

# Learning From Data - worked examples

Mac Radigan

## Chapter 1 Notes

Symbol	Name
$\mathcal{X}$	Input Set
$\mathcal{D}$	Training Set
$\mathcal{Y}$	Target Set
$f$	Target Function
$\mathcal{H}$	Hypothesis Set
$\mathcal{A}$	Learning Algorithm
$M$	Cardinality of Hypothesis Set
$g$	Final Hypothesis
$\epsilon$	Residual Error

### Relations and Sets

$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

$$\{h_1, h_2, \dots, h_M\} \in \mathcal{H}$$

### Hoeffding Inequality

$$P[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N} \forall \epsilon > 0 \quad (1)$$

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N} \forall \epsilon > 0 \quad (2)$$

For the final hypothesis  $g$  selected from  $\mathcal{H}$ ,

$$\begin{aligned}
P[|E_{in}(g) - E_{out}(g)| > \epsilon] &\leq P[|E_{in}(h_1) - E_{out}(h_1)| > \epsilon] \\
&\text{or} \leq P[|E_{in}(h_1) - E_{out}(h_1)| > \epsilon] \\
&\dots \\
&\text{or} \leq P[|E_{in}(h_M) - E_{out}(h_M)| > \epsilon] \\
&\leq \sum_{m=1}^M P[|E_{in}(h_m) - E_{out}(h_m)| > \epsilon] \\
&\leq 2Me^{-1\epsilon^2 N}
\end{aligned} \tag{3}$$

### In-sample Error

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^N P[h(x_n) \neq f(x_n)] \text{ fraction of D where f \& h disagree} \tag{4}$$

### Out-of-sample Error

$$E_{out}(h) = P[h(x) \neq f(x)] \tag{5}$$

## Chapter 1 Additional Notes

### Markov's Inequality

$$P[x \geq \alpha] \leq \frac{E(x)}{\alpha} \text{ for } \alpha > 0$$

Proof:

$$\begin{aligned}
E(x) &= \int xP(x)dx \\
&= \int_0^\alpha xP(x)dx + \int_\alpha^\infty xP(x)dx \\
&\geq \int_\alpha^\infty xP(x)dx \\
&\geq \int_\alpha^\infty \alpha P(x)dx \\
&= \alpha \int_\alpha^\infty P(x)dx \\
&= \alpha P[x \geq \alpha]
\end{aligned} \tag{6}$$

## Information Theory

In information theory, the analog of the law of large numbers is the Asymptotic Equipartition Property (AEP).

The Law Of Large Numbers states that for Independent, Identically Distributed (*i.i.d.*) random variables states:

$$\frac{1}{N} \sum_{k=1}^N X_k \rightarrow E(N) \text{ for sufficiently large } N$$

The Asymptotic Equipartition Property (AEP) states:

$$\frac{1}{N} \log_2 \frac{1}{p(X_1, X_2, \dots, X_N)} \rightarrow H(X) \text{ where } H \text{ is the entropy, } X_k \text{'s are } i.i.d. \text{ and } p(X_1, X_2, \dots, X_N) \text{ is the probability of observing the sequence } X_1, X_2, \dots, X_N.$$

This enables us to divide the set of all sequences into two sets, the *typical set*, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences.

## Entropy

The entropy is a measure of the average uncertainty in a random variable.

The entropy of a random variable  $X$  with a probability mass function  $p(x)$  is defined by

$$H(X) = - \sum_x p(x) \log_2 p(x) \quad (7)$$

## Asymptotic Equipartition Property (AEP)

If  $X_1, X_2, \dots, X_N$  are *i.i.d*  $\sim p(x)$ , then

$$\frac{1}{N} \log_2 p(X_1, X_2, \dots, X_N) \rightarrow H(X) \text{ in probability} \quad (8)$$

Proof:

$$\begin{aligned} \frac{1}{N} \log_2 p(X_1, X_2, \dots, X_N) &= -\frac{1}{N} \sum_k \log_2 p(X_k) \\ &\rightarrow -E \log_2 p(X) \\ &\rightarrow H(X) \end{aligned}$$

**Typical Set**  $A_\epsilon^{(n)}$ 

A *typical set*  $A_\epsilon^{(n)}$  with respect to  $p(x)$  is the set of sequences  $(x_1, x_2, \dots, x_N) \in \mathcal{X}^N$  with the property

$$2^{-N(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_N) \leq 2^{-N(H(X)-\epsilon)}$$

**Problem 1.1**

We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black and a white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball it is black. You now pick the second ball from that same bag. What is the probability that this ball is also black? [Hint: Use Bayes' Theorem:  $P[A \text{ and } B] = P[A|B]P[B] = P[B|A]P[A]$ .]

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

- $B_A$  the event that bag A was selected
- $B_B$  the event that bag B was selected
- $k_A$  the event that a black ball from bag A was selected
- $k_B$  the event that a black ball from bag B was selected
- $k_1$  the event that a black ball was selected on the first selection
- $k_2$  the event that a black ball was selected on the second selection

$$k_1 = B_A k_A + B_B k_B = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{4}$$

$$P(B_A|k_1) = \frac{P(B_A \cap k_1)}{P(k_1)} = \frac{P(k_1|B_B)P(B_A)}{P(k_1)}$$

$$P(B_B|k_1) = \frac{P(B_B \cap k_1)}{P(k_1)} = \frac{P(k_1|B_B)P(B_B)}{P(k_1)}$$

$$P(k_1|B_A) = \frac{1}{2} \cdot 1 = \frac{1}{2}$$

$$P(k_1|B_B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$P(B_A) = \frac{1}{2}$$

$$P(B_B) = \frac{1}{2}$$

$$P(k_1) = \frac{3}{4}$$

$$P(B_A|k_1) = \frac{P(k_1|B_A)P(B_A)}{P(k_1)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

$$P(B_B|k_1) = \frac{P(k_1|B_B)P(B_B)}{P(k_1)} = \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{\frac{1}{8}}{\frac{3}{4}} = \frac{4}{24} = \frac{1}{6}$$

$$P(k_2) = P(k_1|B_B)P(k_B) + P(B_B|k_1)P(k_B) = \frac{1}{3} \cdot 1 + \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{3} + \frac{1}{12} + \frac{4}{12} + \frac{1}{12} = \frac{5}{12}$$

## Problem 1.2

Consider the perceptron in two dimensions:  $h(x) = \text{sign}(w^\top x)$  where  $w = [w_0, w_1, w_2]$  and  $x = [1, x_1, x_2]$ . Technically,  $x$  has three coordinates, but we call this perceptron two-dimensional because the first coordinate is fixed at 1.

- (a) Show that the regions on the plane where  $h(x) = +1$  and  $h(x) = -1$  are separated by a line. If we express this line by the equation  $x_2 = ax_1 + b$ , what are the slope  $a$  and intercept  $b$  in terms of  $w_0, w_1, w_2$ ?

$$w^\top x = 0$$

$$\implies w_0x_0 + w_1x_1 + w_2x_2 = 0$$

$$\implies w_2x_2 = -w_1x_1 - w_0x_0 = -w_1x_1 - x_0$$

$$\implies x_2 = -\frac{w_1}{w_2}x_1 - \frac{w_0}{w_2}$$

$$a = -\frac{w_1}{w_2}$$

$$b = -\frac{w_0}{w_2}$$

- (b) Draw a picture for the cases  $w = [1, 2, 3]$  and  $w = -[1, 2, 3]$ .

