ML Foundation HW2 PA Report

B04505036
李慕家

---

6. (20 points) For Problems 7–8, you will play with the decision stump algorithm. In class, we taught about the learning model of "positive and negative rays" (which is simply one-dimensional perceptron) for one-dimensional data. The model contains hypotheses of the form:

$$h_{s,\theta}(x) = s \cdot \text{sign}(x - \theta).$$

The model is frequently named the "decision stump" model and is one of the simplest learning models. As shown in class, for one-dimensional data, the VC-Dimension of the decision stump model is 2.

In fact, the decision stump model is one of the few models that we could easily minimize E_{in} efficiently by enumerating all possible thresholds. In particular, for N examples, there are at most $2N$ dichotomies (see page 22 of lecture 5 slides), and thus at most $2N$ different E_{in} values. We can then easily choose the dichotomy that leads to the lowest E_{in}, where ties an be broken by randomly choosing among the lowest E_{in} ones. The chosen dichotomy stands for a combination of some "spot" (range of θ) and s, and commonly the median of the range is chosen as the θ that realizes the dichotomy.

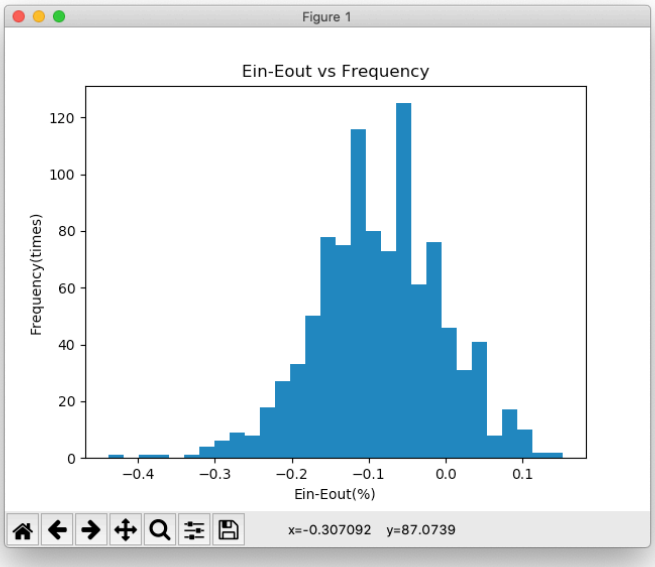
In the next problem, you are asked to implement such an algorithm and run your program on an artificial data set. We shall start by generating a one-dimensional data by the procedure below:

(a) Generate x by a uniform distribution in $[-1, 1]$.

(b) Generate y by $f(x) = \tilde{s}(x)$ + noise where $\tilde{s}(x) = \text{sign}(x)$ and the noise flips the result with 20% probability.

7. (20 points, *) Generate a data set of size 20 by the procedure above and run the one-dimensional decision stump algorithm on the data set. Record E_{in} and compute E_{out} with the formula above. Repeat the experiment 1000 times and plot a histogram of $E_{in} - E_{out}$. Describe your findings.

7.



```
Average Ein rate: 0.170%
Average Eout rate: 0.255%
Average Ein - Eout: -0.085%
```

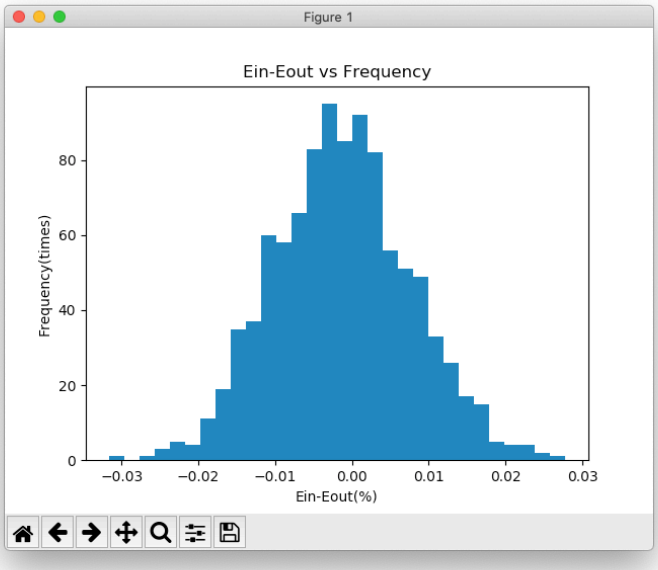
當資料量只有 20 筆時：

E_{out}明顯大於E_{in}，平均是E_{in}的 1.5 倍。$E_{in} - E_{out}$的值大多落在[-0.2,0.0]。

---

8. (20 points, *) Generate a data set of size 2000 by the procedure above and run the one-dimensional decision stump algorithm on the data set. Record E_{in} and compute E_{out} with the formula above. Repeat the experiment 1000 times and plot a histogram of $E_{in} - E_{out}$. Describe your findings and compare the findings with those in the previous problem.

8.




```
Average Ein rate: 0.199%
Average Eout rate: 0.201%
Average Ein - Eout: -0.001%
```

當資料量達到 2000 筆時：

$E_{out} \cong E_{in}$。相較前一題，當資料量由 20 筆增加至 2000 筆，E_{in}上升 17.1%，E_{out}下降 21.2%。

$E_{in} - E_{out}$的值大多落在[-0.015,0.010]。