6. Data Science ML Introduction Exercise

By Robin Hegenberg, Matr. Nr. 355386

1. Read the blog https://towardsdatascience.com/12-useful-things-to-know-about-machine-learning-487d3104e28 and write down one point which you think is important

A major point of machine learning that I derived from the text above is the overall goal of generalization with its challenges regarding overfitting especially. A perfectly generalizing machine learning tool does not overfit on specific data sets, but does well in the real world problem scenario with never-before-seen data too. It's good to see that many approaches tackle this problem already. Most importantly, I agree to split the overall data into training and test sets from the beginning on to ensure evaluation does occur on data that is new for the algorithm (cross-validation). Moreover, it is critical to make sure that the data fits well to the problem task, ideally including many examples of all relevant scenarios and making sure that they are well-distributed into both data sets (plus, feature selection is a big point). Another problem with overfitting is that it comes in many forms that are not so easy to grasp. A good example is the tradeoff regarding bias and variance errors: While comparably complex algorithms like decision trees often suffer from high variance and overfitting errors, easier algorithms like linear regression have a high bias instead and therefore tend to underfit. Adding a regularization term that punishes high complexity or adding a statistical significance test like chi-quare that determines if a more complex structure is needed before adding onto it are very clever approaches here. Combining multiple algorithms instead of having one that presumably fits best to the problem task is another huge improvement. Despite all of this, overfitting remains an unsolved challenge that we are sure about to see more research in.

2. Give a use case in which machine learning can help (you need to describe which type ML is used, supervise or unsupervised, regression or classification)

A common use case is hand-writing recognition: It is difficult to solve this task with traditional logic-based approaches because nearly infinite variations exist and for humans it is difficult to say what exactly makes a graphic become a certain character. However, it can be solved using typically a supervised classification approach: With labelled examples that fall into one of the (ASCII, Unicode, …) character set categories (meaning, representing one of their characters).

3. What is Data-Centric AI? How can you use data-centric concepts to improve your use case?

While traditional AI focusses on getting results based on the provided data, a data-centric approach would be to determine which data is needed to make the AI work best and build processes around this data collection. Improving the quality and quantity of the data is key here. For the handwriting recognition use case above this would mean collecting a large variety of different handwriting styles and label them correctly. Maybe collect words instead of single characters too, (and use different lighting settings and camera angles if you capture them by camera and so on). The importance of those parameters must be closely observed though. Maybe create an app that lets users write things and say what they meant to write while the app collects all those parameters, creating a useful dataset.