In more than two dimensions, the  $+1$  and  $-1$  regions are separated by a hyperplane, the generalization of a line.

## Problem 1.3

Prove that the PLA eventually converges to a linear separator for separable data. The following steps will guide you through the proof. Let  $w^*$  be an optimal set of weights (one which separates the data). The essential idea in this proof is to show that the PLA weights  $w(t)$  get “more aligned” with  $w^*$  with every iteration. For simplicity, assume that  $w(0) = 0$ .

- (a) Let  $\rho = \min_{1 \leq n \leq N} (w^{*\top} x_n)$ . Show that  $\rho > 0$ .

We know  $w^{*\top} x_n > 0 \forall n$ , since  $x$  is linearly separable, and  $w^*$  separates  $x$ .

So by definition,

$$w^{*\top} x_n > 0 \quad \forall n \text{ s.t. } y_n = +1$$

and

$$w^{*\top} x_n < 0 \quad \forall n \text{ s.t. } y_n = -1$$

and thus

$$y_n (w^{\top} x_n) \forall n$$

that is,  $\rho$  is strictly positive.

(b) Show that  $w^{\top}(t)w^* \geq w^{\top}(t-1)w^* + \rho$ , and conclude that  $w^{\top}(t) \geq t\rho$ .

Let  $x'$  be the set of misclassified points, and let  $y'$  be the corresponding truth values

Further, let  $\rho_m = \min_m y_m (w^{\top} x_m)$

Then  $w_t = \sum_{m=0}^t y'_m x'_m$

But since  $(y'_m x'_m) \leq (y'_t x'_t)^{\top} w^* \forall m, t$

then

$$\left( \sum_{m=0}^t y'_m x'_m \right) w^* \geq \left( \sum_{m=0}^{t-1} y'_m x'_m \right)^{\top} w^* + y_t (x_t^{\top} w^*)$$

and thus

$$w^{\top}(t) \geq t\rho$$

(c) Show that  $\|w(t)\|^2 \leq \|w(t-1)\| + \|w(t-1)\|^2$ .

[Hint:  $y(t-1) \cdot (w^{\top}(t-1)x(t-1)) \leq 0$  because  $x(t-1)$  was misclassified by  $w(t-1)$ .]

Note  $w_t = w_{t-1} + y_{t-1}x_{t-1}$

so

$$\|w_{t-1} + y_{t-1}x_{t-1}\|^2 \leq \|w_{t-1}^2\|^2 + \|x_{t-1}^2\|^2$$

then

$$(w_{t-1} + y_{t-1}x_{t-1})^2 \triangleq w_{t-1}^2 + 2w_{t-1}^{\top}y_{t-1}x_{t-1} + y_{t-1}^2x_{t-1}^2$$

but  $2w_{t-1}^{\top}y_{t-1}x_{t-1} < 0$

thus

$$\|w(t)\|^2 \leq \|w(t-1)\| + \|w(t-1)\|^2$$

(d) Show by induction that  $\|w(t)\|^2 \leq tR^2$ , where  $R = \max_{1 \leq n \leq N} \|x_n\|$ .

Base case  $t = 1$ :

$$w_1 = w_0 + x'_m y'_m = 0 + x'_m y'_m = y'_m x'_m, \text{ and } \|x_m\| \leq \|x_n\|$$

(e) Using (b) and (d), show that

$$\frac{w^\top}{\|w(t)\|^2} w^* \geq \sqrt{t} \cdot \frac{\rho}{R},$$

and hence prove that

$$t \leq \frac{R^2 \|w^*\|^2}{\rho^2}.$$

[Hint:  $\frac{w^\top(t)w^*}{\|w(t)\|\|w^*\|} \leq 1$ . Why?]

Note that  $\frac{w_t^\top w^*}{\|w_t\|\|w^*\|} \leq 1$

$$\text{so } \frac{w_t^\top w^*}{\|w_t\|} \leq \|w^*\|$$

$$\text{Now } \|w_t\|^2 \leq tR^2 \Rightarrow \|w_t\| \leq \sqrt{t}R$$

$$\text{Then } \frac{1}{\|w_t\|} \geq \frac{1}{\sqrt{t}R}$$

$$\text{Thus } \frac{w_t^\top w^*}{\|w_t\|} \geq \frac{\sqrt{t}\rho}{R}$$

In practice, PLA converges more quickly than the bound  $\frac{R^2 \|w^*\|^2}{\rho^2}$  suggests. Nevertheless, because we do not know  $\rho$  in advance, we can't determine the number of iterations to convergence, which does pose a problem if the data is non-separable.

## Problem 1.4

In Exercise 1.4, we use an artificial data set to study the perceptron learning algorithm. This problem leads you to explore the algorithm further with data sets of different sizes and dimensions.

```

1  ## pla
2  ## Mac Radigan
3
4  """pla.py
5
6  Perceptron Learning Algorithm (PLA)
7
8
9  initialization:
10
11     N data dimensionality
12     K number of samples in dataset
13
14     W [Nx1] initial weights
15
16
17 random dataset generation:
18
19     X      [NxK] ~ U[a,b] random training samples

```

```

20     W_star [Nx1] ~ U[a,b]  random target function
21
22
23     update step:
24
25     y_hat  = signum( X * W' )
26
27     y_err  = |y - y_hat|
28
29     rhos   = x * y
30     k      = argmax y_err
31
32     err    = exist y" in y_err s.t. |y"| > 0
33
34     W      = W + Y_k*X_k  if err
35
36
37     training:
38
39     update while err
40
41
42     """
43
44     import numpy as np
45     import matplotlib.pyplot as plt
46
47     class PLA:
48
49     def __init__(self, n):
50         """initialize a new PLA:
51             n - dimension
52         """
53         self.new(n)
54
55     def new(self, n):
56         """initialize a new PLA:
57             n - dimension
58         """
59         self.n = n                # data dimensionality
60         self.t = 0                # total iterations
61         self.w = np.random.rand(1,n+1) # initial weight
62
63     def __str__(self):
64         """print insternal state"""
65         return str(self.w)
66
67     def rand_data(self, k):
68         """create random data:
69             returns X, a Kx(N+1) real matrix of samples augmented as
        ⇨ homogeneous coordinates
70         """
71         a = -100                  # randomization lower bound
72         b = +100                  # randomization upper bound
73         return np.concatenate((
74             np.ones((k, 1)),

