MACHINE LEARNING – CLUSTERING, REGRESSION AND CLASSIFICATION

In my last post of this series, I explained the concept of supervised, unsupervised and semi-supervised machine learning. In this post, we will go a bit deeper into machine learning – clustering, regression and classification (but don't worry, it won't be that deep yet!) and look at more concrete topics. But first of all, we have to define some terms, which basically derive from statistics or mathematics.

Features and Labels in Machine Learning

- Features
- Labels

Features are known values, which are often used to calculate results. This are the variables that have an impact on a prediction. If we talk about manufacturing, we might want to reduce junk in our production line. Known features from a machine could then be: Temperature, Humidity, Operator, Time since last service. Based on these Features, we can later calculate the quality of the machine output

Labels are the values we want to build the prediction on. In training data, labels are mostly known, but for the prediction they are not known. When we focus on the machine data example from above, a label would be the quality. So all of the features together make up for a good or bad quality and algorithms can now calculate the quality based on that.

Machine Learning: Clustering, Classification and Regression

The first one is **clustering**. Clustering is an unsupervised technique. With clustering, the algorithm tries to find a pattern in data sets without labels associated with it. This could be a clustering of buying behaviour of customers. Features for this would be the household income, age, … and clusters of different consumers could then be built. The next one is **classification**. In contrast to clustering, classification is a supervised technique. The last technique for this post is **regression.** Regression is often confused with clustering, but it is still different from it. With a regression, no classified labels (such as good or bad, spam or not spam, …) are predicted. Instead, regression outputs continuous, often unbound, numbers. This makes it useful for financial predictions and alike. A common known sample is the prediciton of housing prices, where several values (FEATURES!) are known, such as distance to specific landmarks, plot size,… The algorithms could then predict a price for your house and the amount you can sell it for..