Singapore University of Technology and Design
Information Systems Technology and Design
50.007 Machine Learning

Homework 1 (Due: Sep 30, 2014; in class)

Question 1. Prove that the perceptron learning algorithm (PLA) converges to a linear separator for separable data $\{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), \ldots, (\mathbf{x}^N, y^N)\}$ by following the steps below. Hint: Let $\mathbf{w}^*$ be a set of weights that separates the data, and show that $\mathbf{w}^t$ gets "closer" to $\mathbf{w}^*$ with each iteration. For simplicity, assume that $\mathbf{w}^0 = \mathbf{0}$.

(a) Show that $\rho = \min_{1 \le i \le N} y^i(\mathbf{w}^* \cdot \mathbf{x}^i) \ge 0$. *(5 points)*

(b) Show that $\mathbf{w}^t \cdot \mathbf{w}^* \ge \mathbf{w}^{t-1} \cdot \mathbf{w}^* + \rho$. *(4 points)*

(c) Show by induction that $\mathbf{w}^t \cdot \mathbf{w}^* \ge t\rho$. *(4 points)*

(d) Show that $\mathbf{w}^t \cdot \mathbf{w}^t \le \mathbf{w}^{t-1} \cdot \mathbf{w}^{t-1} + \mathbf{x}^{t-1} \cdot \mathbf{x}^{t-1}$ where $(\mathbf{x}^{t-1}, y^{t-1})$ is the example chosen in the t iteration of PLA. (Hint: $y^{t-1}(\mathbf{w}^{t-1} \cdot \mathbf{x}^{t-1}) \le 0$ because $\mathbf{x}^{t-1}$ was misclassified by $\mathbf{w}^{t-1}$.) *(4 points)*

(e) Show by induction that $\mathbf{w}^t \cdot \mathbf{w}^t \le tR^2$ where $R = \max_{1 \le i \le N} \sqrt{\mathbf{x}^i \cdot \mathbf{x}^i}$ ($R^2 = R \times R$. The 2 is not an index.) *(4 points)*

(f) Using (c) and (e), show that $\frac{\mathbf{w}^t}{\sqrt{\mathbf{w}^t \cdot \mathbf{w}^t}} \cdot \mathbf{w}^* \ge \frac{\sqrt{t}\rho}{R}$. *(4 points)*

(g) Hence, prove that $t \le \frac{R^2 \mathbf{w}^* \cdot \mathbf{w}^*}{\rho^2}$, i.e., the number of iterations t of PLA is bounded by some finite non-negative number. (Hint: $\frac{\mathbf{w}^t \cdot \mathbf{w}^*}{\sqrt{\mathbf{w}^t \cdot \mathbf{w}^t}\sqrt{\mathbf{w}^* \cdot \mathbf{w}^*}} \le 1$. Why?) *(5 points)*

Question 2. Automatic handwritten digit recognition is an important machine learning task. The US Postal Service Zip Code Database (http://www.unitedstateszipcodes.org/zip-code-database/) provides 16 × 16 pixel images preprocessed from scanned handwritten zip codes (US zip codes are the analogues of Singapore postal codes). The task is to recognize the digit in each image. We shall consider the simpler goal of recognizing only two digits: 1 and 5. To simplify our task even further, let's consider only two features: intensity and symmetry. Digit 5 generally occupies more black pixels and thus have higher average pixel intensity than digit 1. Digit 1 is usually symmetric but digit 5 is not. By defining asymmetry as the average difference between an image and its flipped versions, and symmetry as the negation of asymmetry, we can get higher symmetry values for digit 1. **Write a Python implementation of the pocket algorithm.** Train it on the training set (`train_1_5.csv`), and evaluate its accuracy on the test set (`test_1_5.csv`). The training and test sets are posted on eDimension. csv stands for comma-separated values. In the files, each row is an example. The first value is the symmetry, the second is the average intensity, and the third is the label. (The more adventurous among you can download the data from the website and preprocess the data on your own.)

(a) What is the best accuracy of your implemention on the test set? *(5 points)*

(b) At which iteration of the pocket algorithm is this accuracy achieved? *(5 points)*

(c) Submit your code by emailing to TA Dong Fei (fei_dong@sutd.edu.sg) with **crystal** clear instructions on how to run it. If he cannot run it, he has no choice but to give a zero grade. *(20 points)*

Question 3.

(a) You are given a dataset with two positive examples (1,1) and (2,2), and two negative examples (-1,1) and (1,-1). For each of the following hypothesis spaces, find the parameters of a member (classifier) of the hypothesis space that can correctly classify all the examples in the dataset, or explain why no such member exists.

i. Inside or outside of an origin-centered circle with radius r (r is the parameter). *(5 points)*

ii. Inside or outside of an (x,y)-centered circle with radius r (x, y, r are the parameters). *(5 points)*

iii. Above or below a line through the origin with normal (x,y). *(5 points)*

(b) Which of the above hypothesis spaces are linear classifiers? *(5 points)*

Question 4. Let's better understand the two common evaluation measures of *precision* and *recall*. Suppose that a machine learning algorithm can only make *postive/negative* (*true/false*, *yes/no*) predictions. Precision refers to the fraction of all positive predictions that the algorithm makes that are truly positive. Recall refers to the fraction of truly positive examples that the algorithm has predicted to be positive. We shall illustrate these concepts with the following situation. There are ten patients $p^1, p^2, \ldots, p^{10}$. Of these ten patients, p^1, p^2, p^3 and p^4 are known to have cancer; the other patients are cancer-free. A inexperienced junior doctor is asked to diagnose the patients for cancer, and he diagnoses p^1, p^2, p^7, p^8, p^9 to have cancer, and the rest to be cancer free. His precision is 2/5 (only two of the five people he predicted to have cancer truly have cancer). His recall is 2/4 (of the four people who have cancer, he only correctly predicted two of them). Explain how it is possible for the doctor to trade-off precision for recall, and vice versa? *(10 points)*

Question 5. Why is it a bad idea to measure classifier performance by using only the training set? *(10 points)*