```



```

75         a + (b-a) * np.random.rand(k, self.n)
76     ), axis=1)
77
78     def rand_target(self):
79         """create a random target function:
80             returns a random target function
81         """
82         a = -100                                # randomization lower bound
83         b = +100                                # randomization upper bound
84         return a + (b-a) * np.random.rand(1, self.n+1)
85
86     def target(self, x, w_star):
87         """create a training set:
88             x - data set, an 1xK matrix with {-1,+1} entries
89             returns truth training set
90         """
91         return np.sign(x.dot(w_star.T))
92
93     def train(self, x, y):
94         """run the learning algorithm
95             x - training data set, an NxK real matrix, augmented as homogenous
96             ↪ coordinates
97             returns the perceptron weights
98         """
99         while not self.update(x, y):
100             pass
101         return self.w
102
103     def select(self, x, y):
104         """choose a misclassified point
105             x - data set
106             returns a misclassified point
107         """
108         y_hat = np.sign(x.dot(self.w.T))          # classify training set
109         ↪ using current weights
110         y_err = np.abs(y - y_hat)                 # compute error from
111         ↪ truth
112         ks = y_err.nonzero()[0]
113         err = np.any(ks)                          # did the algorithm
114         ↪ converge?
115         if err:
116             k = ks[0]
117         else:
118             k = None
119         return k
120
121     def update(self, x, y):
122         """PLA iterative update step
123             x - data set
124             returns a boolean indicating convergence
125         """
126         self.t = self.t + 1                        # increment iteration
127         y_hat = np.sign(x.dot(self.w.T))          # classify training set
128         ↪ using current weights
129         y_err = np.abs(y - y_hat)                 # compute error from
130         ↪ truth

```

```

125     #ks = y_err.nonzero()[0]
126     #err = np.any(ks)                                # did the algorithm
    ↪ converge?
127     #if err:
128     #     k = ks[0]
129     #     self.w = self.w + y[k] * x[k]                # update weights
130     #     print('t: %d' % self.t)
131     #return not err
132     k = self.select(x, y)
133     if k is not None:
134         self.w = self.w + y[k] * x[k]                # update weights
135         print('t: %d' % self.t)
136         return False
137     else:
138         return True
139
140 def classify(self, x):
141     """classify a dataset using the internal PLA weights
142         x - data set
143         returns the classification
144     """
145     y = np.sign(x.dot(self.w.T))
146     return y
147
148 def plot(self, x, w, w_star, y, title=None, outfile=None):
149     """plot classification results
150         x - data set
151         w_star - target function
152     """
153     #plt.rc('text', usetex=True)
154     #plt.rc('font', family='serif')
155     fig = plt.figure()
156     ax = fig.add_subplot(1, 1, 1)
157     k_a = (y > 0).nonzero()[0]
158     x1_a = x[k_a, 1]
159     x2_a = x[k_a, 2]
160     k_b = (y < 0).nonzero()[0]
161     x1_b = x[k_b, 1]
162     x2_b = x[k_b, 2]
163     ax.scatter(x1_a, x2_a, color='b', marker='o')
164     ax.scatter(x1_b, x2_b, color='r', marker='x')
165     dx1 = np.linspace(-150, 150, 2)
166     ## target function
167     w0 = w_star.T[0]
168     w1 = w_star.T[1]
169     w2 = w_star.T[2]
170     m = -w1/w2
171     b = -w0/w2
172     dx2 = m * dx1 + b
173     ax.plot(dx1, dx2, 'g')
174     ## PLA results
175     w0 = w.T[0]
176     w1 = w.T[1]
177     w2 = w.T[2]
178     m = -w1/w2
179     b = -w0/w2

```

```

180     dx2 = m * dx1 + b
181     ax.plot(dx1, dx2, 'm--')
182     ## decorations
183     ax.set_xlim(-200.0, 200.0)
184     ax.set_ylim(-200.0, 200.0)
185     plt.xlabel(r'x1')
186     plt.ylabel(r'x2')
187     if title is not None:
188         ax.set_title(title)
189     if outfile is not None:
190         print('save: %s' % (outfile))
191         plt.savefig(outfile)
192     else:
193         plt.show()
194     return ax
195
196 def plot1(self, x, w_star, y, title=None, outfile=None):
197     """plot classification results
198         x - data set
199         w_star - target function
200     """
201     #plt.rc('text', usetex=True)
202     #plt.rc('font', family='serif')
203     fig = plt.figure()
204     ax = fig.add_subplot(1, 1, 1)
205     k_a = (y > 0).nonzero()[0]
206     x1_a = x[k_a, 1]
207     x2_a = x[k_a, 2]
208     k_b = (y < 0).nonzero()[0]
209     x1_b = x[k_b, 1]
210     x2_b = x[k_b, 2]
211     ax.scatter(x1_a, x2_a, color='b', marker='o')
212     ax.scatter(x1_b, x2_b, color='r', marker='x')
213     w0 = w_star.T[0]
214     w1 = w_star.T[1]
215     w2 = w_star.T[2]
216     m = -w1/w2
217     b = -w0/w2
218     dx1 = np.linspace(-150, 150, 2)
219     dx2 = m * dx1 + b
220     ax.plot(dx1, dx2, color='m')
221     ax.set_xlim(-200.0, 200.0)
222     ax.set_ylim(-200.0, 200.0)
223     plt.xlabel(r'x1')
224     plt.ylabel(r'x2')
225     if title is not None:
226         ax.set_title(title)
227     if outfile is not None:
228         print('save: %s' % (outfile))
229         plt.savefig(outfile)
230     else:
231         plt.show()
232     return ax
233
234 def plot_misclassified(self, x, w_star, y, k, title=None, outfile=None):
235     """plot a misclassified point

```

```

236         x - data set
237         w_star - target function
238         """
239         #plt.rc('text', usetex=True)
240         #plt.rc('font', family='serif')
241         fig = plt.figure()
242         ax = fig.add_subplot(1, 1, 1)
243         k_a = (y > 0).nonzero()[0]
244         x1_a = x[k_a, 1]
245         x2_a = x[k_a, 2]
246         k_b = (y < 0).nonzero()[0]
247         x1_b = x[k_b, 1]
248         x2_b = x[k_b, 2]
249         ax.scatter(x1_a, x2_a, color='b', marker='o')
250         ax.scatter(x1_b, x2_b, color='r', marker='x')
251         x1_m = k[1]
252         x2_m = k[2]
253         ax.scatter(x1_m, x2_m, color='b', marker='*')
254         w0 = w_star.T[0]
255         w1 = w_star.T[1]
256         w2 = w_star.T[2]
257         m = -w1/w2
258         b = -w0/w2
259         dx1 = np.linspace(-150, 150, 2)
260         dx2 = m * dx1 + b
261         ax.plot(dx1, dx2, color='m')
262         ax.set_xlim(-200.0, 200.0)
263         ax.set_ylim(-200.0, 200.0)
264         plt.xlabel(r'x1')
265         plt.ylabel(r'x2')
266         if title is not None:
267             ax.set_title(title)
268         if outfile is not None:
269             print('save: %s' % (outfile))
270             plt.savefig(outfile)
271         else:
272             plt.show()
273         return ax
274
275     def training_movie(self, x, w_star, y, title=None, outfile=None):
276         """save a movie of the PLA training process
277         returns the perceptron weights
278         """
279         t = 0
280         err = True
281         base = outfile
282         #while err:
283         # k = self.select(x, y)
284         # outfile = "%s_%03.3d.png" % (base, self.t)
285         # title = "%s, t=%d" % (title, self.t)
286         # if k is not None:
287         #     self.plot_misclassified(x, w_star, y, k, title, outfile)
288         #     self.w = self.w + y[k] * x[k] # update weights
289         #     print('t: %d' % self.t)
290         #     err = True
291         # else:

