Problem Solutions

e-Chapter 7

Pierre Paquay

Problem 7.1

To solve this problem, we first begin by separating the positive decision region into two components : the lower one corresponding to $x_2 \in [-1, 1]$ and the upper one corresponding to $x_2 \in [1, 2]$. To define the decision region, we need 7 perceptrons, namely

$$h_1(x) = \text{sign}(x_2 - 2),\ h_2(x) = \text{sign}(x_2 - 1),\ h_3(x) = \text{sign}(x_2 + 1),$$

for the horizontal lines, and

$$h_4(x) = \text{sign}(x_1 + 2),\ h_5(x) = \text{sign}(x_1 + 1),\ h_6(x) = \text{sign}(x_1 - 1),\ h_7(x) = \text{sign}(x_1 - 2)$$

for the vertical lines. We are now able to define the lower decision region by $\overline{h_2} h_3 h_4 \overline{h_7}$, and the upper decision region by $\overline{h_1} h_2 h_5 \overline{h_6}$, which means that the total decision region is defined by

$$f = \overline{h_2} h_3 h_4 \overline{h_7} + \overline{h_1} h_2 h_5 \overline{h_6}$$

which actually characterizes a 3-layer perceptron.

Problem 7.2

(a) Let x and x' be two points from the same region. If we consider a set of M hyperplanes defined by $\{x : w_i^T x = 0\}$, we have that

$$(\text{sign}(w_1^T x), \cdots, \text{sign}(w_M^T x)) = (\text{sign}(w_1^T x'), \cdots, \text{sign}(w_M^T x'));$$

or put more simply that $\text{sign}(w_i^T x) = \text{sign}(w_i^T x') = s_i$ for $i = 1, \cdots, M$ where $s_i = \pm 1$. We begin by the case where $s_i = 1$. Here, we know that $w_i^T x > 0$ and $w_i^T x' > 0$, consequently we have that, for $\lambda \in [0, 1]$,

$$w_i^T(\lambda x + (1 - \lambda)x') = \lambda w_i^T x + (1 - \lambda) w_i^T x' > 0$$

and

$$\text{sign}(w_i^T(\lambda x + (1 - \lambda)x')) = 1.$$

Now, we consider the case where $s_i = -1$. Here, we know that $w_i^T x < 0$ and $w_i^T x' < 0$, consequently we have that, for $\lambda \in [0, 1]$,

$$w_i^T(\lambda x + (1 - \lambda)x') = \lambda w_i^T x + (1 - \lambda) w_i^T x' < 0$$

and

$$\text{sign}(w_i^T(\lambda x + (1 - \lambda)x')) = -1.$$

So, in conclusion, the region is actually convex.

(b) A region is defined as the following set

$$\{x : (\text{sign}(w_1^T x), \cdots, \text{sign}(w_M^T x)) = (s_1, \cdots, s_M); s_i \in \{-1, 1\}\};$$

thus a region is characterized by a particular M-uple $(s_1, \cdots, s_M)$. Since there are at most 2^M of such M-uples, we have at most 2^M different regions.

(c) Let $B(N, d)$ be the maximum number of regions created by M hyperplanes in d-dimensional space. Now, consider adding an $(M + 1)$th hyperplane; this hyperplane can obviously be viewed as a $(d - 1)$-dimensional space, so if we project the initial M hyperplanes into this space, we obtain M hyperplanes in a $(d-1)$-dimensional space. These hyperplanes can create at most $B(M, d-1)$ regions in this space, and for each of these regions, we get two regions in the original d-dimensional space. Thus, this means that the $(M+1)$th hyperplane intersects at most $B(M, d-1)$ of the regions created by the M hyperplanes in the d-dimensional space, and so

$$B(M+1, d) \leq B(M, d) + B(M, d-1).$$

Now, we will prove that

$$B(M, d) \leq \sum_{i=0}^{d} \binom{M}{i}$$

by induction. We begin by evaluating the boundary conditions, we have

$$B(M, 1) = M + 1 \leq \sum_{i=0}^{1} \binom{M}{i} = \binom{M}{0} + \binom{M}{1} = M + 1$$

for all M, and

$$B(1, d) = 2 \leq \sum_{i=0}^{d} \binom{1}{i} = \binom{1}{0} + \binom{1}{1} = 2$$

