

# Tutorial 3

CS 337 Artificial Intelligence & Machine Learning, Autumn 2021

September, 2021

**Problem 1.** Given a set of data points  $\{x_n\}$  (for  $n = 1 \dots N$ ) in some  $d$  dimensional space ( $\mathbb{R}^d$ ), we can define the convex hull to be the set of all points  $x$  given by

$$x = \sum_n \alpha_n x_n$$

where  $\alpha_n \geq 0$  and  $\sum_n \alpha_n = 1$ . Consider a set of points  $\{x_m\}$  together with its convex hull and a second set of points  $\{y_m\}$  together with its corresponding convex hull. By definition, the two sets of points will be linearly separable if there exists a vector  $\hat{w}$  and a scalar  $w_0$  such that  $\mathbf{w}^T x_n + w_0 > 0$  for all  $x_n$  and  $\mathbf{w}^T y_m + w_0 < 0$  for all  $y_m$ . Show that if the convex hulls of  $\{x_n\}$  and  $\{y_n\}$  intersect, the two sets of points cannot be linearly separable.

**Optional and advanced:** Also prove the converse statement that if they are not linearly separable, their convex hulls must intersect.

**Problem 2.** Let  $X$  have a uniform distribution over integers in an interval  $[0, \theta)$ :

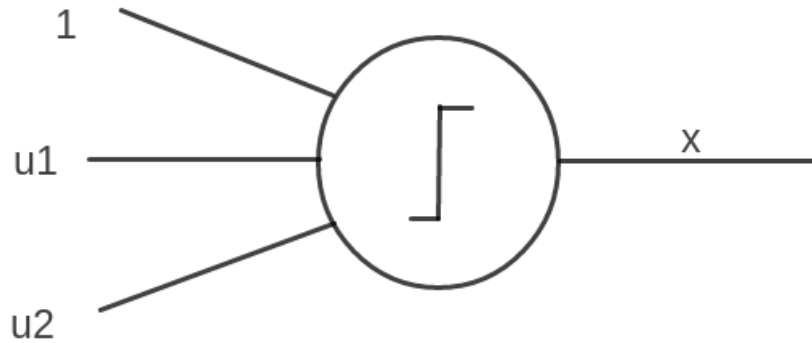
$$p(X = x; \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x < \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose  $n$  samples  $x_1, \dots, x_n$  are drawn i.i.d based on  $p(x; \theta)$ . What is the MLE estimate of  $\theta$ ?

**Problem 3. Computing power of perceptrons.** Perceptrons can only separate linearly separable data as discussed in class. Given  $n$  variables we can have  $2^{2^n}$  boolean functions, but not all of these can be represented by a perceptron. For example when  $n=2$  the XOR and XNOR cannot be represented by a perceptron. Given  $n$  boolean variables how many of  $2^{2^n}$  boolean functions can be represented by a perceptron?

**Problem 4.** Consider a perceptron for which  $u \in \mathbb{R}^2$  and

$$f(a) = \begin{cases} 1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$



Let the desired output be 1 when elements of class  $A = \{(1,2), (2,4), (3,3), (4,4)\}$  is applied as input and let it be -1 for the class  $B = \{(0,0), (2,3), (3,0), (4,2)\}$ . Let the initial connection weights  $w_0(0) = +1, w_1(0) = -2, w_2(0) = +1$  and learning rate be  $\eta = 0.5$ .

This perceptron is to be trained by perceptron convergence procedure, for which the weight update formula is  $w(t+1) = w(t) + \eta(y^k - x^k(t))u^k$

1. (a) Mark the elements belonging to class A with x and those belonging to class B with o on input space.  
 (b) Draw the line represented by the perceptron considering the initial connection weights  $w(0)$ .  
 (c) Find out the regions for which the perceptron output is +1 and -1  
 (d) Which elements of A and B are correctly classified, which elements are misclassified and which are unclassified?
2. If  $u=(4,4)$  is applied at input, what will be  $w(1)$  ?
3. Repeat a) considering  $w(1)$ .
4. If  $u=(4,2)$  is then applied at input, what will be  $w(2)$ ?

5. Repeat 1) considering  $w(2)$ .
6. Do you expect the perceptron convergence procedure to terminate? Why?

**Problem 5.** In the class, we discussed the probabilistic binary (class) logistic regression classifier. How will you extend logistic regression probabilistic model to multiple (say  $K$ ) classes? Are there different ways of extending? What is the intuition behind each? Discuss and contrast advantages/disadvantages in each.