Abstract

For the course Big Data you are required to write an essay. The goal of this essay is to demonstrate your knowledge of the content of this course. The target audience for this essay is a fellow master student that did not follow this course. Therefore you can assume that the reader knows basic concepts from probability theory and computer science, but you cannot assume that the reader knows any term used in the course.

Arguably, "demonstrating knowledge" is a rather vague description for a product that determines whether or not you pass a course. Therefore this document aims to specify the informal requirements that we have for this essay. The Oxford dictionary describes the word *knowledge* as "Facts, information and theoretical understanding of a subject". Since this is a master-level course, the content of your essay must entirely focus on the latter; do not spend time or space on copying definitions, formulas or even computations that are presented in the slides. Each chapter of the essay should rather focus on the *understanding* part of knowledge e.g.: what is the use of the *Apriori property*? What are closures and why do we need them? What does the term *agnostic learning* mean and what are its consequences?[1]

[1] Your own essay should not have an abstract, and footnotes are for small remarks not citations.

Requirements for Big Data

Ivor van der Hoog

January 2019

1 Introduction

To make your life (and our life) easier, we provide very specific parameters to which your essay must comply. The first parameter, is that we provide you with this style file in LaTeX. This allows to enforce a page-limit in an objective way. Most scientific journals have their own LaTeX style file, so you can regard using this style file as practice for the future.

The easiest option for you, is to copy the content of this tex document. This contains all the LaTeX code you need to include sections, figures, theorems and equations. If you are really not comfortable in LaTeX and would rather work with a text-editor like Word, we ask you to set the document settings to match the sizes and margins in this document.

Given this style file, we ask that the document that you hand in between six and seven pages. We are also providing a skeleton which your essay must follow. The section headers that you see in this document, must be present in your own essay in the same order. You are free to choose for each (sub-)section what the size of this section will be. However, in previous years we experienced that our students found it difficult to match our expectations for the essay. To help you navigate in this direction, we also provide a suggestion for how long each section could be. For example: it would be wise to not let your introduction be much longer than this introduction. Skip ahead to Figure 1 and 2 for a visible example.

Problem statement. Anyone will tell you that Big Data is a current-day problem in Computer Science (there is a reason that Big Data is a mandatory course). But what problem are we trying to tackle with this course? This paragraph is meant to (briefly!) specify what this essay and the course is about. What *is* data analysis? What is the result of data analysis and what problems arise if a database becomes too large? Your introduction together with the problem statement should be at most half a page.

2 Frequent item set mining

At various points we discussed frequent item set mining. Do not make this section a large list of definitions and facts. A common mistake is that a student constricts this dissertation to only contain a list of facts: "a *lattice* is a structure with three properties: *idempotency*, *commutativity*, *associativity* - (end of section)". It is O.K. to spend time on discussing properties, but focus on the following: what power do these definitions give you? What is the purpose of a property like commutativity? You can also provide examples of where these properties occur in data.

Another common mistake is that a student only focuses on the formal (mathematical) definitions surrounding item set mining. e.g.: "We say that there is a *Galois connection* between two lattices $(P(\mathcal{I}, \subseteq)$ and $P(\mathcal{D}, \subseteq)$ if for $I \in P(\mathcal{I}, \subseteq)$ and $E' \in P(\mathcal{D}, \subseteq)$ we have that (Lecture 3, slide 31):"

$$[E' \subseteq F(I)] \Leftrightarrow [\forall t \in E' \mid I \subseteq t]$$
$$\Leftrightarrow [I \subseteq \cap_{t \in E'} t] \Leftrightarrow [I \subseteq G(E')]$$

However, this only demonstrates that you are able to copy a definition from a slide. If you decide to discuss a property such as a Galois connection, then describe the semantics of the definition and focus on what each a property is good for: what kind of problem is this property trying to solve or circumvent?

This section is at most one-and-a-half page. Below is a list of terms that you should mention. Whenever you mention a definition, always focus on the intuition behind the definition and on what "power" the definition gives you.

- Transactional databases.
- Frequent items sets.
- (Optional) You could be more general and describe frequent patterns.
- Apriori.
- Closures.
- Range spaces of frequent item sets.
- From frequent item set mining to classification.

3 PAC-learning

The bulk of the lectures (and the exercises) discussed results around PAC-learnability of hypothesis classes. It should be no surprise that a large portion of the essay should also cover this subject. This section should be between 2 and 3 pages (depending on how large the previous sections are).

3.1 Classification, Quality and convergence

This subsection should explain what we mean with classification and hypothesis classes in the context of sampling and big data. You could also discuss learning algorithms and concepts such as: the halving-learning algorithm, empirical risk minimization, empirical loss, true loss and the notion of convergence.

3.2 Realizable vs agnostic PAC-learning

This subsection should discuss both variants of PAC-learning that were covered in the course. These results make use of the Union bound and the Hoeffding bound, and this was covered in the exercise class. Also discuss the difference between the two PAC-bounds.

3.3 VC-dimension and the fundamental theorem

This subsection should explain the details of the fundamental theorem.

3.4 (Optional) Bias and no free lunch

The aforementioned subsections could fill plenty of pages on their own. But if you manage to present PAC-learning (or frequent item set mining) in a concise manner then it could be that you have some space left. In this case, you could choose to add one of the two optional subsections. One option is to discuss bias from the hypothesis class and how this relates to the no free lunch theorem. These two subsections could be very interesting in their own right, but remember that it is always better to have 3 complete subsections than 4 incomplete ones!

3.5 (Optional) Examples of PAC-learnable hypothesis classes

The lectures contain a few elaborate examples of non-trivially PAC-learnable classes. If you want, you can spend time and space to explain why a certain hypothesis class is or is not PAC-learnable. You could even choose to discuss a hypothesis class that is not mentioned on the lecture slides.