for all d. Now, we assume the statement is true for $M = M_0$ and all d, we will prove that the statement is still true for $M = M_0 + 1$ and all d. We have that

$$\begin{aligned}
B(M_0+1, d) &\leq B(M_0, d) + B(M_0, d-1) \\
&\leq \sum_{i=0}^{d} \binom{M_0}{i} + \sum_{i=0}^{d-1} \binom{M_0}{i} \\
&= \binom{M_0}{0} + \sum_{i=1}^{d} \binom{M_0}{i} + \sum_{i=1}^{d} \binom{M_0}{i-1} \\
&= 1 + \sum_{i=1}^{d} \underbrace{\left[\binom{M_0}{i} + \binom{M_0}{i-1} \right]}_{=\binom{M_0+1}{i}} \\
&= \sum_{i=0}^{d} \binom{M_0+1}{i}.
\end{aligned}$$

We have thus proved the induction step, so the statement is true for all M and d.

Problem 7.3

We begin by proving the following equivalence relation

$$h_m(x) = c_m \Leftrightarrow h_m^{c_m}(x) = +1.$$

The condition is necessary because if $c_m = +1$, we have

$$h_m^{c_m}(x) = h_m(x) = c_m = +1;$$

and if $c_m = -1$, we have

$$h_m^{c_m}(x) = \overline{h}_m(x) = \overline{c}_m = +1.$$

Now the condition is also sufficient because if $c_m = +1$, we have

$$+1 = h_m^{c_m}(x) = h_m(x),$$

which means that $h_m(x) = +1 = c_m$; and if $c_m = -1$, we have

$$+1 = h_m^{c_m}(x) = \overline{h}_m(x),$$

which implies that $h_m(x) = -1 = c_m$.

Now we are able to write that

$$\begin{array}{rc} & x \in r \\ \Leftrightarrow & (h_1(x), \cdots, h_M(x)) = (c_1, \cdots, c_M) \\ \Leftrightarrow & h_m^{c_m}(x) = +1, \ \forall m \\ \Leftrightarrow & \prod_{m=1}^M h_m^{c_m}(x) = +1 \\ \Leftrightarrow & t_r(x) = +1. \end{array}$$

The above relation also implies that

$$x \notin r \Leftrightarrow t_r(x) = -1.$$

Now if x is in a positive region ($f(x) = +1$), we know that there exists i such that $x \in r_i$, and consequently that $t_{r_i}(x) = +1$ which means that

$$t_{r_1}(x) + \cdots + t_{r_k}(x) = +1 = f(x).$$

And if x is in a negative region ($f(x) = -1$), we know that $x \notin r_i$ for all i, so $t_{r_i}(x) = -1$ for all i which means that

$$t_{r_1}(x) + \cdots + t_{r_k}(x) = -1 = f(x).$$

Problem 7.4

Since $f = t_{r_1} + \cdots + t_{r_k}$, we may write that

$$f = \text{sign}(k - \frac{1}{2} + \sum_{i=1}^{k} t_{r_i}),$$

which characterizes the penultimate layer of our perceptron. For the layer before, we have that $t_{r_i} = h_1^{c_1^{(i)}} \cdots h_M^{c_M^{(i)}}$, and consequently

$$t_{r_i} = \text{sign}(-M + \frac{1}{2} + \sum_{m=1}^{M} h_m^{c_m^{(i)}});$$

moreover, the previous layer may be characterized with

$$h_m^{c_m^{(i)}} = \text{sign}(c_m^{(i)} w_m^T x).$$

Putting all this together, we obtain the following characterization of a 3-layer perceptron

$$f = \text{sign}(k - \frac{1}{2} + \sum_{i=1}^{k} \text{sign}(-M + \frac{1}{2} + \sum_{m=1}^{M} \text{sign}(c_m^{(i)} w_m^T x)))$$

whose structure is given by $[d, kM, k, 1]$.

