236606 - Deep Learning - Homework 1

Eyal Ben David - 305057416

April 26, 2018

1 Distance of Point to Hyperplane

A.

The optimization problem will obtain the following form:

$\min_x \|x - x_0\|^2$

sub to $w^T x + b = 0$

B.

solving the optimization problem we get

$L(\lambda, x) = \|x - x_0\|^2 - \lambda(w^T x + b)$

$\frac{\partial}{\partial x} L(\lambda, x) = 2(x - x_0) - \lambda w = 0$

$\frac{\partial}{\partial \lambda} L(\lambda, x) = w^T x + b = 0$

we will isolate x from the first equation and get $x = \frac{1}{2}\lambda w + x_0$ and put it into the second equation to get:

$\frac{1}{2}\lambda \|w\|^2 + w^T x + b = 0 \Longrightarrow \lambda = \frac{-(w^T x + b)}{\frac{1}{2}\|w\|^2}$.

next we will put this expression back into the first equation and get

$x = \frac{-(w^T x_0 + b)}{\|w\|^2} w + x_0 \Longrightarrow$

$\Longrightarrow \|x - x_0\| = \left\| \frac{-(w^T x_0 + b)}{\|w\|\|w\|} w \right\| = \left\| \frac{w}{\|w\|} \right\| \left\| \frac{w^T x_0 + b}{\|w\|} \right\| = \left\| \frac{w^T x_0 + b}{\|w\|} \right\|$

2 Perceptron

A.

suppose that w* is a vector which holds: $yw * x \lneqq 0$, and w* is not normalized.

lets show that we can normalize w* and X to be on the unit hyper sphere and still without changing the predictions made by w*.

define $w' = \frac{w*}{\|w*\|}$ and define $x_{max} = max\,|x|$and then $x' = \frac{x}{x_{max}}$

now, lets show that: $yw * x \lneqq 0$ => $yw'x' \lneqq 0$:

$yw'x' = y\frac{w*}{\|w*\|}\frac{x}{max|x|} \lneqq 0$.

as can be seen above, normalizing the data and w* to be on the unit hyper sphere does not change the prediction made by w*.

B.

lets show that:

$\|W_{t+1}\| \leq \|W_t\| + \frac{1}{2\|W_t\|} + \frac{\gamma}{2}$

in the case where we are not updating, this equation is trivial.

if we are updating then:

$\|W_{t+1}\|^2 = \|W_t \pm X\|^2 = \|W_t\|^2 \pm 2W_tX + \|X\|^2 = \|W_t\|^2 \pm 2W_tX + 1$

if we are updating because $yW_tX < 0$:

$\|W_{t+1}\|^2 \leq \|W_t\|^2 + 1$, because if y is 1 then $-2W_tX < 0$ and if y equals $+1$ then$+2W_tX > 0$.

but in the case we are updating because $-\frac{\gamma}{2} < \frac{W_tX}{\|W_t\|} < \frac{\gamma}{2}$:

$\|W_{t+1}\|^2 \leq \|W_t\|^2 \pm (\mp 2\frac{\gamma}{2}\|W_t\|) + 1 = \|W_t\|^2 + (2\frac{\gamma}{2}\|W_t\|) + 1 =$

$= \|W_t\|^2 + 2\|W_t\|(\frac{\gamma}{2} + \frac{1}{2\|W_t\|}) \leq$

$\leq \|W_t\|^2 + 2\|W_t\|(\frac{\gamma}{2} + \frac{1}{2\|W_t\|}) + (\frac{\gamma}{2} + \frac{1}{2\|W_t\|})^2 = (\|W_t\| + \frac{1}{2\|W_t\|} + \frac{\gamma}{2})^2$

hence:

$\|W_{t+1}\|^2 \leq (\|W_t\| + \frac{1}{2\|W_t\|} + \frac{\gamma}{2})^2$ and $\|W_{t+1}\| \leq \|W_t\| + \frac{1}{2\|W_t\|} + \frac{\gamma}{2}$.

now, if we can claim that there is a 't' such that $\|W_t\| \geq \frac{2}{\gamma}$,

we would like to max up M, number of updating iterations.

we will notice that:

$\|W_{t+1}\| \leq \|W_t\| + \frac{3\gamma}{4}$.

from lecture (2) we that:

*$\|W_{t+1}\| \leq \|W_t\| + 1$, for every t>0.

**$t\gamma \leq \|W_{t+1}\|\,\|W*\| = \|W_{t+1}\|$, for every t>0..

It can easily be seen that for $K \geq \frac{2}{\gamma^2}$:

$\frac{2}{\gamma} \leq K\gamma \Longrightarrow \frac{2}{\gamma} \leq K\gamma \leq \|W_{K+1}\|$ (by using **).

