Summary#1

Hector Lopez CAP6673 FAU SPRING 2018

Twitter is a platform where about 20% of the internet engages in providing real time data. It is a great tool for applying data analysis and training learners to extract information from a live stream of consciousness. The Twitter API provides many data mining benefits. We can use natural language processing to extract features from the streaming data and ascertain information such as sentiment analysis. Emoticons in the data can be extracted as features containing the sentiment of each tweet. The text itself can also be used to express information about the sentiment. When there are many features to in a dataset it is called "Dimensionality" . With greater dimensionality it is harder to model a system to learn the emotion from the data. Tests such as ANOVA can measure the different factors and their importance of a set of data. This feature selection technique can help to improve learners. Using feature selection and other ensemble methods greater levels of performance can be achieved than just bagging and boosting alone. There is a large benefit in data sampling. Anytime we incorporate many features we must perform feature selection, or we will not see any benefit from data with high dimensionality.

Computational hardware has progressed to do massively parallel computations, so with the availability of large data sets we can create models for complex nonlinear functions. This is known as deep learning. It was first designed around neural networks based on a single biological neuron. By varying the size of the synapse analog connections can embed pathways for calculations. Layering these neurons can simulate more complex computations. Activation functions allow signals to be passed to other neurons, these activation functions can be stepped but are more useful as non-linear so that the network can create non-linear models. Forward feed networks, are trained in one direction, and retrained as necessary. But a Back-feed network ties the output into the input and trains itself.

Tools like, numpy and scikit learn can help with machine learning analysis. These tools allow for more programmatic benefits than the Weka tool. NLTK (natural language tool kit) is used for feature extraction from text. The browser based software, Jupyter notebooks can also help to build programs to perform machine learning and data analysis. When looking for tools for deep neural network processing the most widely used and best supported deep neural network back-end is called TensorFlow by Google, CTK ,is Microsoft's version. When dealing with the back-end services a higher level interface like Keras allows for the interoperability of tensor flow or ctk while only utilizing a python based API.