Problem 7.5

First, we decompose the unit hypercube $[0, 1]^d$ into $1/\epsilon^d$ ϵ-hypercubes (hypercube whose sides have length equal to ϵ), thus we get a grid-like structure of our unit hypercube. Now, if we consider a decision region (which may be composed by disconnected regions) whose boundary surfaces are smooth, this decision region partition the unit hypercube into two regions : one labelled $+1$ and one labelled -1. We now have k_ϵ ϵ-hypercubes labelled $+1$ which are formed by $2d$ hyperplanes each defined by $h_m^{(i)} = \text{sign}(w_m^{(i),T} x)$ where $m = 1, \cdots, 2d$ and $i = 1, \cdots, k_\epsilon$. So, the first layer whose task is to activate the hyperplanes involved in the positive ϵ-hypercubes is characterized by

$$h_m^{(i)} = \text{sign}(w_m^{(i),T} x).$$

Now to activate the positive ϵ-hypercubes H_i themselves we characterize the second layer by

$$t_{H_i} = (h_1^{(i)})^{c_1^{(i)}} \cdots (h_{2d}^{(i)})^{c_{2d}^{(i)}},$$

where the $c_m^{(i)}$ are defined as in Problem 7.3 and 7.4; or

$$t_{H_i} = \text{sign}(-2d + \frac{1}{2} + \sum_{m=1}^{2d} (h_m^{(i)})^{c_m^{(i)}}).$$

And finally to activate all the positive ϵ-hypercubes, we define the MLP output h by

$$h = t_{H_1} + \cdots + t_{H_{k_\epsilon}};$$

or

$$h = \text{sign}(k_\epsilon - \frac{1}{2} + \sum_{i=1}^{k_\epsilon} t_{H_i}).$$

Putting all this together, we obtain the following characterization of a 3-layer perceptron

$$h = \text{sign}(k_\epsilon - \frac{1}{2} + \sum_{i=1}^{k_\epsilon} \text{sign}(-2d + \frac{1}{2} + \sum_{m=1}^{2d} \text{sign}(c_m^{(i)} w_m^{(i),T} x))).$$

Now, it remains to see that the above MLP can aribtrarily closely approximate the initial positive decision region D_+ (and consequently the negative decision region also); to do so, we first note that

$$\text{Vol}(H_i) = \epsilon^d \to 0 \text{ and } k_\epsilon \to \infty$$

when $\epsilon \to 0$. So, the ϵ-hypercubes can be made arbitrarily small, which obviously means that the total volume of the positive ϵ-hypercubes can be made arbitrarily close to the volume of the positive decision region (because of its smoothness). Mathematically, we may write that

$$\text{Vol}(H_1 \cup \cdots \cup H_{k_\epsilon}) = \sum_{i=1}^{k_\epsilon} \epsilon^d \to \text{Vol}(D_+)$$

when $\epsilon \to 0$. This means that the region where our 3-layer perceptron will output $+1$ (resp. -1) converges to the positive (resp. negative) decision region in our unit hypercube.

Problem 7.6

For a specific layer l, if we replace the weight $w_{ij}^{(l)}$ with $w_{ij}^{(l)} + \epsilon$, we need to recompute the corresponding node output of that layer and also the node outputs for the subsequent layers (which are the ones numbered from $l + 1$ to L). Consequently, for each weight $w_{ij}^{(l)}$, we have

$$\sum_{k=l+1}^{L} d^{(l)}(d^{(l-1)} + 1) + 1 + \sum_{k=l+1}^{L} d^{(l)}$$

multiplications and θ-evaluations respectively; this means that the computational complexity of obtaining the partial derivatives is overall equal to

$$
\begin{aligned}
&\underbrace{2\sum_{l=1}^{L} d^{(l)}(d^{(l-1)}+1)}_{=|W|}\left(\sum_{k=l+1}^{L} d^{(l)}(d^{(l-1)}+1)+1+\sum_{k=l+1}^{L} d^{(l)}\right) \\
\leq\quad & 2|W|\left(\underbrace{\sum_{k=1}^{L} d^{(l)}(d^{(l-1)}+1)}_{=|W|}+1+\sum_{k=1}^{L} \underbrace{d^{(l)}}_{\leq d^{(l)}(d^{(l-1)}+1)}\right) \\
\leq\quad & 2|W|(2|W|+1)=O(|W|^2)
\end{aligned}
$$

since we need to compute the derivatives corresponding to $w_{ij}^{(l)}+\epsilon$ and also to $w_{ij}^{(l)}-\epsilon$.