If we will choose the first K that fit this equation, and by choosing $M > K$, we can write:

$\|W_{M+1}\| \leq \|W_M\| + \frac{3\gamma}{4} \leq \|W_{M-1}\| + (1+1) * \frac{3\gamma}{4} \leq ... \leq$

$\leq \|W_{K+1}\| + (M-K)\frac{3\gamma}{4} \leq \|W_{K+1}\| + M * \frac{3\gamma}{4}$.

Using (*), $\|W_{K+1}\| \leq \|W_K\| + 1$, its easy to show that:

$\|W_{M+1}\| \leq \|W_{K+1}\| + M * \frac{3\gamma}{4} \leq \|W_K\| + 1 + M * \frac{3\gamma}{4}$.

while remembering that $\|W_K\| \leq \frac{2}{\gamma}$:

hence:

$\|W_{M+1}\| \leq \|W_K\| + 1 + M * \frac{3\gamma}{4} \leq \frac{2}{\gamma} + 1 + M * \frac{3\gamma}{4} \Longrightarrow$

$\Rightarrow M\gamma \leq \|W_{M+1}\| \leq \frac{2}{\gamma} + 1 + M * \frac{3\gamma}{4} \Longrightarrow$

$\Rightarrow M * \frac{\gamma}{4} \leq 1 + \frac{2}{\gamma} \Longrightarrow M \leq \frac{4}{\gamma} + \frac{8}{\gamma^2}$.

Finally, because the data is in the unit hyper sphere, in worst case scenario - the two points are on the sphere and their distance is 2, and then $\gamma = 1$. So we can say that: $0 < \gamma \leq 1$ and:

$M \leq \frac{4}{\gamma} + \frac{8}{\gamma^2} \leq \frac{4}{\gamma^2} + \frac{8}{\gamma^2} = \frac{12}{\gamma^2}$.

3 Ridge RegressionM

A.

Given the cost function $L(w,b) = \frac{1}{2M}\left\|Xw + b1 - y\right\|^2 + \lambda\left\|w\right\|^2$

$\frac{\partial}{\partial b}L(w,b) = \frac{1}{M}1^T(Xw + b1 - y) \overset{1}{=} \frac{1}{M}(1^Tb1 - 1^Ty) = \frac{1}{M}(Mb - 1^Ty)$

$\implies \frac{\partial}{\partial b}L(w,b) = \underbrace{b - \frac{1}{M}1^Ty}.$

*1 is correct because we remember that $1^Tx = 0$

B.

$\frac{\partial}{\partial w}L(w,b) = \frac{1}{M}X^T(Xw + b1 - y) + 2\lambda Iw =$

$= \frac{1}{M}X^TXw + \frac{1}{M}X^Tb1 - \frac{1}{M}X^Ty + 2\lambda Iw =$

$= (\frac{1}{M}X^TX + 2\lambda I)w + \frac{b}{M}X^T1 - \frac{1}{M}X^Ty = \underbrace{(\frac{1}{M}X^TX + 2\lambda I)w - \frac{1}{M}X^Ty}.$

C.

$b - \frac{1}{M}1^Ty = 0 \implies \underbrace{b = \frac{1}{M}1^Ty}$, which means that b is the empirical avarage of y.

$(\frac{1}{M}X^TX + 2\lambda I)w - \frac{1}{M}X^Ty = 0 \implies \underbrace{w = (\frac{1}{M}X^TX + 2\lambda I)^{-1}\frac{1}{M}X^Ty}$

These values will obtain the global minimum for the objective function. This is true because the objective function is a convex function and it is known that for a convex function there is only one minima which is the global minima.

D.

we will prove that $(\frac{1}{M}X^TX + 2\lambda I)$ is positive definite.

First note that a matrix of the form X^TX is psd.

we will prove by definition:

$v^T(\frac{1}{M}X^TX + 2\lambda I)v = v^T\frac{1}{M}X^TXv + v^T2\lambda Iv > 0$

The first term is greater or equal to 0 because X^TX is psd.

The second term is greater than 0 because it evaluates to:

$v^T2\lambda Iv = \sum_{i=1}^{D} 2v_i^2\lambda > 0, \forall\lambda > 0, \forall v$

Hence $(\frac{1}{M}X^TX + 2\lambda I)$ is positive definite.

E.

The main advantage is that there is always an inverse matrix for the analytic solution and we don't need to use the pseudo inverse.

4 Binary Classification Using Ridge Regression

A+B.

added code in python.

C.

