Project supervision (optional)

We offer up to two 15-minute supervision meetings to each project group. If you wish to have such meetings, you must let us know (please email susannep[at]chalmers.se) **no later than Friday 14 February 2020**, so we have time to assign each group a supervisor and schedule the supervision meetings.

2 Project ideas

2.1 Text analysis and generation

Texts can be analysed in several ways using information theory. Apply some of the concepts of Chapter 3 in the lecture notes to this end. Examples could include a comparison of the information present at different correlation lengths as well as the redundancy in texts by Hemingway and Joyce or automatically generating your own Shakespeare sonnet.

You will need to calculate conditional probabilities for sequences of characters. The number of such probabilities scales with the sequence length n as ν^n where ν is the number of characters in your "alphabet". However, for large n most of these are zero. It may thus pay off to think a bit about how to represent them on the computer.

You will also need a corpus of text to analyze. Text files of a large number of classical works can be found at the Project Gutenberg website, `http://www.gutenberg.org`, but you can surely come up with other ideas too.

2.2 Data compression

Human languages are highly redundant. This realization leads to the idea of compression: computers do not need this redundancy so we can design codes which does away with it, thus reducing required resources for storage and communication.

This lossless compression is purely information theoretical and can be performed on any data regardless of whether the meaning of the data is known. If the intended interpretation of the data is known, it is often possible to accept some losses in the less significant parts of it. This allows for lossy compression, where the loss of invertibility buys larger reduction in size. A common example is the JPEG compression of photographic images which mainly depends on the human inability to perceive fast variations in brightness. Study how the information in a file changes under use of one or more compression schemes. Begin with a lossless one, such as the Ziv-Lempel algorithm. Compute the entropy and correlations and estimate the redundancy before and after the compression. Find a suitable lossy compression algorithm and compare its results with the ones from the lossless one. How does the compression ratios differ and can you estimate the information loss induced?

2.3 Image analysis

Apply the techniques of chapter 3.5 and/or chapter 6 of the lecture notes to analyze the information content in a few pictures. See also chapter 7. Formulate an interesting question and answer it. You could for example try to use the resolution dependent information measure to implement edge detection. Draw images showing the structural information for different resolutions.

If you have studied statistical mechanics (or just are interested) you may want to analyze the Ising model. Study the correlations in a snapshot of a two dimensional Ising model at different temperatures. What happens when you approach the critical temperature? Can you say something about the possibility to generate similar images from conditional probabilities (as is done in text generation)?

2.4 Sound (image) generation

The techniques described in chapter 3 can be used to generate not just texts similar to a particular one. Any data set can be imitated in a similar way. Examine the possibilities of generating sound from an example. Think of how you want to represent the sound. If you want to generate music, each symbol may consist of a particular note and its duration, while if you are interested in sound effects a set of frequency pulses may be more appropriate. Again, note that the larger the alphabet you work with, the harder it will be to handle long correlations.

You can also try to generate images. Start with a small number of colors to keep the computation times reasonable. Note that the correlations are local so you will probably not be able to capture large scale structure. It may be possible to solve this with a hierarchical scheme. If you are ambitious enough to try it and want some ideas, you are welcome to discuss it with us.

2.5 Cryptography

One way to solve ciphers is to try to find patterns or correlations. If you have read about cryptography you might have learned that the go to approach when solving simple substitution ciphers is to look at the frequency of occurrence for the different letters in the crypto compared to unencrypted text. Using information theory we can take this one

step further, by including correlations at longer distances. You might also want to consider approaching more advanced ciphers, for example the Vigenère cipher . This cipher is much more difficult but the periodicity enforced by the keyword length could possibly be obtained by looking at correlations. The goal would be to create a program that automatically solves encrypted texts. If you need sample texts, a large number of classical works can be found at the Project Gutenberg website, http://www.gutenberg.org/.

2.6 Clustering in networks

To analyze a large network, e.g., a network of scientific papers linked by citations, or a network of cities linked by flight travels, it is often useful to find clusters of similar nodes, to make a simplified description of the network. One information-theoretic approach to this problem is provided in the map equation (Rosvall et al., 2009). There is also a variant called the hierarchical map equation (Rosvall and Bergström, 2011), and a number of other publications apply or develop the concept further (see http://www.mapequation.org/publications.html for a list).

Pick one or two of these papers. What do they say? What sort of problems does the method solve? Can you implement the method and apply it to some data set? What did you learn? This is a rather open-ended project idea, where you have to further specify the goal of the project. You may choose to go more into theory or into practical applications.

2.7 Cross-entropy minimization in spatial datasets

If a spatial dataset has too low resolution, can one estimate a higher resolution with the help of other, related datasets? The short answer is yes, but there are many different ideas about how to do this.

One example is the method of cross-entropy minimization (You and Wood, 2005; You et al., 2014). Roughly speaking, these two papers demonstrate how regional statistics on crop production areas (a low-resolution dataset) can be disaggregated, i.e., transformed into a higher-resolution image, by adding higher-resolution information on where the cropland is located, how suitable different locations are for different crops, etc. As noted by You and Wood (2005), for this "we need an approach that can utilize all such information, recognizing that the information may be limited, partially correct, and sometimes conflicting."

What do these papers say? Do you think the method is reasonable? What limitations does it have? What assumptions are needed to make it work? What sort of errors can be expected in the higher-resolution estimates?

It might be both fun and instructive to implement a simple version of such a calculation. If you are interested in doing this, Rasmus can help to pick out some relevant data that is less complicated than the examples given in the papers above. We can set up simple datasets so you do not have to deal with geographical information systems.

References

Rosvall, M., D. Axelsson, and C. T. Bergstrom (Nov. 1, 2009). The Map Equation. *The European Physical Journal Special Topics* **178** (1), pp. 13–23. DOI: 10.1140/epjst/e2010-01179-1.

Rosvall, M. and C. T. Bergström (Apr. 8, 2011). Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems. *PLOS ONE* **6** (4), e18209. DOI: 10.1371/journal.pone.0018209.

You, L. and S. Wood (Dec. 2005). Assessing the Spatial Distribution of Crop Areas Using a Cross-Entropy Method. *International Journal of Applied Earth Observation and Geoinformation*. Bridging Scales and EpistemologiesLinking Local Knowledge with Global Science in Multi-Scale Assessments **7** (4), pp. 310–323. DOI: 10.1016/j.jag.2005.06.010.

You, L., S. Wood, U. Wood-Sichra, and W. Wu (May 2014). Generating Global Crop Distribution Maps: From Census to Grid. *Agricultural Systems* **127**, pp. 53–60. DOI: 10.1016/j.agsy.2014.01.002.