```

```

292 # self.plot1(x, w_star, y, title, outfile)
293 # print('t: %d' % self.t)
294 # err = False
295 # t = t + 1
296 #return self.w
297 #####
298 while err:
299     k = self.select(x, y)
300     self.t = self.t + 1 # increment iteration
301     outfile = "%s_%03.3d.png" % (base, self.t)
302     title = "%s, t=%d" % (title, self.t)
303     y_hat = np.sign(x.dot(self.w.T)) # classify training
↪ set using current weights
304     y_err = np.abs(y - y_hat) # compute error from
↪ truth
305     ks = y_err.nonzero()[0]
306     err = np.any(ks) # did the algorithm
↪ converge?
307     if err:
308         k = ks[0]
309         self.w = self.w + y[k] * x[k] # update weights
310         self.plot_misclassified(x, w_star, y, k, title, outfile)
311         print('t: %d' % self.t)
312     else:
313         self.plot1(x, w_star, y, title, outfile)
314         print('t: %d' % self.t)
315     return not err
316
317 ## *EOF*

```

- (a) Generate a linearly separable data set of size 20 as indicated in Exercise 1.4. Plot the examples  $(x_n, y_n)$  as well as the target function  $f$  on a plane. Be sure to mark the examples from different classes differently, and add labels to the axes of the plot.

```

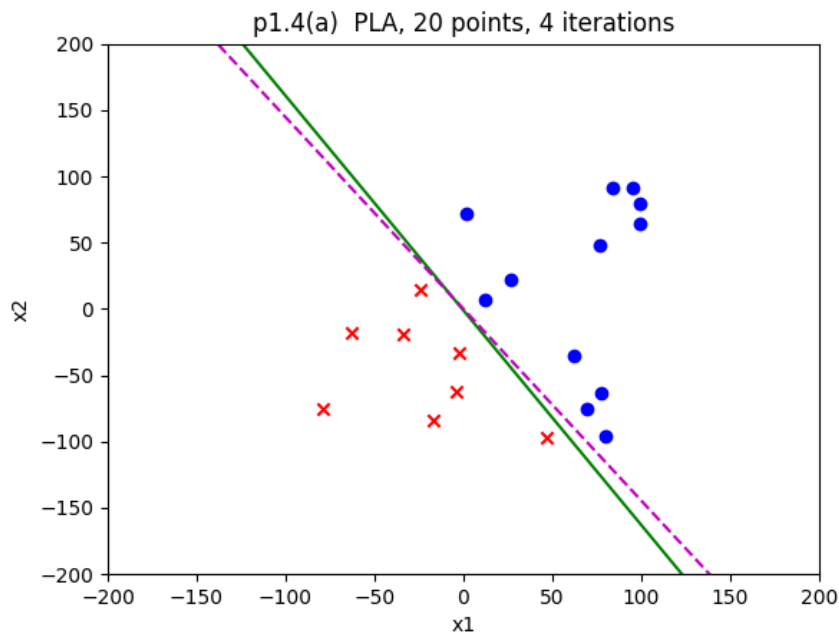
1 #!/usr/bin/env python
2 ## problem_1_4_a.py
3 ## Mac Radigan
4
5 """problem_1_4_a.py
6 """
7
8 import argparse
9 from sys import stdout
10 from perceptron import pla
11 import numpy as np
12
13 def main(verbose):
14     'problem 1.4'
15     dim = 20
16     test(dim)
17
18 def case(dim):
19     'problem 1.4'

```

```

20     outdir = 'figures/'
21     p = pla.PLA(2)
22     x = p.rand_data(dim)
23     w_star = p.rand_target()
24     y = p.target(x, w_star)
25     #outfile = '%s/%s' % (outdir, 'p1.4a_target.png')
26     #title = 'p1.4(a) Target Function, %d points' % (dim)
27     #ax = p.plot1(x, w_star, y, title, outfile)
28     w = p.train(x, y)
29     y_hat = p.classify(x)
30     y_err = y_hat - y;
31     err = 0 != np.sum(y_err)
32     #outfile = '%s/%s' % (outdir, 'p1.4a_classify.png')
33     #title = 'p1.4(a) PLA Results, %d iterations' % (p.t)
34     #ax = p.plot1(x, w, y_hat, title, outfile)
35     outfile = '%s/%s' % (outdir, 'p1.4a.png')
36     title = 'p1.4(a) PLA, %d points, %d iterations' % (dim, p.t)
37     ax = p.plot(x, w, w_star, y_hat, title, outfile)
38     return (p, x, w_star, y, w, y_hat, y_err, err)
39
40 def prn(name, value):
41     print('%s:\n%s\n' % (name, value))
42
43 def test(dim):
44     'test'
45     (p, x, w_star, y, w, y_hat, y_err, err) = case(dim)
46     prn("x", x)
47     prn("w_star", w_star)
48     prn("y", y.T)
49     prn("w", w)
50     prn("y_hat", y_hat.T)
51     #prn("y_err", y_err.T)
52     prn("err", err)
53
54 if __name__ == '__main__':
55     parser = argparse.ArgumentParser(description='problem 1.4 (a)')
56     parser.add_argument('-v', '--verbose', action='store_true',
57         ↪ dest='verbose', default=False, help='verbose output to stdout')
57     args = parser.parse_args()
58     main(args.verbose)
59
60 ## *EOF*

```



- (b) Run the perceptron learning algorithm on the data set above. Report the number of updates that the algorithm takes before converging. Plot the examples  $(x_n, y_n)$ , the target function  $f$ , and the final hypothesis  $g$  in the same figure. Comment on whether  $f$  is close to  $g$ .

```

1  #!/usr/bin/env python
2  ## problem_1_4_b.py
3  ## Mac Radigan
4
5  """problem_1_4_b.py
6  """
7
8  import argparse
9  from sys import stdout
10 from perceptron import pla
11 import numpy as np
12
13 def main(verbose):
14     'problem 1.4'
15     dim = 20
16     test(dim)
17
18 def case(dim):
19     'problem 1.4'
20     outdir = 'figures/'
21     p = pla.PLA(2)

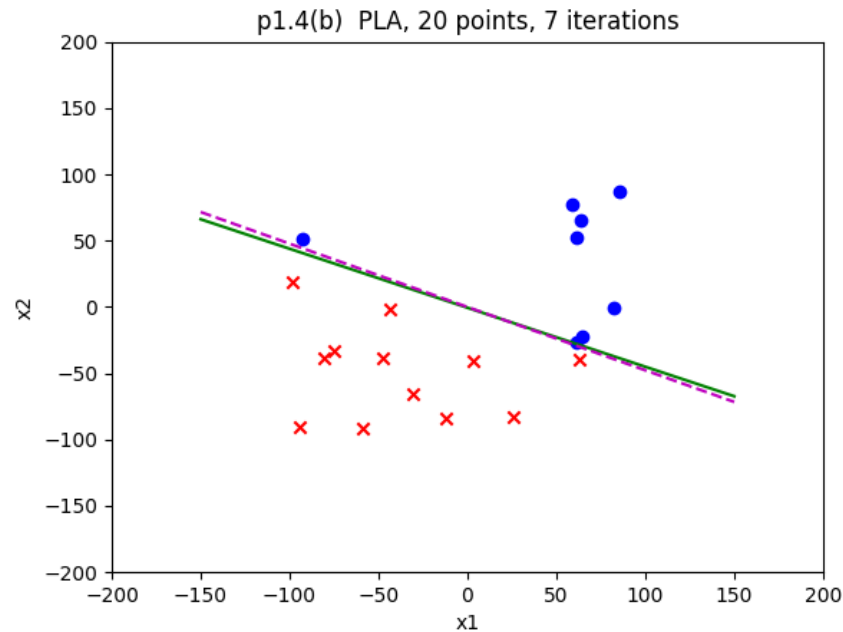
```

```

22 x = p.rand_data(dim)
23 w_star = p.rand_target()
24 y = p.target(x, w_star)
25 w = p.train(x, y)
26 y_hat = p.classify(x)
27 y_err = y_hat - y;
28 err = 0 != np.sum(y_err)
29 outfile = '%s/%s' % (outdir, 'p1.4b.png')
30 title = 'p1.4(b) PLA, %d points, %d iterations' % (dim, p.t)
31 ax = p.plot(x, w, w_star, y_hat, title, outfile)
32 return (p, x, w_star, y, w, y_hat, y_err, err)
33
34 def prn(name, value):
35     print('%s:\n%s\n' % (name, value))
36
37 def test(dim):
38     'test'
39     (p, x, w_star, y, w, y_hat, y_err, err) = case(dim)
40     prn("x", x)
41     prn("w_star", w_star)
42     prn("y", y.T)
43     prn("w", w)
44     prn("y_hat", y_hat.T)
45     #prn("y_err", y_err.T)
46     prn("err", err)
47
48 if __name__ == '__main__':
49     parser = argparse.ArgumentParser(description='problem 1.4 (a)')
50     parser.add_argument('-v', '--verbose', action='store_true',
51         ↪ dest='verbose', default=False, help='verbose output to stdout')
52     args = parser.parse_args()
53     main(args.verbose)
54     ## *EOF*

