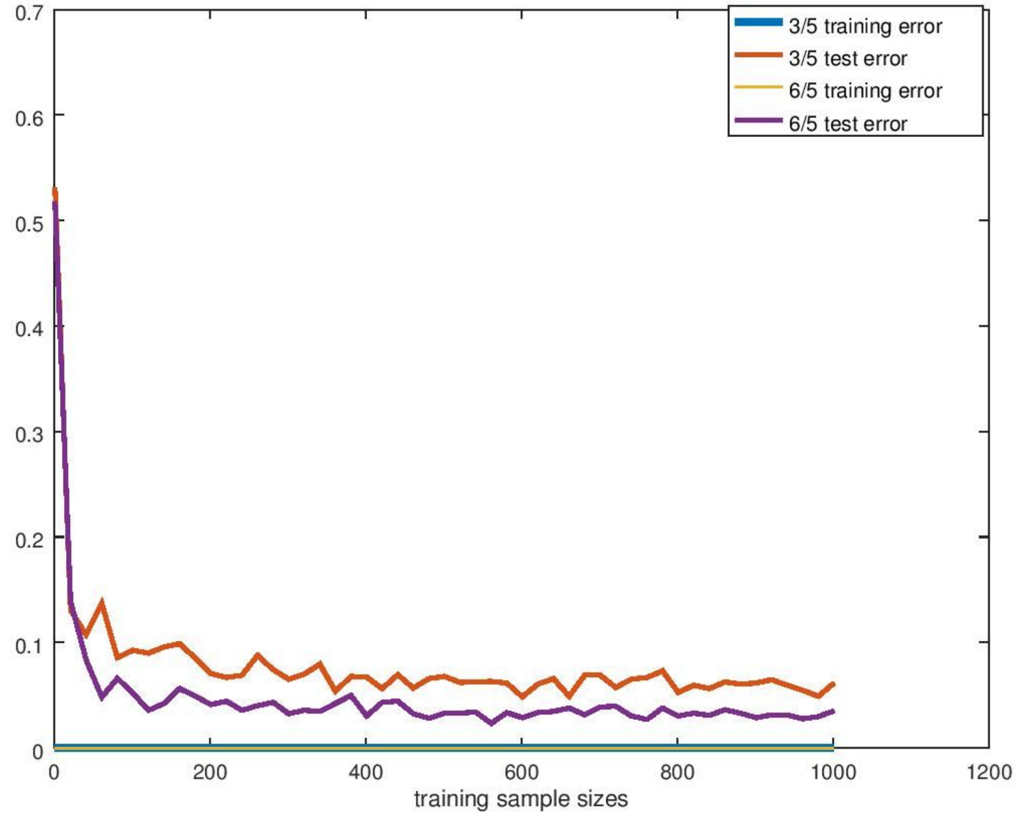
Assignment 1

Answers 2. (a) Perceptron error rate on test sample of 3/5 and 6/5 figures, by training samples:



Answers 2. (b) The trends in the graphs are as following:
The trend for the training error is constant linar line equal to zero.
The trend for the test error is at the beginning (0 to 50 samples) decreasing significantly after that point the function vibrates but the trend stay constant between 0 to 0.1.
The trend for the difference between them is at the beginning (0 to 50 samples) decreasing significantly after that point the function vibrates but the trend stay constant between 0 to 0.1.
Answers 2. (c) The reason for the trends observed in the graph:
We can observe an 'overfitting' at the beginning of the graph. This is because the error rate on the training sample is low (actually 0), but on the training sample the error is relatively high. The phenomena of 'overfitting' we are experiencing is not surprising, this is since the less train sample we had, the less it represented the distribution. This is because it couldn't see the whole picture and returned a result that is far from being optimal for the distribution but was perfect for the sample. We will notice that we can't tell a strong fact about the approximation error, since we didn't cover all existed separators, but we know it is not more than the lowest error we found which was around 0.05. Since the approximation error is a constant number, we can conclude that the more samples we have, the estimation error decreases ($err_{est} = err(h_s, D) - inf_{h \in H} err(h, D)$).
Although both graphs (5/6 and 3/5) have same trends, we can see that the graph of 6/5 needed a lower number of samples to get lower error compared to 3/5. Since the error of 6/5 is lower from 3/5 for all sample sizes, we can assume that its approximation error is lower, but we cannot know for sure, since we didn't cover all existed separator and we don't know the accurate distribution for both problems. This means that the problem of differentiating between 5 to 6 is probably relatively easier than differentiating between 3 to 5 for a machine learning algorithm using a linear separator attitude.

Question 3. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a finite set of examples, and let $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that $\mathcal{D}$ has a Bayes-error of zero and that it is c-Lipschitz.

(a) Let $S \sim \mathcal{D}^m$. Prove that for any two pairs $(x_1, y_1), (x_2, y_2) \in S$, if $y_1 \neq y_2$ then $\|x_1 - x_2\| \geq 1/c$.

