EE3-23: Assignment 1

Spring, 2017

Instructions:

Solve all problems.

You need to submit a single zip file on Blackboard, named `assignment1_<name>.zip` where `<name>` is your name. The file should contain a report in PDF (write your name on your solution), named `assignment1_<username>_<name>.pdf` where `<username>` is your blackboard user name, and your code and other relevant files. If your code is in matlab or python, name it `assignment1_<username>.m` or `assignment1_<username>.py`, otherwise provide a make file which produces an executable `assignment1_<username>` on a standard linux system. Running your code should generate all the plots required in the solution. It should assume that the files required in Problem 3(c) are named `features.train, features.test, zip.train, zip.test`, as you can download them, and they are located in the directory `../data/` relative to the source code.

Justify your answers. You are allowed to consult with others (mention their names at the beginning of your solution), but you need to work out and write up your solution alone. The use of any written sources (e.g., books or the web) is permitted.

Include the following pledge into your answer:

I, <YOUR NAME>, pledge that this assignment is completely my own work, and that I did not take, borrow or steal work from any other person, and that I did not allow any other person to use, have, borrow or steal portions of my work. I understand that if I violate this honesty pledge, I am subject to disciplinary action pursuant to the appropriate sections of Imperial College London.

If you don't include your pledge, you will get 0%.

Deadline: February 6, 18:00. Late submissions are not accepted.

Problem 1. On Jun 9, 2016, ORB published a public poll result about Brexit: they asked 2,052 people, 55% of them supported leaving the EU while 45% wanted to stay. Assuming that the people's decisions are independent and identically distributed, with what confidence level can you reject the hypothesis that the probability of choosing to stay in the EU is at least 50% (i.e., the majority of people would like to stay in the EU). **[3 points]**

Hint: Use Hoeffding's inequality. For help, the one-sided version is provided here (this is more general then the one we learned in class). Let $a_1 \le b_1$, $a_2 \le b_2$, $\ldots$, $a_n \le b_n$ be real numbers, and let $X_1, X_2, \ldots, X_n$ be a sequence of independent real-valued random such that with probability one $X_t \in [a_t, b_t]$ for every $t = 1, 2, \ldots, n$. For any $\varepsilon \ge 0$,

$$\Pr\left(\sum_{t=1}^{n} X_t - \sum_{t=1}^{n} \mathbb{E}[X_t] \ge \varepsilon\right) \le \exp\left(\frac{-2\varepsilon^2}{\sum_{t=1}^{n}(b_t - a_t)^2}\right) .$$

Equivalently, for any $0 < \delta \le 1$, with probability at least $1 - \delta$,

$$\sum_{t=1}^{n} X_t < \sum_{t=1}^{n} \mathbb{E}[X_t] + \sqrt{\frac{1}{2}\ln(1/\delta)\sum_{t=1}^{n}(b_t - a_t)^2} .$$

Problem 2 (Convergence of the Perceptron Learning Algorithm)**.** Assume we are given n data points $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ for a binary classification problem, where $x^{(i)} \in \mathbb{R}^{d+1}, y^{(i)} \in \{1, -1\}, i = 1, \ldots, n$, and there exists $w^* \in \mathbb{R}^{d+1}$ such that $\mathrm{sign}({w_*}^\top x^{(i)}) = y^{(i)}$ for all $i \in \{1, \ldots, n\}$.[1] Consider the perceptron algorithm:

```
PERCEPTRON

Initialize $w_0 = \mathbf{0} \in \mathbb{R}^{d+1}$.
For each round $t = 1, 2, \ldots$
    • If there exists $(x_{t-1}, y_{t-1}) \in \mathcal{D}$ such that $\mathrm{sign}(w_{t-1}^\top x_{t-1}) \neq y_{t-1}$
        – set $w_t = w_{t-1} + y_{t-1} x_{t-1}$;
    • otherwise return $w_* = w_{t-1}$.
```

(a) Define $\rho = \min_{i \in \{1,\ldots,n\}} y^{(i)} w_*^\top x^{(i)}$. Show that $\rho > 0$. **[2 points]**

(b) Show that for any t before the algorithm stops, $w_*^\top w_t \geq t\rho$.