```





- (c) Repeat everything in (b) with another randomly generated data set of size 20. Compare your results with (b).

```

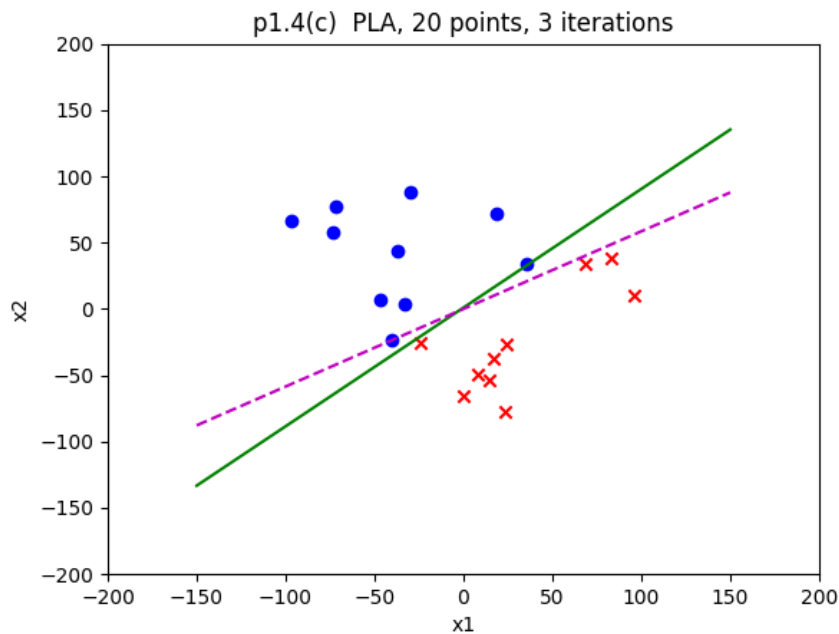
1  #!/usr/bin/env python
2  ## problem_1_4_c.py
3  ## Mac Radigan
4
5  """problem_1_4_c.py
6  """
7
8  import argparse
9  from sys import stdout
10 from perceptron import pla
11 import numpy as np
12
13 def main(verbose):
14     'problem 1.4'
15     dim = 20
16     test(dim)
17
18 def case(dim):
19     'problem 1.4'
20     outdir = 'figures/'
21     p = pla.PLA(2)
22     x = p.rand_data(dim)
23     w_star = p.rand_target()

```

```

24     y = p.target(x, w_star)
25     w = p.train(x, y)
26     y_hat = p.classify(x)
27     y_err = y_hat - y;
28     err = 0 != np.sum(y_err)
29     outfile = '%s/%s' % (outdir, 'p1.4c.png')
30     title = 'p1.4(c) PLA, %d points, %d iterations' % (dim, p.t)
31     ax = p.plot(x, w, w_star, y_hat, title, outfile)
32     return (p, x, w_star, y, w, y_hat, y_err, err)
33
34 def prn(name, value):
35     print('%s:\n%s\n' % (name, value))
36
37 def test(dim):
38     'test'
39     (p, x, w_star, y, w, y_hat, y_err, err) = case(dim)
40     prn("x", x)
41     prn("w_star", w_star)
42     prn("y'", y.T)
43     prn("w", w)
44     prn("y_hat'", y_hat.T)
45     #prn("y_err'", y_err.T)
46     prn("err'", err)
47
48 if __name__ == '__main__':
49     parser = argparse.ArgumentParser(description='problem 1.4 (a)')
50     parser.add_argument('-v', '--verbose', action='store_true',
51                         dest='verbose', default=False, help='verbose output to stdout')
52     args = parser.parse_args()
53     main(args.verbose)
54     ## *EOF*

```



- (d) Repeat everything in (b) with another randomly generated data set of size 100. Compare your results with (b).

```

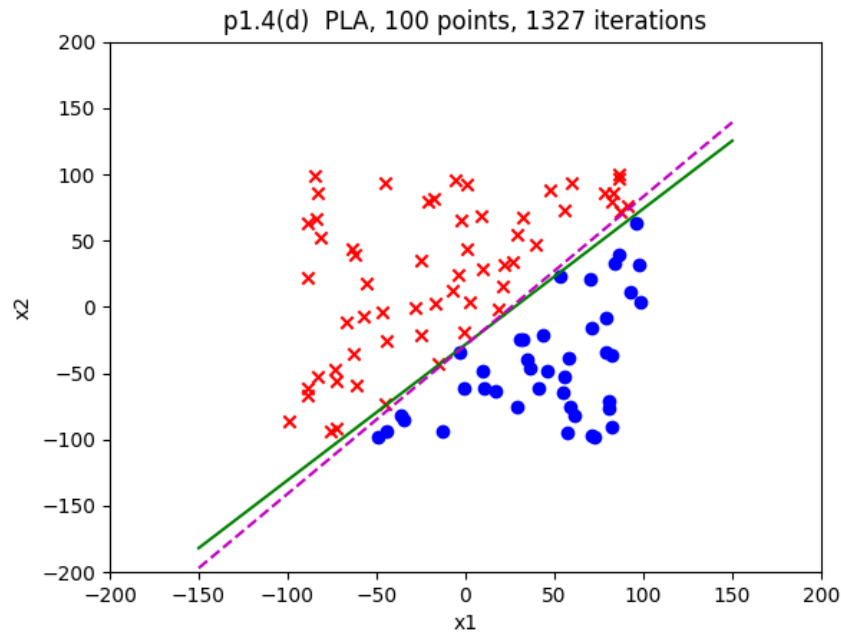
1  #!/usr/bin/env python
2  ## problem_1_4_d.py
3  ## Mac Radigan
4
5  """problem_1_4_d.py
6  """
7
8  import argparse
9  from sys import stdout
10 from perceptron import pla
11 import numpy as np
12
13 def main(verbose):
14     'problem 1.4'
15     dim = 100
16     test(dim)
17
18 def case(dim):
19     'problem 1.4'
20     outdir = 'figures/'
21     p = pla.PLA(2)
22     x = p.rand_data(dim)
23     w_star = p.rand_target()

