Natnael Tsige

CPTS 315

Homework 4

4/ 8 / 2022

CptS 315: Introduction to Data Mining

Homework 3

(Due date: April 5, midnight PST)

1. A. It is non-linear. suppose we have two weight vectors (0,1) and (−1,0) and are trying to distinguish between positive and negative examples. Both weight vectors have a "survival time" of 1 (i.e., they both last the same number of iterations during the training phase). The only instances which are labeled as positive will be in the upper left quadrant. Thus it is nonlinear.

   B. Yes, the decision boundary of the averaged perceptron is linear. Here if we have the 2 weighted vectors as in the above-mentioned answer, then we have averaged those 2 weighted vectors which form a single weighted vector. A single weighted vector always defines a linear boundary.

2. In this case something like voted perceptron is a good solution because it enables us to obtain more accurate result.

3. The main problem in this situation is the vast majority of our example is negative and we care more about the accuracy of the positive example. One way to deal with the imbalance is to oversample the positives and under-sample the negatives.

4. Suppose we are given score(k) = w * (x_k) and score (l) = w * (x_l). If it is the case score(k) > Score(l) then it implies that the score confirms to score(k > l). In addition, score (k > l) implies that score(k) - score (l) > 0. Should the case be otherwise we adjust the weight.

5. CLOSED boundary is a hyperplane that is linearly separatable one side positive examples and other side positive examples.

   A direct proof would be to showing that
   C+ > 0 = 1 and C- < 0 = 0

6.

A few useful things to know About Machine Learning
Summary

The reading addresses some issues and misconception about Machine Learning via 12 key pointes. Throughout the reading the author use classifiers as an example.

The author defines what learning is as the by stating learning = representation + evaluation + optimization. Representation simply means classifier must be represented in some formal language that the computer can handle. Evolution is needed to distinguish good classifiers from bad ones, and Optimization is need a method to search among the classifiers in the language for the highest-scoring one. The three elements help us to not get lost into a sea of algorithms when deciding which algorithm to apply.

Another points the author touched on generalization. Obviously, the goal of Machine Learning to be able to generalize that is if a Machine Learning is successful it should work beyond the scope of the training sample. In this action the author also addresses a misconception beginners make in Machine Learning which if ML works on a training data then it works, which is not the case.

Another key point addressed in the reading is that data alone by itself is not enough. Every learner must embody some knowledge or assumptions beyond the data it is given in order to generalize beyond it.

A major problem addressed in the reading is overfitting. Overfitting occurs when there is no enough data to make a proper classification. Such problems can be addressed by ensuring our weight and bias are being properly managed.

The second major problem in Machine Learning is the curse of dimensionality. To put it simply curse of dimensionality means as our number of futures and dimension growth the amount of data need to accurately generalize exponentially and we run in to a multitude of problems. The paper proposes blessing of nonuniformity as a solution to the problem.

Overall I have learned the most misconceptions people make about Machine Learning and I also learned the issue that are within Machine Learning.