Elijah Andrushenko

CPTS 315

Data Mining

Jana Doppa

Homework #3 (Analytical Part)

1A) **No**, Voted Perceptron tries to discern between positive and negative examples leaving quadrants of positives and negatives rather than a linear pattern.

1B) **Yes**, Averaged Perceptron finds the average of two weight vectors which leads to a linear decision boundary.

2) I would consider using voted perceptron if the importance of the training data is emphasized. Although it sacrifices space and time complexity, it will yield more accurate results. This also lets us adjust the weight vector until an optimal solution is found.

3) I would modify the perceptron algorithm by including geometric margins and max margin classifiers. Since the data is linearly separable, we can get better accuracy and confidence using the concept of margin.

4) First, we subtract the score of l from k. This lets us know if we have a positive or negative score which tells us which score it higher between k and l. We then would compare the output with the desired output. If it matches move on to the next comparison, otherwise we make changes.

5) The CLOSE boundary is a hyperplane that is linearly separable. The distinction is made with positive and negative training examples. When the positive is on one side of a linear line and the negative being on the other. The training examples that are considered as support vectors are the ones that are closest to that linear separator.

6) Machine learning is the process of learning a program from data, and it is used in all sorts of computational fields from search algorithms to stock trading. The most used machine learning type if classification. In machine learning the data must be properly represented. One that is done, then it can be evaluated and after evaluation it can finally be optimized. A fundamental goal of machine learning is to generalize, so that the program works correctly on more than just the provided training set. A common mistake programmers make when it comes to machine learning is overfitting the training data. Thus, defeating the goal of generalization and making a program that can only reliably predict the training data. Steps to take to avoid this scenario would be large and diverse sets of training data and cross validation. More data is better than a clever algorithm. A weaker algorithm with a lot of data to provide the program with, will end up being more reliable than a strong algorithm with very limited data. Learning many models rather than just one is useful in many cases. Some models are more suitable for solving certain problems and will help make a better program.