IFT6390 Final Project proposal

Analysis and comparison of main machine learning algorithms in fashion MNIST and income prediction data sets

5 members:

November 12 , 2018

Lifeng Wan matricule:20108546.
Jinfang Luo matricule:20111308.
Yan Ai matricule:20027063
Ying Xiao. matricule :20111402
Qiang Ye matricule: 20139927

There are too many machine learning algorithms, such as classification, regression, clustering, recommendation, image recognition and so on. It is really not easy to find a suitable algorithm. Usually, we start with commonly accepted algorithms, such as SVM, Adaboost or neural network. But we wanted to know if these algorithms were equally effective for all different data sets. So we chose three representative algorithms to compare on two different data sets. Compare their differences to see which algorithm works better on the two data sets and why.

The three classifiers we have chosen:

- Naive Bayes :This model from the classical mathematics theory and has a stable classification efficiency. It's relatively simple, often used for text classification. If the conditional independence assumption is made, the Naive Bayes classifier will converge faster than the discriminant model, such as logistic regression. So we only need less training data. Even if the NB conditional independence assumption is not true, the NB classifier still performs well in practice. Its main drawback is that it cannot learn the interaction between features. we used it rare before, and we want to know it more, so we chose it.

- Support Vector Machine: High accuracy, and even if the data is linearly inseparable in the original feature space, given a proper kernel function, it will work well. It is especially popular in text categorization with high dimension. But it's not very efficient when you have a lot of samples, and sometimes it's hard to find a proper kernel. But it's also Very popular, we chose it to see how good its performance is.

- Multilayer perceptron: High classification accuracy; It can fully approximate complex nonlinear relations; It has the function of associative memory. But it takes a lot of parameters; You can't observe the learning process, and the learning process takes so long that you may not even reach the goal of learning. We chose it, because it's very hot now, we want to see the different with SVM.

The two datasets we have chosen:

- Fashion MNIST : consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes.

- Income prediction: Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using some conditions. Its prediction task is to determine whether a person makes over 50K a year.