

Homework VI-KEY

1. Consult the AdaBoost algorithm given in Bishop Chapter 14. Suppose there are two weak learners h_1 and h_2 , and a set of 17 points.

a) Let say h_1 makes one mistake and h_2 makes four mistakes on the dataset. Which learner will AdaBoost choose in the first iteration (namely $m=1$)?

Will choose h_1 .

b) What is α_1 ?

$$\epsilon_m = 1/17. \alpha_m = \ln\{(1 - 1/17)/1/17\} = \ln(16)$$

c) Calculate the data weighting co-efficient w_2 for the following two cases (1) the points on which chosen learner made a mistake and (2) the points on which the chosen learner did not make a mistake. **[10 Points]**

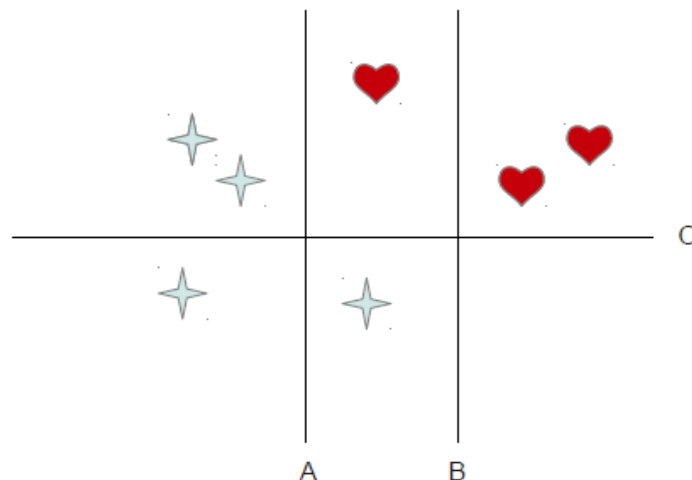
Case 1: Error made

$$w_2 = 1/17 \times 16 = 16/17$$

Case 2: No Error

$$w_2 = 1/17.$$

2. The diagram shows training data for a binary concept where a heart denotes positive examples. Also shown are three decision stumps (A, B and C) each of which consists of a linear decision boundary. Suppose that AdaBoost chooses A as the first stump in an ensemble and it has to decide between B and C as the next stump. Which will it choose? Explain. What will be the ϵ and α values for the first iteration? **[5 Points]**



It will choose B because the only example mis-classified by A is correctly classified by B (the misclassified examples are assigned higher weight in the next iteration). B also makes another one error C makes 2 errors.

In the first iteration $\epsilon = 1/7$

and $\alpha = \ln\left\{\frac{1-\epsilon}{\epsilon}\right\} = \ln(6)$

3. Consider cluster 1D data with a mixture of 2 Gaussian using the EM algorithm. You are given the 1D data points $x = [1 \ 10 \ 20]$. Suppose the output of the E step is the following matrix

$$R = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

where entry $r_{i,c}$ is the probability of observation x_i belonging to cluster c (the responsibility of cluster c for data point i). You have to compute the M step. You may state the equations for maximum likelihood estimates of these quantities (which you should know) without proof; you just have to apply the equations to this data set. You may leave your answer in fractional form.

- Write down the likelihood function you are trying to optimize.
- After performing M step for the missing weights π_1, π_2 , what are the new values?
- After performing M step for the means μ_1, μ_2 , what are the new values?

[10 Points]

a. Likelihood

$$p(D|\theta) = \prod_{i=1}^3 \sum_{k=1}^2 \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k)$$

b. Mixing Wights

$$\pi_1 = \frac{(1+0.4+0)}{3} = \frac{1.4}{3} = 0.46$$

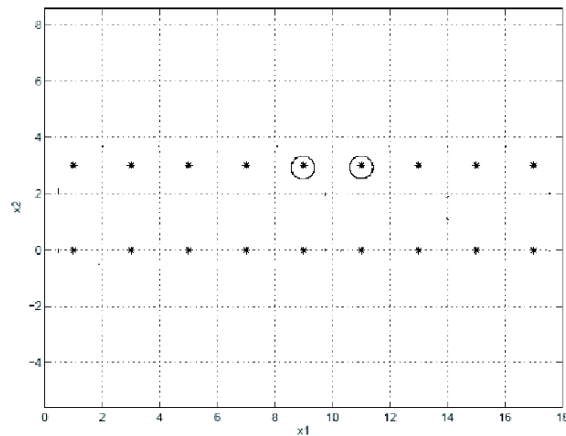
$$\pi_2 = \frac{(0+0.6+1)}{3} = \frac{1.6}{3} = 0.53$$