(b) Let $\mathcal{B}$ be a set of balls of radius $1/(3c)$ that *cover* the space of points $\mathcal{X}$ (so that every point from $\mathcal{X}$ is in at least one ball in $\mathcal{B}$). Let $S \sim \mathcal{D}^m$. Suppose that for every ball $B \in \mathcal{B}$, there is some pair $(x, y) \in S$ which satisfies $x \in B$. Prove that under this assumption, and the other assumptions on $\mathcal{D}$ given above, $\mathrm{err}(f_S^{nn}, \mathcal{D}) = 0$, where f_S^{nn} is the 1-nearest-neighbor function defined in class.

Answers 3. (a) We assume that the conditional probability function is c-Lipschitz for some c > 0: Namely, for all $x_1,\ x_2\ \in X,\ |\eta(x_1) - \eta(x_2)| \leq c \cdot \|x_1 - x_2\|$ since c > 0 than: for all $x_1,\ x_2\ \in X,$ $\frac{|\eta(x_1) - \eta(x_2)|}{c} \leq \|x_1 - x_2\|$. without limitaion of generality $y_1 = 1,\ y_2 = 0$, because the mistake of Bayes is 0 than $\eta(x)$ beave as deterministic function that return 0 or 1 for all x $\in X$, since we know that the labels are different and the probability to get x_1 is more than 0, we can applies: $\eta(x)\ =\ P[y\ =\ 1|x_1] = 1$ and $\eta(x_2)\ =\ P[y\ =\ 1|x_2] = 0$, hence $\frac{|\eta(x_1) - \eta(x_2)|}{c} \leq \|x_1 - x_2\|\ \rightarrow\ \frac{|1-0|}{c} \leq \|x_1 - x_2\|\ \rightarrow\ \frac{1}{c} \leq \|x_1 - x_2\|$.

Answers 3. (b) Lets assume by contradiction that exists a pair (x,y) s.t. $f_s^{nn}(x)\ \neq y$. Since $f_s^{nn}(x)\ \neq y$, it can’t be that $x\ \in\ S$, otherwise x is the nearest neighbor of itself and $f_s^{nn}(x)\ = y$. Hence, exists some $(x', y')\ s.t.\ x'\ \in\ S \wedge x'\ is\ the\ nearest\ neighbor\ of\ x\ \wedge\ y' \neq y$. From proof 3(a) we can conclude that the distance between x and x’ is at least 1/c (Its easy to see that proof 3(a) is true for all subgroup of D rather than just the any sample S). Now, lets observe the ball $b\ \in\ B\ s.t\ x\ \in\ b$. From the assumption given in the question, since $x \notin S$ there must be other $x''\ s.t.\ x''\ \in\ b \wedge x''\ \in\ S$. Since x’ and x’’ are on the same ball of radius $\frac{1}{3c}$ the maximum distance between them is $\frac{2}{3c} < \frac{1}{c}$ for $c > 0$. Hence x’’ is nearer to x then x’ by contradiction. Therefore there is no pair on which the algorithm f_s^{nn} returns a wrong answer, meaning $err(f^{nn}{}_S\ ,\ D)\ =\ 0$.

Answers 4. We basically want to find two linear separators w_1, w_2 that applies the conditions of $h_{w1,w2}(x)$. We will find those separators with single run of the linear programming solver. In order to accomplish that, we will put both constraints, those of w1 and w2 in the matrix A and expect to receive an output vector w which is combination of w1 and w2.
In a formal way, for sample S (|S|=m) find w1 and w2 as following:

1. Run a linear programming solver with the following configuration:
   $u = (0, ...0)$
   $v = (1, ..., 1)$
   $A \in M_{2m\,X\,2d}$
   $1 \le i \le m,\ 1 \le j \le d:\ A_{i,j} = sign(2 - y_i)x_{i,j}\ ,\ A_{i,d+m} = 0$
   $m < a \le 2m,\ d < b \le 2d:\ A_{a,b-d} = 0\ ,\ A_{a,b} = g(y_{a-m})x_{a-m,b-d}\ s.t.\ g(x) = (-1)\ iff\ x\ is\ even,\ otherwise\ g(x) = 1$
   (The matrix A is two blocks matrix with the constraints of w1 in the first block and the constraints of w2 in the second block)
2. If the linear solver returns a vector w return:
   $w_1 = first\ d\ coordinates\ of\ w,\ w_2 = second\ d\ coordinates\ of\ w$.
   Else, return "no solution".