4 Frequent item set mining on big data

This course aims to explain the results from two influential papers (Toivonen 1996, Riondato and Upfal 2014) on the subject of frequent item set mining. In this section, you demonstrate your knowledge of the course by describing both results and by discussing the difference between the two results. All the relevant content for this section was covered during the course, there is no need to read the two papers.

The previous sections are designed to contain all the definitions, prior results and concepts that are needed to understand the result from Toivonen and Riondato and Upfal. Hence, this section can be rather short. Each subsection should be at most half a page.

4.1 Toivonen and frequent item sets

Toivonen was one of the first to describe how to find frequent item sets on big data sets. Specifically, Toivonen specifies how large a sample should be, before you can "accurately" learn frequent item sets from a large database. In this subsection you should discuss this theoretical result.

4.2 PAC-learning and frequent item sets

Riondato and Upfal showed that the problem of frequent item set mining is PAC-learnable. This subsection should discuss two things: Why is it possible to PAC-learn frequent item sets and how large should your sample size be (according to Riondato and Upfal) and why? Refer to explanations and definitions from prior sections such as range spaces.

4.3 Toivonen vs PAC-learning

We have seen two approaches for sampling for frequent item set mining which gives us two different bounds for the sample size, and different assurances for each result. In this subsection you should discuss what these differences are.

5 (Optional) extra section

The previous section ended in a theoretical comparison of the bounds provided by Toivonen and those by Riondato and Upfal. In this optional final section you present an *experimental* evaluation.

More precisely, you

- download a appropriate data set from, e.g., Kaggle or the UCI machine learning repository
- you determine the sample sizes as given by both methods
- you take n samples of the determined sizes and compare the frequent item sets computed from

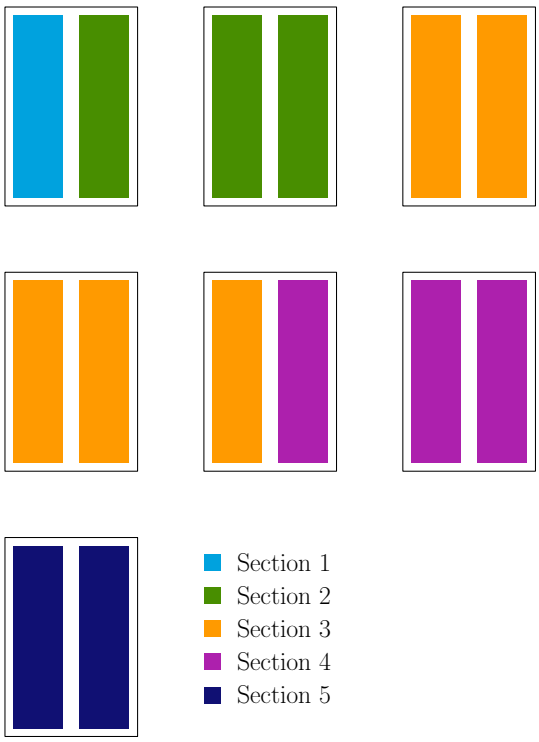

Figure 1: Student 1 is a regular student aiming for a sufficient. Student 1 filled every section up to the maximal limit but did decide to write the optional final section just to be sure.

these samples with those computed on the full data set

- are the guarantees met?

Describe your experiments, your results and the conclusions you draw in a reproducible manner.

6 Useful information:

Here is some useful information for if you want to get started on your essay. This section should be especially useful if you make it in LaTeX. Here is how you include a figure and reference to a figure. Figure 1 and 2 each show an example for how your essay could be structured:

The distribution is denoted by $\mathcal{D}$. A sample of size m is denoted by $D \in \mathcal{D}^m$. We have a loss function on the database $\mathcal{L}_D$ and a loss function on the complete distribution $\mathcal{L}_{\mathcal{D}}$.

We have quantifiers $\epsilon, \delta, \sigma, \theta, m, n$, we write the expected probability of X as $\mathbb{E}(X)$.

Equations with fractions and square roots are written as follows:

$$a^2 + b^2 = \frac{c^3\sqrt{a^2 + b^2}}{c^2} \tag{1}$$

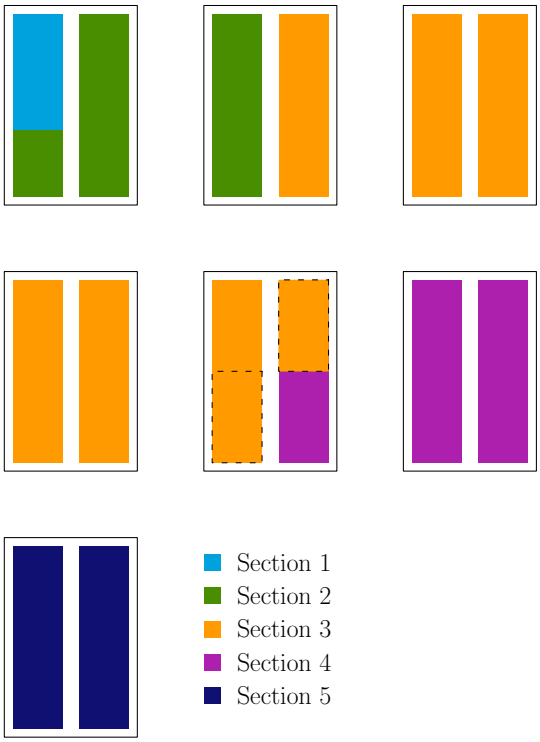

Figure 2: Student 2 tries to get a very high grade. Student 2 manages to describe the problem statement, frequent item set mining and PAC-learning very concisely, without missing essential information! Since student 2 had some space to spare, student 2 chose to implement an optional subsection of Section 2.