```

```

24     y = p.target(x, w_star)
25     w = p.train(x, y)
26     y_hat = p.classify(x)
27     y_err = y_hat - y;
28     err = 0 != np.sum(y_err)
29     outfile = '%s/%s' % (outdir, 'p1.4d.png')
30     title = 'p1.4(d) PLA, %d points, %d iterations' % (dim, p.t)
31     ax = p.plot(x, w, w_star, y_hat, title, outfile)
32     return (p, x, w_star, y, w, y_hat, y_err, err)
33
34 def prn(name, value):
35     print('%s:\n%s\n' % (name, value))
36
37 def test(dim):
38     'test'
39     (p, x, w_star, y, w, y_hat, y_err, err) = case(dim)
40     prn("x", x)
41     prn("w_star", w_star)
42     prn("y'", y.T)
43     prn("w", w)
44     prn("y_hat'", y_hat.T)
45     #prn("y_err'", y_err.T)
46     prn("err'", err)
47
48 if __name__ == '__main__':
49     parser = argparse.ArgumentParser(description='problem 1.4 (a)')
50     parser.add_argument('-v', '--verbose', action='store_true',
51                         dest='verbose', default=False, help='verbose output to stdout')
52     args = parser.parse_args()
53     main(args.verbose)
54     ## *EOF*

```



- (e) Repeat everything in (b) with another randomly generated data set of size 1,000. Compare your results with (b).

```

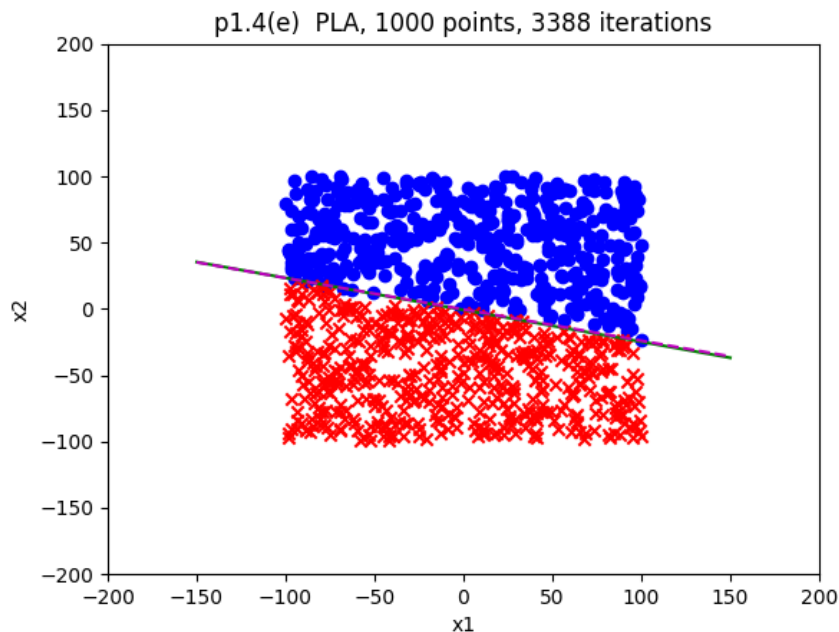
1  #!/usr/bin/env python
2  ## problem_1_4_d.py
3  ## Mac Radigan
4
5  """problem_1_4_d.py
6  """
7
8  import argparse
9  from sys import stdout
10 from perceptron import pla
11 import numpy as np
12
13 def main(verbose):
14     'problem 1.4'
15     dim = 100
16     test(dim)
17
18 def case(dim):
19     'problem 1.4'
20     outdir = 'figures/'
21     p = pla.PLA(2)
22     x = p.rand_data(dim)
23     w_star = p.rand_target()

```

```

24     y = p.target(x, w_star)
25     w = p.train(x, y)
26     y_hat = p.classify(x)
27     y_err = y_hat - y;
28     err = 0 != np.sum(y_err)
29     outfile = '%s/%s' % (outdir, 'p1.4d.png')
30     title = 'p1.4(d) PLA, %d points, %d iterations' % (dim, p.t)
31     ax = p.plot(x, w, w_star, y_hat, title, outfile)
32     return (p, x, w_star, y, w, y_hat, y_err, err)
33
34 def prn(name, value):
35     print('%s:\n%s\n' % (name, value))
36
37 def test(dim):
38     'test'
39     (p, x, w_star, y, w, y_hat, y_err, err) = case(dim)
40     prn("x", x)
41     prn("w_star", w_star)
42     prn("y'", y.T)
43     prn("w", w)
44     prn("y_hat'", y_hat.T)
45     #prn("y_err'", y_err.T)
46     prn("err'", err)
47
48 if __name__ == '__main__':
49     parser = argparse.ArgumentParser(description='problem 1.4 (a)')
50     parser.add_argument('-v', '--verbose', action='store_true',
51                         dest='verbose', default=False, help='verbose output to stdout')
52     args = parser.parse_args()
53     main(args.verbose)
54     ## *EOF*

```



- (f) Modify the algorithm such that it takes  $x_n \in \mathbb{R}^n$  instead of  $\mathbb{R}^2$ . Randomly generate a linearly separable data set of size 1,000 with  $x_n \in \mathbb{R}^{10}$  and feed the data set to the algorithm. How many updates does the algorithm take to converge?
- (g) Repeat the algorithm on the same data set as (f) for 100 experiments. In the iterations of each experiment, pick  $x(t)$  randomly instead of deterministically. Plot a histogram of the number of updates that the algorithm takes to converge.
- (h) Summarize your conclusions with respect to accuracy and running time as a function of  $N$  and  $d$ .

## Problem 1.8

The Hoeffding Inequality is one form of the law of large numbers. One of the simplest forms of that law is the Chebyshev Inequality, which you will prove here.

