Review of TDT4173 Method paper

Paper ID: **Unsupervised_Learning-16.pdf**

- Write a short (2-3) sentence about the paper
    - The overall topic of the paper is unsupervised learning within machine learning. The paper focuses on clustering and the K-means algorithm. Alternative algorithms are also presented. The last section presents current applications of the K-means algorithm, presenting current research on lung cancer and analyzation of seismic data.

- Relevance: Does the paper fit in the overall topic?
    - The introduction explains differences between supervised and unsupervised learning, and clearly states that the core of unsupervised learning is clustering. Because of this, the paper will focus on clustering and the K-means algorithm. In section 2, there is an in-depth explanation of what clustering is, and what the K-means algorithm used for clustering are. Later, in section 3, the paper informs the reader of drawbacks of the K-means algorithm and presents alternatives to it, still clinging to the topic of clustering and unsupervised learning. Section 4 is also focused around clustering in applications of current research.
      From my understanding, the paper fit excellent in the overall topic.

Relevance:	1 (poor)	2	3	4	[✓ 5 (excellent)]

- Technical Soundness
    - Core formulas and theoretical foundations are presented. Section 2 starts out by explaining that it is important to have a small within-cluster variation and an equation showing how this can be obtained. The equation does not speak for itself, and more explanation of the equation would have been preferred. Specifically, what all variables represent and how the equation work (what are the x'es? The p?), and how this equation is indeed minimizing the with-cluster variation.
      Further, the section are presenting equations and algorithm for the K-means algorithm. These equations and algorithm are well explained and fairly easy to follow and understand. Nice!
      The section then lists quality functions, showing different methods of checking the K-means quality, which is nice.
      The explanation of overfitting and underfitting are nice, with a good example that is easy to follow! The illustration (figure 4) is also helping with understanding the theory.

Technical soundness:	1 (poor)	2	3	[✓ 4]	5 (excellent)

- Clarity and Presentation
    - The paper uses an easy language, and are for the most part very clear and understandable, though I think section 3 is a bit messy. The first part is great, with a nice example to follow (the square-example). When reaching the K-means++, I would have liked a bit more explanation than only presenting the algorithm. Then the paper suddenly switches to talk about K-medios (when I was expecting more on the K-means++). This paragraph of the K-medios algorithm, however, is well-structured and easy to follow and understand. K-modes and CLARA are also

easy to follow. I think the problem I had with this section is that there are a lot of different information on just two pages, making it seem a bit much.

  - Then again, the paper is for the most part well-written, well-structured, easy to follow and easy to understand! The format of intro - foundations - alternatives - current applications - a conclusion is making the paper easy to follow and gives the paper a natural progression of information.

Clarity:	1 (poor)	2	3	[✓ 4]	5 (excellent)

- Does the paper points out current (successful) use of the method?
  - The paper points out several different real-life problems where unsupervised learning can be applied. It also goes deeper into two current applications - analyzation of seismic data and lung cancer.
  - Seismic data:
    The paper explains the problem well, why unsupervised learning can be used and that the research used the Gaussian mixture model for their learning process. Given that the research is not yet finished, it is understandable that results are not presented.
  - Lung cancer:
    The paper explains the problem well, and why unsupervised learning can be used. It also explains what kind of algorithms that were used in the study (K-means) and the results the research found.

Current applications:	1 (poor)	2	3	4	[✓ 5 (excellent)]

- List 3 good points of the paper and make 3 suggestions how to improve the paper
- Suggestions for improvements:
  1. In section 3, maybe cut down on how many improvements you present and go deeper into some of them.
  2. Explain equation 1 in more detail. Escpecially what all variables mean.
  3. Explain more around the algorithm for K-means++.
- Good points:
  1. Very nice examples and figures that are easy to follow and understand!
  2. Very natural progression of information, starting with an introduction, presenting K-means, improvements to K-means and then current applications!
  3. You explain each new term you introduce, so that it is easy for the reader to follow along.

- List of suggestions / recommendations
  - I really enjoy the possibility to click on figure numbers to "jump" to the figure. It does not seem to work in this paper. It would have been nice to have that feature. (In LaTeX, put a \label{fig:something} in the figure and then when referring to the it, use "As you can see in figure \ref{fig:something}")
  - Missing an `e` on the end of `therefore` on bottom of page `3` and the last sentence of paragraph 2 on page `5`
  - The figure text on figure 5 are a bit weird, the space between the texts should be bigger. The first 3-4 times I read it, I did not realize it was three different texts.