Correctness:

The algorithm returns $w_1 w_2$ for some sample S, we will show that for some $(x_i, y_i) \in S\ h_{w_1 w_2}(x_i) = y_i$:

case $y_i = 1$

From the linear solver of w_1 we know that $A * w\ \ge (1.....1)$ therefore the i-th row applies:
$A_i * w \ge 1\ =>\ (sign(2 - y_i)x_{i,1}, ...sign(2 - y_i)x_{i,d},\ 0.....0) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \ge\ 1 =>$
$< sign(2 - y_i)x_i\ ,\ w_1 >\ \ge 1\quad =>\quad < sign(1)x_i,\ w_1\ >\ \ge 1\quad =>\quad < x_i, w_1 >\ \ge 1 => sign(< x_i, w_1 >) = 1$
And the (i+m)-th row applies:
$A_{i+m} * w \ge 1\ =>\ (\ 0.....0, g(y_i)x_{i,1}......g(y_i)x_{i,d}) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \ge 1\ =>$
$< g(y_i)x_i\ ,\ w_2 >\ \ge 1\quad =>\quad < x_i\ ,\ w_2\ >\ \ge 1\quad => sign(< x_i, w_2 >) = 1$.
$h_{w_1 w_2}(x_i\ )\ =\ 1$

case $y_i = 2$

From the linear solver of w_1 we know that $A * w\ \ge (1.....1)$ therefore the i-th row applies:
$A_i * w \ge 1\ =>\ (sign(2 - y_i)x_{i,1}, ...sign(2 - y_i)x_{i,d},\ 0.....0) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \ge\ 1 =>$
$< sign(2 - y_i)x_i\ ,\ w_1 >\ \ge 1\quad =>\quad < sign(0)x_i,\ w_1\ >\ \ge 1\quad =>\quad < x_i, w_1 >\ \ge 1 => sign(< x_i, w_1 >) = 1$
And the (i+m)-th row applies:
$A_{i+m} * w \ge 1\ =>\ (\ 0.....0, g(y_i)x_{i,1}......g(y_i)x_{i,d}) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \ge 1\ =>$
$< g(y_i)x_i\ ,\ w_2 >\ \ge 1\quad =>\quad < -x_i\ ,\ w_2\ >\ \ge 1\quad => < x_i, w_2 >\ \le\ -1 => sign(< x_i, w_2 >) =\ -1$.
$h_{w_1 w_2}(x_i\ )\ =\ 2$

case $y_i = 3$

From the linear solver of w_1 we know that $A * w\ \ge (1.....1)$ therefore the i-th row applies:
$A_i * w \ge 1\ =>\ (sign(2 - y_i)x_{i,1}, ...sign(2 - y_i)x_{i,d},\ 0.....0) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \ge\ 1 =>$
$< sign(2 - y_i)x_i\ ,\ w_1 >\ \ge 1\quad =>\quad < sign(-1)x_i,\ w_1\ >\ \ge 1\quad =>\quad < x_i, w_1 >\ \le\ -1 => sign(< x_i, w_1 >) =\ -1$
And the (i+m)-th row applies:
$A_{i+m} * w \ge 1\ =>\ (\ 0.....0, g(y_i)x_{i,1}......g(y_i)x_{i,d}) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \ge 1\ =>$
$< g(y_i)x_i\ ,\ w_2 >\ \ge 1\quad =>\quad < x_i\ ,\ w_2\ >\ \ge 1\quad => sign(< x_i, w_2 >) = 1$.
$h_{w_1 w_2}(x_i\ )\ =\ 3$

case $y_i = 4$

From the linear solver of w_1 we know that $A * w \geq (1.....1)$ therefore the i-th row applies:

$A_i * w \geq 1$ => $(sign(2-y_i)x_{i,1}, ...sign(2-y_i)x_{i,d},\ 0.....0) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \geq\ 1$ =>

$< sign(2-y_i)x_i\ ,\ w_1 >\ \geq 1$ => $< sign(-1)x_i,\ w_1\ >\ \geq 1$ => $< x_i, w_1 >\ \leq\ -1$ => $sign(< x_i, w_1 >) =\ -1$

And the (i+m)-th row applies:

$A_{i+m} * w \geq 1$ => $(\ 0.....0, g(y_i)x_{i,1}......g(y_i)x_{i,d}) * ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d) \geq 1$ =>

$< g(y_i)x_i\ ,\ w_2 >\ \geq 1$ => $< -x_i\ ,\ w_2\ >\ \geq 1$ => $< x_i, w_2 >\ \leq\ -1$ => $sign(< x_i, w_2 >) =\ -1$

$h_{w_1 w_2}(x_i\ )\ =\ 4$

We will now show that in case the algorithm returns “no solution” there is indeed no w_1, w_2 as required:

Assuming the algorithm couldn’t find w - Assuming in contradiction that exists a w_1, w_2 that applies the conditions. That means that for all (x,y) $sign(< x, w_1 >) = 1\ if\ y \in \{1,2\}\ \wedge sign(< x, w_1 >) =-1\ if\ y \in \{3,4\}$ and $sign(< x, w_2 >) = 1\ if\ y \in \{1,3\}\ \wedge sign(< x, w_2 >) =-1\ if\ y \in \{2,4\}$. We will notice that the vector $w = ((w_1)_1....(w_1)_d, (w_2)_1....(w_2)_d)$ applies $A * w\ > (0.....0)$, from the definition of A. We showed in class that if there is a vector w’ that A*w’>(0….0) then exists a vector w s.t. $A * w \geq (1....1)$. But Because of the validity of the linear programming solver and the fact that it didn’t return any vector, there is no such w and hence our assumption is false and there are no w1 and w2 as required.