Hint: Show that $w_*^\top w_t \geq w_*^\top w_{t-1} + \rho$ and use induction. **[3 points]**

(c) Show that for any t before the algorithm stops, $\|w_t\|^2 \leq tR^2$, where $R = \max_{i \in \{1,\ldots,n\}} \|x^{(i)}\|$.

Hint: Show that $\|w_t\|^2 \leq \|w_{t-1}\|^2 + \|x_{t-1}\|^2$ and use induction. **[3 points]**

(d) Show that for any t before the algorithm stops, $t \leq \frac{R^2 \|w_*\|^2}{\rho^2}$, and conclude that the algorithm stops after at most $R^2 \|w_*\|^2 / \rho^2$ updates.

Hint: Use the Cauchy-Schwarz inequality: $a^\top b \leq \|a\| \|b\|$ for any $a, b \in \mathbb{R}^k$. **[4 points]**

Problem 3 (Experiments with the Perceptron Learning Algorithm)**.** This problem guides you through the steps of implementing a machine learning algorithm, testing some of its properties on carefully designed synthetic data, and finally using it on a real-life dataset.

(a) Implement the perceptron learning algorithm and test it on some small synthetic datasets (e.g., of size 2, 4, 10, 100, resp.). Describe (i) what method you use in the learning algorithm to select the next point to update; and (ii) the datasets: how you generated your data and why it is useful in testing the correctness of your implementation. **[2 points]**

Hint: Do not forget to add the extra coordinate for the offset.

(b) Consider the problem that $x \in \mathbb{R}^2$ is uniformly distributed in the unit square $[0, 1]^2$, and for any point $x \in [0, 1]^2$, the corresponding label is 1 if $x_2 - x_1 - 0.1 \geq 0$ and -1 otherwise.

(1) Give a closed form formula on the test error of any separator line $ax + b$ (any other characterization of a linear separator is acceptable). **[3 points]**

[1] Here $\mathrm{sign}(z)$ is defined the usual way, i.e., $\mathrm{sign}(z) = 1$ if $z > 0$, -1 if $z < 0$ and 0 if $z = 0$.

(2) Train your algorithm on $100, 200, \ldots, 500$ points and plot the resulting test error. **[2 points]**

(3) Are the above test error values random? Explain your answer. If yes, repeat it 100 times, plot the average and the 90% confidence intervals (leave out the best and worst 5% of the runs). **[1 point]**

(4) Repeat problems (b1)-(b3) when the classes are separated by some margin: using the same separator line as above, consider the case when the feature vectors are uniformly distributed over the region $K_\gamma = \{(x_1, x_2) \in [0,1]^2 : |x_2 - x_1 - 0.1| > \gamma\}$ for $\gamma = 0.3, 0.1, 0.01, 0.001$, respectively. Comment on how the test error depends on γ. **[4 points]**

Also, compare the number of updates to the theoretical bound of Problem 2. Find at least two viable choices for w_* in the definition of ρ. **[3 points]**

(c) Handwritten digit classification: Download the handwritten digit database provided with the assignment.

(1) Apply some modifications of the perceptron algorithm to learn a linear classifier between classes 2 and 8, using first the raw dataset and then the one with 2-D features. What modifications did you apply? **[2 points]**

(2) Compare the classification error on the training and test data of the two cases. Also, plot the change of the training and test error as a function of the number of updates of your algorithm and the original perceptron learning algorithm. **[3 points]**

(3) Compute the optimal linear regression weights (minimizing the squared error) for the training set using the 2-D features, and repeat (c2) with your learning algorithm initialized with the optimal linear regression weights. Compare with the previous results. **[3 points]**

(4) Assuming all data points are generated in an i.i.d. fashion, give a high-probability upper bound on the difference of the empirical test error and the ideal test error (defined as the expectation over the generating distribution). **[3 points]**