TTIC 31020: Introduction to Machine Learning

Problem Set 1

Hung Le Tran

08 Jan 2024

Problem 1.1 (Problem 1)

Solution

(a) Medical diagnosis using X-ray images. The end system should be able to receive an X-ray image of a certain part of the body as input and output if the patient has/is likely to have a certain disease. The metric can be the accuracy of the model, that is the ratio of true positives and true negatives.

(b) Yes. Take the lungs for example. Perhaps an expert and a programmer can program where to identify common patterns that appear on a lung X-ray scan when a patient has lung cancer. Expert knowledge is required for where and in what shape these common patterns appear for the program to identify (through the pixel values).

(c) Database of lung X-ray scans of patients who have and do not have lung cancer, possibly along with other data points such as age, medical history, etc.

(d) Advantages: Early detection, mass detection, ease of distribution (e.g. to places where diagnosis might be limited, or expertise is not offered).

Problems: Reliability, privacy, might still need a human doctor to verify the prediction.

□

Problem 1.2 (Problem 2)

Solution

(a) We can use the **HALVING** learning rule. Its mistake bound is

$$M = \log_2|\mathcal{H}| = \log_2 d \ll d - 1$$

(b) We construct a sequence of labeled samples realized by $h_1(x) = \text{sgn}(x[1])$ (1-index). Fix d.

Let A_k be the set of tuples of length $(d-1)$ with k (-1)'s and $(d-1-k)$ 1's in any order. For instance, for $d = 4$,

$$A_2 = \{(-1, -1, 1), (-1, 1, -1), (1, -1, -1)\}$$

Then we can construct the following sample sequence as below:

```
arr = []
for k in range(0, d, 2):
    for s in A_k:
        arr.append((1, *s))
        arr.append((-1, *s))
```

In this sample sequence, we claim that every sample of the form $(-1, *s)$ is closest to the one immediately prior to it, namely $(1, s)$, since they only differ in 1 index. This is because $(-1, *s)$ differs in at least 2 indices with all other samples before it.

To show this, let us only consider previous samples starting with (-1); the same reasoning works for the other samples, even better.

Note that by the construction of A_k (and by only iterating through k even), $(-1, *s)$ differs from other $(-1, *s')$ for s' coming from an earlier $A_l (l \leq k-2)$ in at least 2 indices, because their number of (-1)'s differs by at least 2. If s' comes from the same A_k, s and s' are different strings so they must differ in at least 1 index. However, they have the same number of (-1)'s so this difference in 1 index implies a difference in at least another index, totaling up to at least 2. Our claim is thus demonstrated.

Coming back, this implies that our NN predictor would predict the label for **every** $(-1, *s)$ as the label of the one coming before it:

$$\hat{y}((-1, *s)) = h_1((1, *s)) = 1 \neq h_1((-1, *s))$$

thus making a mistake. The number of such mistakes is at least:

$$|A_0| + |A_2| + \ldots = \binom{d-1}{0} + \binom{d-1}{2} + \binom{d-1}{4} + \ldots = \frac{2^{d-1}}{2} = 2^{d-2} = 2^{\Omega(d)}$$

□