# CS 5350/6350: Machine Learning Spring 2020

Homework 3
Britton Gaul
u0915408

Handed out: 25 Feb, 2020
Due date: 11:59pm, 7 Mar, 2020

## 1 Paper Problems [40 points + 10 bonus]

1.

(a)

Yes there is a margin, because the no points are located on the hyperplane and are classified correctly.
formula used: $d(x,h) = \frac{|w^T x+b|}{|w|}$
$minimum\ margin = \frac{1}{\sqrt{13}}$

(b)

The hyperplane does not have a margin for this dataset, because the last training example is missclassified.

2.

(a) The margin can be calculated from this dataset.
$margin = \frac{\sqrt{2}}{2}$

(b)

The margin cannot be calculated from this dataset, becasue the data is not seperable.

3.

(a)

The upper bound for the number of mistakes made by the Perceptron algorithm is $\frac{R}{\gamma^2}$.
If $u$ is not a unit vector then the bound becomes $\frac{R^2||u||^2}{\gamma^2}$
This is becasue of the equation:
$u^T w_t = ||u||||w_t||cos\theta$ with t mistakes
The equation:
$\sqrt{t}R \geq ||w_t|| \geq \frac{u^T w_t}{||u||} \geq \frac{t\gamma}{||u||} = t \leq \frac{R^2||u||^2}{\gamma^2}$ can be formed

(b)

The margin requirement should be increased, $\forall i, y_i(u^T x_i) \geq ||u||\gamma$.

(c)

By using the second assumption it can be stated that there exists $h : w^T x = 0$, which can seperate the data with margin $\gamma$.
$\frac{|w^T x^i|}{||w||} = \frac{y_i(w^T x^i)}{||w||} \geq \gamma$
Which as a result makes the mistake bound $\frac{R^2}{\gamma^2}$

4.

$R = \sqrt{n}$
$u = (-\frac{1}{\sqrt{2k}}, ..., \frac{1}{\sqrt{2k}}, ..., 0)$
postive samples: $y_i(u^T x^i) \geq -\frac{1}{\sqrt{2k}}$
negative samples: $y_i(u^T x^i) \geq \sqrt{2k}$
If $\gamma = \frac{1}{\sqrt{2k}}$ is chosen then it is guarenteed that for any $x^i, y_i(u^T x^i) \geq \gamma$
Therefore the upper bound for the number of mistakes is, $\frac{R^2}{\gamma^2} = \frac{\sqrt{n}^2}{\frac{1}{\sqrt{2k}}^2} = 2kn$

5.

(a)

VC dimension of $\mathcal{H} = VC(\mathcal{H}) \leq 10$

(b)

There are $2^m$ label distributions for a set of $m$ data points. Each of which correspond to an output of $h \in \mathcal{H}$. If $2^m > |\mathcal{H}|$, then there are at most $|\mathcal{H}|$ outputs. Because the number of possible partitions is $2^m$ is greater than the number of functions in $\mathcal{H}$, there is no way to shatter the data points if $m > log_2(|\mathcal{H}|)$. Therefore, generally, $VC(\mathcal{H}) \leq log_2(|\mathcal{H}|)$.

6.

For this to be the case there needs to be at least one splitting that cannot be shattered by a line, for any arrangement of 4 points. The arrangment of the 4 points can either be where one point is located in the convex closure of the 3 other points, or all 4 points are corner points of the convex closure. The first splitting cannot be shattered by a line, becasue any line that classifies the points with one label, will misclassify the points with a different label. The second splitting cannot be shattered for a similar reason. Therefore linear classifiers cannot shatter 4 points.

7. [**Bonus**] [10 points] Consider our infinite hypothesis space $\mathcal{H}$ are all rectangles in a plain. Each rectangle corresponds to a classifier — all the points inside the rectangle are classified as positive, and otherwise classified as negative. What is VC($\mathcal{H}$)?

# 2 Practice [60 points ]

1.

https://github.com/BritGaul/CS5350

2. (a)

T = 1, w = (-35.523314, -18.60129 , -23.305885, -8.57917 ), b = 23.0, error = 0.024
T = 2, w = (-37.4281297, -25.4913978, -29.340175 , -15.347344 ), b = 29.0, error = 0.028
T = 3, w = (-35.263374 , -28.07227 , -33.012615 , -11.4340566), b = 33.0, error = 0.034
T = 4, w = (-45.428818 , -37.91767 , -26.868495 , -12.4211916), b = 37.0, error = 0.038
T = 5, w = (-56.479182, -29.08942 , -35.15271 , -15.393275), b = 40.0, error = 0.028
T = 6, w = (-51.0585655, -32.22879 , -42.7972676, -9.064403 ), b = 43.0, error = 0.038
T = 7, w = (-46.83874 , -33.768615, -38.759985, -3.39268 ), b = 44.0, error = 0.02
T = 8, w = (-57.367873, -34.5365 , -37.8838 , -16.709465), b = 53.0, error = 0.018
T = 9, w = (-60.72953 , -39.21511 , -40.191914, -11.150653), b = 57.0, error = 0.014
T = 10, w = (-59.49004 , -34.84013 , -40.92277 , -13.150159), b = 60.0, error = 0.01

(b)

Becasue there are so many weight vectors being used the output is very long. For this reason the output is located in the weights-counts-errors-partb.txt file.

(c)

T = 1, w = (-1147.65044 -672.19275 -432.12782 -245.189073), b = 404.0, error = 0.074
T = 2, w = (-2434.9438585 -1496.23658 -1213.365613 -640.603101), b = 1412.0, error = 0.034
T = 3, w = (-3259.955521 -2367.11727 -2044.45811 -1005.62992), b = 2152.0, error = 0.022
T = 4, w = (-4401.110835 -2735.615095 -2428.082668 -939.644328), b = 2690.0, error = 0.026
T = 5, w = (-5339.843497 -3552.23921 -3210.62105 -1264.499558), b = 3751.0, error = 0.018
T = 6, w = (-5872.8532305 -4064.77446 -3762.32626 -1406.224791), b = 4348.0, error = 0.018
T = 7, w = (-6680.49729 -4087.99826 -4029.607708 -1189.622225), b = 5260.0, error = 0.018
T = 8, w = (-8373.824228 -5300.13512 -5191.961385 -1598.8459572), b = 5931.0, error = 0.018
T = 9, w = (-8000.8376 -5153.39226 -5149.81775 -1422.730017), b = 6528.0, error = 0.014
T = 10, w = (-10228.003859 -6551.34039 -6132.2837484 -1848.4629562), b = 8523.0, error = 0.018

(d)

Table for percentage errors:

| T | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| standard | 1.2 | 2.6 | 1.2 | 1.0 | 1.4 | 1.8 | 1.4 | 1.2 | 1.8 | 1.4 |
| voted | 3.6 | 2.2 | 1.6 | 1.6 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 |
| average | 6.8 | 2.8 | 1.8 | 2.0 | 1.6 | 1.8 | 1.8 | 1.6 | 1.8 | 1.4 |

From the table it can be seen that as the number of epochs increases the error for each method generally decreases. It can also be seen that the standard perceprton has the smallest test error, followed by the voted perceptron, and then the average perceptron generally has the highest test error.