- (a) If  $t$  is a non negative random variable, prove that for any  $\alpha > 0$ ,  $P[t \geq \alpha] \leq \frac{E(t)}{\alpha}$

By definition,

$$P[t \geq \alpha] \triangleq \int_{\alpha}^{\infty} p(x)dx$$

and

$$E(t) \triangleq \int_{-\infty}^{\infty} xp(x)dx$$

so evaluate

$$\alpha \int_{\alpha}^{\infty} p(x)dx = \alpha P[t \geq \alpha] \leq E(t) = \int_{-\infty}^{\infty} xp(x)dx$$

since  $t$  is strictly positive, this can be written as

$$\int_{\alpha}^{\infty} \alpha p(x)dx \geq \int_0^{\alpha} xp(x)dx + \int_{\alpha}^{\infty} xp(x)dx$$

$$\text{note that } \int_{\alpha}^{\infty} \alpha p(x)dx \geq \int_{\alpha}^{\infty} xp(x)dx$$

$$\text{and because } t > 0, \int_0^{\alpha} xp(x)dx \geq 0$$

$$\text{so it holds that } \alpha \int_{\alpha}^{\infty} p(x)dx = \int_{-\infty}^{\infty} xp(x)dx$$

$$\text{and thus } P[t \geq \alpha] \leq \frac{E(t)}{\alpha}$$

- (b) If  $u$  is any random variable with mean  $\mu$  and variance  $\sigma^2$ , prove that for any  $\alpha > 0$ ,  $P[(u - \mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$ . [Hint: Use (a)]

By definition,

$$\sigma^2 \triangleq E[(u - \mu)^2]$$

and thus with substitution of  $t := (u - \mu)^2$  and using (a), we have

$$P[(u - \mu)^2 \geq \alpha] \leq \frac{\sigma^2}{\alpha}$$

- (c) If  $u_1, \dots, u_N$  are iid random variables, each with mean  $\mu$  and variance  $\sigma^2$ , and  $u = \frac{1}{N} \sum_{n=1}^N u_n$ , prove that for any  $\alpha > 0$ ,  $P[(u - \mu)^2 \geq \alpha] \leq \frac{\sigma^2}{N\alpha}$ .

Notice that the RHS of this Chebyshev Inequality goes down linearly in  $N$ , while the counterpart in Hoeffding's Inequality goes down exponentially. In Problem 1.9, we develop an exponential bound using a similar approach.

## Chapter 2 Notes

	Symbol	Name	
	$h(x_k) \exists x_k \in (X)$	Dichotomy	
	$N$	Number of Data Points	
	$m_{\mathcal{H}}(N)$	Growth Function	
	$m_{\mathcal{H}}(N)$	Growth Function	
	$B(N, k)$	Growth Bound with Breakpoint $k$	
v	$d_{vc}$	Vapnik-Chervonenkis Dimension	24



## Generalization Error

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \quad (9)$$

## Growth Function

$$m_{\mathcal{H}}(N) = \max_{\vec{x} \in \mathcal{X}} |\mathcal{H}(\vec{x})| \quad (10)$$

## Bounding the Growth Function

$B(N, k)$  is the maximum number of dichotomies on  $N$  points such that no subset of size  $k$  of the  $N$  points can be shattered by these dichotomies.

## Vapnik-Chervonenkis Dimension

$$d_{VC}(\mathcal{H}) = \begin{cases} \max N \text{ s.t. } m_{\mathcal{H}}(N) = 2^N & \text{if } m_{\mathcal{H}}(N) < 2^N \exists N \\ \infty & \text{if } m_{\mathcal{H}}(N) = 2^N \forall N \end{cases} \quad (11)$$

$k = d_{VC} + 1$  is a breakpoint for  $m_{\mathcal{H}}(N)$

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$$

$$m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1$$

## Sauer's Lemma

$$B(N, k) \leq \sum_{i=0}^{k-1} \binom{N}{i} \quad (12)$$

## Problem 2.1

In Equation (2.1), set  $\delta = 0.03$  and let

$$\sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \leq \epsilon \implies 2N \geq \frac{\ln \frac{2M}{\delta}}{\epsilon^2} \implies N \geq \frac{1}{2\epsilon^2} \ln \frac{2M}{\delta}$$

(a) For  $M = 1$ , how many examples do we need to make  $\epsilon \leq 0.05$ ?

$$N \geq \frac{1}{2(0.05)^2} \ln \frac{2 \cdot 1}{0.03}$$

(b) For  $M = 100$ , how many examples do we need to make  $\epsilon \leq 0.05$ ?

$$N \geq \frac{1}{2(0.05)^2} \ln \frac{2 \cdot 100}{0.03}$$

(c) For  $M = 10,000$ , how many examples do we need to make  $\epsilon \leq 0.05$ ?

$$N \geq \frac{1}{2(0.05)^2} \ln \frac{2 \cdot 10,000}{0.03}$$

### Problem 2.3

Compute the maximum number of dichotomies,  $m_{\mathcal{H}}(N)$ , for these learning models, and consequently compute  $d_{VC}$ , the VC dimension.

(a) Positive or negative ray:  $\mathcal{H}$  contains the functions which are +1 on  $[a, \infty)$  (for some  $a$ ), together with those that are +1 on  $(-\infty, a]$  (for some  $a$ ).

$$m_{\mathcal{H}}(N) = 2N + 1 \text{ dichotomies}$$

(b) Positive or negative interval.  $\mathcal{H}$  contains the functions which are +1 on an interval  $[a, b]$  and -1 elsewhere, or -1 on an interval  $[a, b]$  and +1 elsewhere.

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = N^2 + N + 1 \text{ dichotomies}$$

### Problem 2.8

Which of the following are possible growth functions  $m_{\mathcal{H}}(N)$  for some hypothesis set:

(a)  $1 + N$

yes, example positive ray

(b)  $1 + N + \frac{N(N-1)}{2}$

no, must be integer-valued (cannot have a fractional number of dichotomies)

(c)  $2^N$

yes

(d)  $2^{\lfloor N \rfloor}$

yes

(e)  $2^{\lfloor \sqrt{N} \rfloor}$

yes

(f)  $2^{\lfloor \frac{N}{2} \rfloor}$

yes

(g)  $1 + N + \frac{N(N-1)(N-2)}{6}$

$$1 + N + \frac{N(N-1)(N-2)}{6} = \frac{N^3}{6} + \frac{N}{2} + \frac{4}{3}$$

no, must be integer-valued (cannot have a fractional number of dichotomies)

## Problem 2.13

- (a) Let  $\mathcal{H} = \{h_1, h_2, \dots, h_M\}$ . Prove that  $d_{VC}(\mathcal{H}) \leq \log_2 M$ .

Finite  $\mathcal{H}$  can represent at most  $|\mathcal{H}| = M$  dichotomies. Thus,  $2^{v_{DC}} \leq M$ , or  $v_{DC} \leq \log_2 M$ .

- (b) For hypothesis sets  $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_K$  with finite dimensions  $d_{VC}(\mathcal{H}_K)$ , derive and prove the tightest upper and lower bound that you can get on  $d_{VC}\left(\bigcap_{k=1}^K \mathcal{H}_k\right)$ .
- (c) For hypothesis sets  $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_K$  with finite dimensions  $d_{VC}(\mathcal{H}_K)$ , derive and prove the tightest upper and lower bound that you can get on  $d_{VC}\left(\bigcup_{k=1}^K \mathcal{H}_k\right)$ .

## Chapter 3 Notes

### Problem 3.1

Consider the double semi-circle “toy” learning task below (see figure). There are two semi-circles of width  $thk$  with inner radius  $rad$ , separated by  $sep$  as shown (red is -1, blue is +1). The center of the top semi-circle is linearly separable when  $sep \geq 0$ , and not so for  $sep < 0$ . Set  $rad = 10$ ,  $thk = 5$ , and  $sep = 5$ . Then, generate 2e3 examples uniformly, which means you will have approximately 1e3 examples for each class.

