Assignment 4

COMP 599 Fall 2020 McGill University
Assignment prepared by: Prakash Panangaden

Due 4^{th} November 2020

Question 1[20 points] Find a hypothesis class and a sequence of examples so that the mistake bound of the halving algorithm is tight.

Question 2[30 points] For $i = 1, \ldots, n$, define $\vec{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$ by

$$\vec{x}_i = (\underbrace{(-1)^i, \ldots, (-1)^i, (-1)^{i+1}}_{i \text{ first components}}, 0, \ldots, 0) \quad \text{and} \quad y_i = (-1)^{i+1}\,.$$

Suppose that the perceptron algorithm is run cyclically over the sequence $S = (\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ until it makes no more mistakes. We are using perceptron in off-line mode. We run it again and again on this data *in the given order* until it makes no more mistakes.

Show that the total number of mistakes is *at least* 2^{n-3}.

Just to make sure that you understand the $\vec{x}_i$, here are the first few:
$\vec{x}_1 = (1, 0, 0, \ldots, 0)$
$\vec{x}_2 = (1, -1, 0, 0, \ldots, 0)$
$\vec{x}_3 = (-1, -1, 1, 0, 0, \ldots, 0)$
$\vec{x}_4 = (1, 1, 1, -1, 0, 0, \ldots, 0)$

Hint. Note that this algorithm will only ever assign integer values to the weights so it suffices to consider vectors in $\mathbb{Z}$. Let $\vec{w} = (w_1, \ldots, w_n) \in \mathbb{Z}^n$ be (a normal vector of) any linear separator. Give lower bounds on the magnitude of the w_i. Then argue that the n-th component is incrased by at most 1 in every update so the size of w_n gives your lower bound.

Question 3[50 points] In this question we will consider how to modify the perceptron algorithm to deal with non-separable data. We start by introducing an apparently new algorithm called the *dual perceptron*. Instead of a weight vector $\mathbf{w}$ we maintain a set of coefficients $\{\alpha_i\}_{i=1}^n$ where each α_i is $+1, 0$ or -1. The basic loop of the algorithm runs for $t = 1$ up to $t = T$:

- Receive $\mathbf{x}_t$. If $\displaystyle\sum_{i=1}^{i=t-1} \alpha_i(\mathbf{x}_i \cdot \mathbf{x}_t) \geq 0$ then predict $+1$, otherwise predict 1.
- Receive correct label y_t. If there is a mistake then $\alpha_t \leftarrow y_t$ otherwise $\alpha_t \leftarrow 0$.

Note that this algorithm makes *exactly the same* predictions as perceptron. However, instead of maintaining $\mathbf{w}$ explicitly we maintain it implicitly as the vector $\sum_{i=1}^{i=t-1} \alpha_i \mathbf{x}_i$.

Consider a sequence of labelled examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$. We do not assume that these data are necessarily separable. Fix $\gamma > 0$, you can think of this as the desired margin, if you like. Now to cope with the fact that there might not be a separating hyperplane we introduce a *loss function*. Suppose that we have a threshold function defined by a hyperplace with unit normal $\mathbf{u}$. We define the loss ℓ_t on input $(\mathbf{x}_t, y_t)$ as:

$$\ell_t = \max(0, \gamma - y_t(\mathbf{u} \cdot \mathbf{x_t})).$$

Notice that ℓ_t is non-negative and is zero if and only if the threshold function has margin at least γ at $(\mathbf{x}_t, y_t)$. The cumulative loss L of $\mathbf{u}$ over the sequence $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_T, y_T)$ is defined by $L^2 = \sum_{t=1}^{t=T} \ell_t^2$. Then $L = 0$ if and only if $\mathbf{u}$ represents a threshold function with margin at least γ on the above sequence. So the loss measures how far we are from our desired margin. We can now state our mistake bound, which specialises to the bound in the separable case by taking $L = 0$.

Prove the following statement:

> Suppose that there is a linear threshold function with squared loss L^2 on $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_T, y_T)$. We assume that the norms are bounded: there exists $D > 0$, such that $\|\mathbf{x}_t\|^2 \leq D$ for all t. Then the perceptron algorithm makes at most $(\frac{D+L}{\gamma})^2$ mistakes.

Hint: Embed the data in a higher-dimensional space $\mathbf{R}^{n+T}$ by changing each vector $\mathbf{x}_t$ to a new vector

$$\tilde{\mathbf{x}}_t = (\mathbf{x}_t, 0, 0, \ldots, \Delta, 0, \ldots, 0),$$

where $\Delta > 0$ is some fixed positive number. Show that if the labels are not changed the predictions are not changed. Show how to define a modified version of $\mathbf{u}$ so that the new data are separable with a suitable margin (which won't just be γ). Now use the separable mistake bound to derive a mistake bound in terms of γ, Δ, L. Show that by choosing Δ to minimise this expression we get the claimed mistake bound.