IN3062 module - Introduction to Artificial Intelligence

Individual report

Jean-Louis Gosselin

As previously mentioned in thre main report, if unsupervised learning has the "advantage" of having (for us students at least) a very limited number of algorithms to process on unlabelled data, and if one of these algorithms (K-means) is fairly comprehensible at first glance, the challenge lay however with everthing related to PCA!

Indeed, from the "curse of dimensionality" to the concepts of variance and bias, and from Eigenvectors to axis rotation in PCA, I clearly had my work cut out, since I found the maths behind PCA quite daunting to understand in their application.

Fortunately, I found great online tools to help me visualize the bigger picture more clearly:

- a series of Youtube tutorial videos ("StatQuest", Josh Stormer, 2018)
- a series of articles on Medium.com (which I am now subscribed to) related to all things AI and specifically to unsupervised learning
- another series of tutorial videos on datacamp.com (which I am also now subscribed to), taught by Ben Wilson

Another challenge was that I only had at my disposal two specific case studies:

- randomly generated data points
- hand-written digit recognition

…with each case study somehow circumventing the intricacies of working with a "normal" generic dataset containing nothing other than a variety of linear values – intricacies I was left to untangle by myself!

Frustratingly, searching for snippets of code on PCA was made more difficult by the fact that two recurring examples (often used by online mentors and writers) demonstrating unsupervised learning techniques concerned the Iris or breast-cancer datasets – both loaded from Sklearn! To start with, these datasets are labelled (which added an element of confusion as to whether or not PCA really is a technique specific to unsupervised learning!), and they are already given class-specific functions and attributes – unuseable for the sort of generic spreadsheet-based dataset I was looking to work with.

On a last note, I've yet to master some of the fundamental basics of Python, notably: the difference (and the possible interchangeability) between numpy arrays and panda dataframes – a notion which caused me much grief at times.

Overall, the plan we agreed on with Negar was that I was to tackle PCA and K-means, both as abstract notions and in their Python implementation and possible correlation. Any findings were shared at our weekly meetings, to ensure that we steered the project in the right direction.