- (a) Run the PLA starting from  $w = 0$  until it converges. Plot the data and final hypothesis.

```

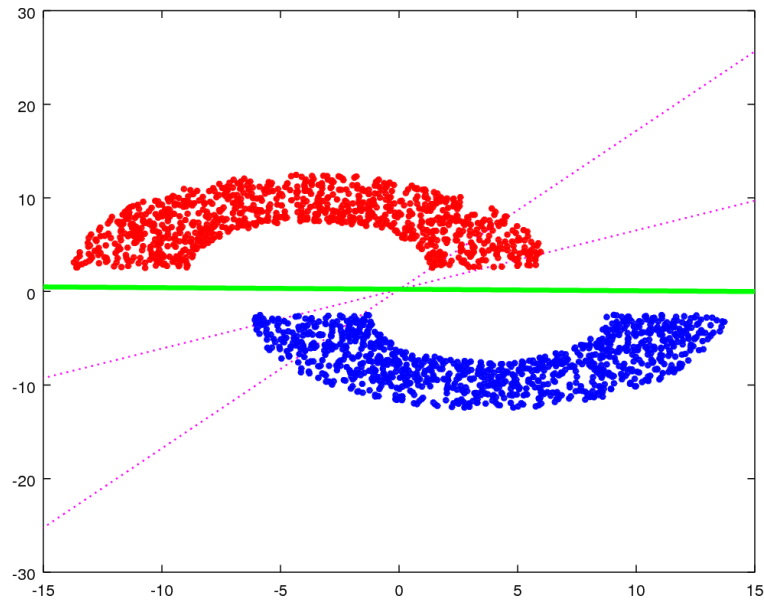
1  #!/usr/bin/env octave
2  ## problem_3_1_a.m
3  ## Mac Radigan
4
5  FORCES__SCRIPT_FILE=1;
6  ux=false;
7
8  function ax = do_plot(ux, rad, thk, c1, c2, x1, x2, w, k_err)
9
10     if ux
11         show = 'on';
12     else
13         show = 'off';
14     end
15
16     %% plot input dataset
17     ax = figure(30111);
18     set(ax, 'visible', show);
19     plot(x1(c1), x2(c1), 'LineWidth', 1, 'r.');
20     hold on;
21     plot(x1(c2), x2(c2), 'LineWidth', 1, 'b.');
22     hold on;
23
24     %% plot final hypothesis, g(x)
25     ext = rad + thk;
26     m = -w(2)/w(3);
27     b = -w(1)/w(3);
28     dx1 = linspace(-ext, ext, 2);
29     dx2 = m * dx1 + b;
30     ax = figure(30111);
31     set(ax, 'visible', show);
32     if k_err > 0
33         plot(dx1, dx2, 'Color', 'magenta', 'LineWidth', 1, 'LineStyle',
34             ↪ ':');
35     else
36         plot(dx1, dx2, 'Color', 'green', 'LineWidth', 3);
37     end
38     hold on;
39     drawnow();
40
41 end % function do_plot
42
43 N = 2e3; % number of training samples
44 rad = 10; % radius of semi-circle
45 thk = 5; % thickness of semi-circle
46 sep = 5; % separation between semi-circles
47
48 %% uniformly distributed data
49 z = (rad - thk * (1 - rand(1,N))) .* exp(-j*2*pi*rand(1,N));
50
51 %% target function

```

```

52 f = @(z) sign(angle(z));
53
54 %% target set
55 y = f(z);
56
57 %% cartesian basis representation
58 x1 = -f(z)*(rad/2-thk/4) + abs(z) .* cos(angle(z));
59 x2 = f(z)*sep/2 + abs(z) .* sin(angle(z));
60
61 %% categorical index
62 c1 = y > 0;
63 c2 = y < 0;
64
65 %% Perceptron Learning Algorithm (PLA)
66 t = 0; % training step counter
67 w = zeros(1,3); % initial weights
68 %w = randn(1,3); % initial weights
69 X = [ones(size(y)); x1; x2]; % training data
70 k_err = inf; % convergence criteria
71 while k_err > 0
72     t = t + 1; % update training step counter
73     ax = do_plot(ux, rad, thk, c1, c2, x1, x2, w, k_err);
74     if ux
75         fprintf(1, 'error: %f\n', k_err);
76         t
77         w
78     end
79     y_hat = sign(w*X); % classify
80     y_err = 0.5*abs(y - y_hat); % residual
81     [k_err, k] = max(y_err); % select misclassified point
82     if k_err <= 0, continue, end
83     x_k = X(:,k); % in the training data
84     y_k = y(k); % in the target set
85     w = w + y_k*x_k'; % update weights
86     if ux
87         disp('...')
88         input('...')
89     end
90 end % training
91
92 ax = do_plot(ux, rad, thk, c1, c2, x1, x2, w, k_err);
93 if ux
94     disp('training done.')
95 else
96     saveas(ax, 'figures/p3.1a.png');
97 end
98
99 ## *EOF*

```



- (b) Repeat part (a) using the linear regression (for classification) to obtain  $w$ . Explain your observations.

```

1  #!/usr/bin/env octave
2  ## problem_3_1_b.m
3  ## Mac Radigan
4
5  FORCES__SCRIPT_FILE=1;
6  ux=false;
7
8  function ax = do_plot(ux, rad, thk, c1, c2, x1, x2, w)
9
10     if ux
11         show = 'on';
12     else
13         show = 'off';
14     end
15
16     %% plot input dataset
17     ax = figure(30112);
18     set(ax, 'visible', show);
19     plot(x1(c1), x2(c1), 'LineWidth', 1, 'r. ');
20     hold on;
21     plot(x1(c2), x2(c2), 'LineWidth', 1, 'b. ');
22     hold on;
23
24     %% plot input dataset

```

```

25     ax = figure(30112);
26     set(ax, 'visible', show);
27     plot(x1(c1), x2(c1), 'rx');
28     hold on;
29     plot(x1(c2), x2(c2), 'bx');
30     hold on;
31
32     %% plot final hypothesis, g(x)
33     ext = rad + thk;
34     m = -w(2)/w(3);
35     b = -w(1)/w(3);
36     dx1 = linspace(-ext, ext, 2);
37     dx2 = m * dx1 + b;
38     ax = figure(30112);
39     set(ax, 'visible', show);
40     plot(dx1, dx2, 'Color', 'green', 'LineWidth', 3);
41     hold off;
42
43     drawnow();
44
45 end % function do_plot
46
47 N = 2e3; % number of training samples
48 rad = 10; % radius of semi-circle
49 thk = 5; % thickness of semi-circle
50 sep = 5; % separation between semi-circles
51
52 %% uniformly distributed data
53 z = (rad - thk * (1 - rand(1,N))) .* exp(-j*2*pi*rand(1,N));
54
55 %% target function
56 f = @(z) sign(angle(z));
57
58 %% target set
59 y = f(z);
60
61 %% cartesian basis representation
62 x1 = -f(z)*(rad/2-thk/4) + abs(z) .* cos(angle(z));
63 x2 = f(z)*sep/2 + abs(z) .* sin(angle(z));
64
65 %% categorical index
66 c1 = y > 0;
67 c2 = y < 0;
68
69 %% Linear Regression
70 X = [ones(size(y)); x1; x2]; % training data
71 w = y * pinv(X);
72
73 %% plot results
74 ax = do_plot(ux, rad, thk, c1, c2, x1, x2, w);
75
76 if ~ux
77     saveas(ax, 'figures/p3.1b.png');
78 end
79
80 ## *EOF*

```