c. Means

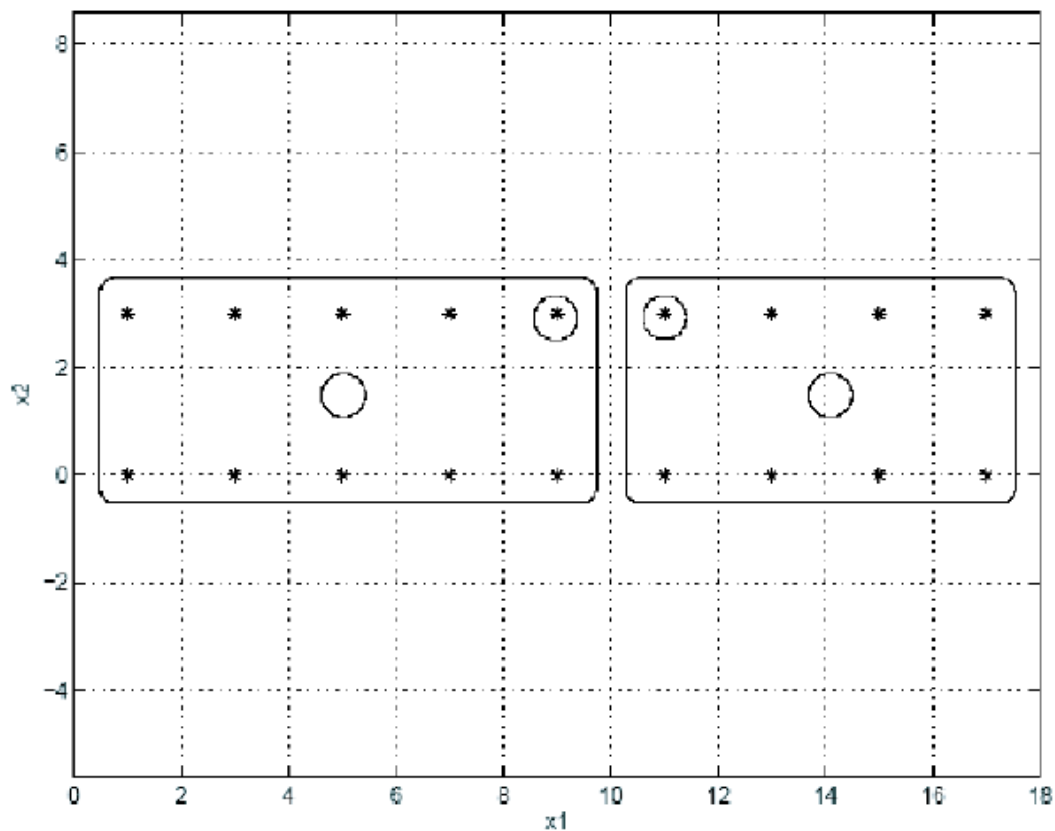
$$\mu_1 = \frac{(1)1 + 0.4(10)}{1.4} = \frac{5}{1.4} = 3.57$$

$$\mu_2 = \frac{0.6(10) + 1(20)}{1.6} = \frac{26}{1.6} = 16.25$$

4. In the following figure some data points are shown which lie on integer grid. Suppose we apply the K-means algorithm to this data, using $K = 2$ and with the centers initialized at the two circled data points. Draw the final clusters obtained after K-means converges. [5 Points]



Solution: K-means algorithm converges in 2 steps. The following figure shows the final means and clusters.



5. Consider training a two-input perceptron. Give an upper bound on the number of training examples sufficient to assure with 90% confidence that the learned perceptron will have true error of at most 5%? **[5 Points]**

$$m \geq \frac{1}{\epsilon} (4 \log_2(2/\delta) + 8VC(H) \log_2(13/\epsilon))$$

that makes

$$m \geq \frac{1}{0.05} \left(4 * \log_2\left(\frac{2}{0.1}\right) + 8 * 3 * \log_2\left(\frac{13}{0.05}\right) \right)$$

$$m \geq 20 * (4 * 4.32193 + 8 * 3 * 8.0224) = 4196.496$$

$$m \geq 4197$$

6. The VC dimension is always less than size of the hypothesis space. True/False?

[5 Points]

True VC dimension of a hypothesis is at most the log of the hypothesis space.

7. Computational Learning Theory

[10 Points]

- (a) Consider the class C of concepts of the form: $(a \leq x_1 \leq b) \wedge (c \leq x_2 \leq d)$. Note that each concept in this class corresponds to a rectangle in 2-dimensions. Let a, b be integers in the range $[0, 199]$ and c, d be integers in the range $[0, 99]$. Give an upper bound on the number of training examples sufficient to assure that for any target concept $c \in C$, any consistent learner using $H = C$ will, with probability 0.99, output a hypothesis with error at most 0.05.

Since, the learner is consistent, we will use the formula
 $m \geq (1/\epsilon) * (\ln(1/d) + \ln |H|)$ to get a bound on m

where, $d = 0.01$ and $\epsilon = 0.05$

$|H|$ is the number of rectangles = $[(n_1 * (n_1 - 1)) / 2] * [(n_2 * (n_2 - 1)) / 2]$ where n_1 is 200 and n_2 is 100.

$$|H| = (200 * 199 * 100 * 99) / 4 = 98505000$$

Therefore, $m \geq (1/0.05) * (\ln(1/0.01) + \ln(98505000))$

$$m \geq 460.216$$

Number of training examples sufficient to satisfy the required conditions is 461.

- (b) Consider the class C of concepts of the form: $(a \leq x_1 \leq b) \wedge (c \leq x_2 \leq d) \wedge (e \leq x_3 \leq f)$. Note that each concept in this class corresponds to a hyper-rectangle in 3-d. Now suppose that a, b, c, d, e, f take on real values instead of integers. Give an upper bound on the number of training examples sufficient to assure that for any target concept $c \in C$, a learner will, with probability 0.95, output a hypothesis with error at most 0.01.

We will use the formula

$$m \geq (1/\epsilon) * [4 * \log_2(2/d) + 8 * VC(H) * \log_2(13/\epsilon)]$$

where, $d = 0.05$, $\epsilon = 0.01$

$$VC(H) = 2 * \text{Number of dimensions} = 2 * 3 = 6$$

$$m \geq (1/0.01) * [4 * \log_2(2/0.05) + 8 * 6 * \log_2(13/0.01)]$$

$$m \geq (1/0.01) * [4 * \log_2(40) + 48 * \log_2(1300)]$$

$$m \geq 100 * (4 * 5.3219 + 48 * 10.3443)$$

$$m \geq 51781.4$$

Number of training examples sufficient to satisfy the required conditions is 51782.