Analytic Loss:

λ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^{-2}
Loss 0-1 training set	0.1080	0.1082	0.1098	0.1173	0.1364	0.1631	0.1846	0.2728
Loss 0-1 validation set	0.0995	0.0998	0.0996	0.1052	0.1235	0.1541	0.1760	0.2711
Squared Loss training set	0.3885	0.3890	0.3908	0.4029	0.4471	0.5885	0.8789	0.9839
Squared Loss validation set	0.3721	0.3719	0.3712	0.3790	0.4241	0.5760	0.8763	0.9837

Gradient Descent Loss:

λ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^{-2}
Loss 0-1 training set	0.2121	0.1933	0.2118	0.2348	0.2033	0.2049	0.2243	0.1951
Loss 0-1 validation set	0.199	0.1844	0.1966	0.2302	0.1991	0.1979	0.2124	0.1938
Squared Loss training set	0.8100	0.7803	0.8185	0.8092	0.7718	0.7983	0.8938	0.9841
Squared Loss validation set	0.8050	0.7761	0.8100	0.8054	0.7677	0.7935	0.8915	0.9838

D.

As we saw, the model which performed the best score on the validation set in terms of 0-1 Loss is - $\lambda_{opt-analytic} = 10^{-5}$, $\lambda_{opt-gd} = 10^{-4}$.
using this model on the test set gave these scores:

Analytic Loss:

λ	10^{-5}
Loss 0-1test set	0.1074
Squared Loss test set	0.39118

Gradient Descent Loss:

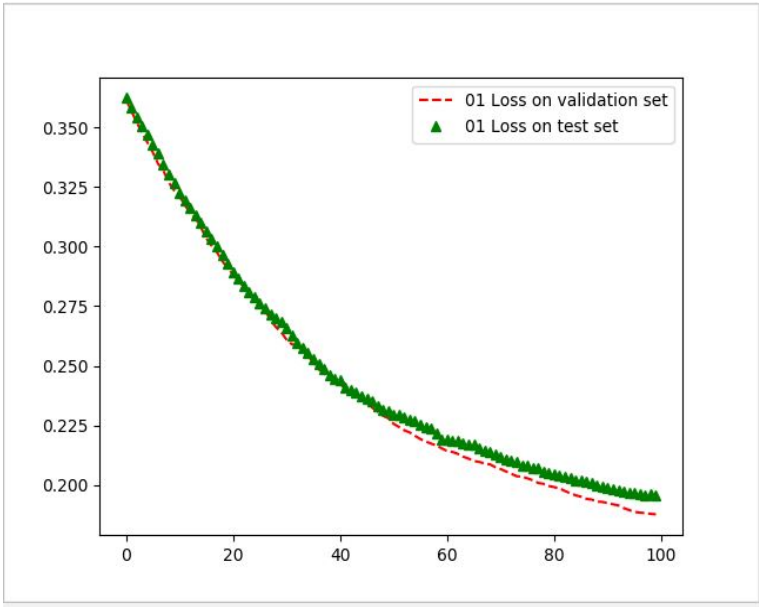
λ	10^{-4}
Loss 0-1 test set	0.1853
Squared Loss test set	0.77001

E.

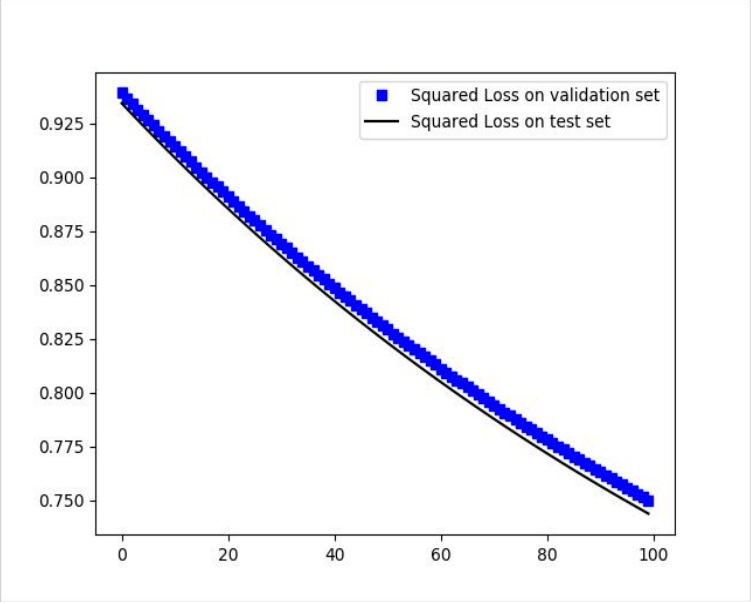
As we saw, the model which performed the best score on the validation set in terms of 0-1 Loss using the Gradient Descent scheme is - $\lambda_{opt-gd} = 10^{-4}$.
For this model we will plot the learning curves consisting of the train

and test Losses which are defined in 4(C), as a function of the number of gradient descent steps:

Graph- 0-1 Loss Graph as a function of number of iterations



Graph- Squared Loss Graph as a function of number of iteratins



F.

The simpliest possible scenario for which we think the use of linear regression for classification is a bad idea is that in lots of cases the data is unseperable with a linear classifier.