CS 189 Introduction to Machine Learning
Fall 2017 HW10

This homework is due **Monday, November 6 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit all code needed to reproduce your results, "HW[n] Code".
3. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

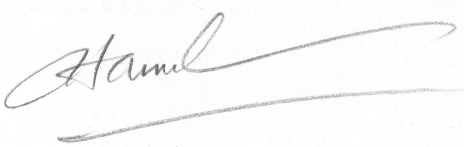
(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None
Comments: N/A

(b) Please copy the following statement and sign next to it:

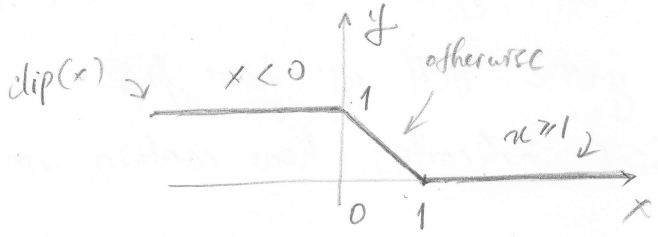
I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solutions are entirely in my words & that I have not looked @ another student's solutions. I have credited all external sources in this write up



Problem # 2

(a)



The function is not convex.

The convex condition is:

$$\alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha) x_2)$$

$\forall x_1, x_2 ; \alpha \in [0, 1]$

Take $\alpha = 0.5$, $x_1 = -1$, $x_2 = 1$

$$LHS = 0.5\, clip(-1) + 0.5\, clip(1) = 0.5 \times 1 + 0.5 \times 0 = 0.5$$

$$RHS = clip(0.5 \times (-1) + 0.5 \times 1) = clip(0) = 1 - 0 = 1$$

$LHS < RHS \rightarrow$ not convex

(b) The loss function checks the signs of y & $w^T x$. If y and $w^T x$ have opposite signs → there is a miss → penalty if they have the same signs → two scenarios:

1.) if $w^T x$ is close to 0 → penalty a little bit by $1 - y w^T x = 1 - \| w^T x \|$ (since y & $w^T x$ have same sign)

2.) If $w^T x$ is large → it is clear that it has label y → no penalty

∴ the region from $0 \rightarrow 1$ of $y w^T x$ can be considered uncertain area.

It is reasonable to look at yw^Tx because if we have only two classes, x has label y iff y & w^Tx have the same signs and the magnitude of yw^Tx tell us how far x is away from the boundary, which indicates how certain we are about the label of x.

(c)

$$R_S[w] = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} \text{loss}(w^Tx_i, y_i) = 0$$

$$\sum_{i=1}^{n} \text{clip}(y_i w^T x_i) = 0 \qquad (*)$$

clip() is a positive function (see graph)

$(*) \Rightarrow$ $\text{clip}(y_i w^T x_i) = 0 \quad \forall (x_i, y_i) \in S$

$\Rightarrow$ $y_i w^T x_i \geq 1$

$\Rightarrow$ $w^Tx_i \geq 1$ or $w^Tx_i \leq -1$

i.e $|w^Tx_i| \geq 1 \quad \forall x_i$

The classification margin is the distance from the boundary to the nearest data point. Consider data point x, the distance from x to the hyperplane boundary is

$$d = \left| \frac{w^Tx}{\|w\|_2^2} \right| \quad \leftarrow \text{absolute value}$$

$$= \frac{|w^Tx|}{\|w\|_2^2} \quad \begin{matrix} \geq 1 \\ < 1 \end{matrix} \quad \Rightarrow \quad d > 1$$

(d) $E_{\mathcal{D}}[R[w]] = E_{\mathcal{D}}\left[\frac{1}{n}\sum_{1}^{n} \text{loss}(w^T x_i, y_i)\right]$

$= \frac{1}{n}\sum_{1}^{n} \underbrace{E_{\mathcal{D}}[\text{loss}(w^T x_i, y_i)]}_{R[w]}$ $\quad (x_i, y_i)$ sampled i.i.d from $\mathcal{D}$

$= \frac{1}{n}\, n\, R[w] = R[w]$

(e) $\text{Var}[R_S[w]] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \text{loss}(w^T x_i, y_i)\right]$

$= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(\text{clip}(y_i w^T x_i)\right)$

$\text{clip}(y_i w^T x_i) \le 1 \quad \forall \ (x_i, y_i)$

$\Rightarrow \text{Var}(\text{clip}(y_i w^T, x_i) \le 1$

$\Rightarrow \sum_{i=1}^{n} \text{Var}(\text{clip}(y_i w^T x_i)) \le n$

$\Rightarrow \text{Var}[R_S[w]] \le \frac{1}{n^2}\, n = \frac{1}{n}$

(f) It is possible. Consider S has only one point (x_1, y_1) and this point is correctly classified $\Rightarrow \text{loss}(w^T x_1, y_1) = 0$

$\Rightarrow R_S[w] = \frac{1}{n}\sum_{i=1}^{n} \text{loss}(w^T x_i, y_i) = \frac{1}{1}\text{loss}(w^T x_1, y_1) = 0$

but if the classifier is not perfect $\rightarrow$ there are some losses over $\mathcal{D}$ $\Rightarrow$ $R[w] = E_{\mathcal{D}}[\text{loss}(w^T x, y)] > 0$

since loss is a positive function

(thus its expected value is positive)

Problem #3

(a) See code attached

(b) See code attached.

If we increase the learning rate, there are two scenarios.

1.) If the learning rate is still less than the critical value, the error converges faster

2.) If the learning rate is greater than the critical value, the error blows up.

Comparison: batch GD converges faster than full GD (see plots)

(c) See code attached

Comparison: Full LS is the best if it hits all features next winner is stochastic gradient descent for all features, then full GD for all features. GDs for one feature are uncertain but stochastic is better than full (see plots)

(d) See code attached

Comparison: Kaczmarz SGD is worse than full GD for all features but it seems to be stable. (see plots)

(e) See code attached

Comparison: Basically, the non-linear cases have the same behaviors as the linear cases for all methods but with larger error initially.

Problem #4

(a) See code attached

(b) See code attached

(c) See code attached

(d) See code attached

(e) See code attached

(f) See code attached

(g) See code attached

(h) See code attached

For k small, sum works well but for k larges, QDA is better

=> Choose QDA w/ $k = 800$

Problem #5

To continue problem 2(c), assume the hyperplane not only defined by w but having bias b. that is the boundary hyperplane is given by $w^T x = b$

(a) Find the distance from a data point x^* to this plane

(b) What is the new classification margin

Solution

we know that without bias, the plane is

$$w^T x = 0 \qquad (P)$$

having unit normal $\frac{w}{\|w\|_2^2}$, and this plane going through the origin $\Rightarrow$ the distance from x^* to (P) is $\frac{w^T x^*}{\|w\|_2^2} = d$

Consider

$$w^T x = b$$

$$w_1 x_1 + w_2 x_2 + \dots + w_n x_n - w_n \cdot \frac{b}{w_n} = 0$$

$$w^T x' = 0 \qquad \text{where } x' = \begin{bmatrix} x_1 \\ \vdots \\ x_n - \frac{b}{w_n} \end{bmatrix} = x - \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{b}{w_n} \end{bmatrix}}_{\bar{x}}$$

For x^* : $x' = x^* - \bar{x}$

$\Rightarrow$ new distance : $\frac{w^T(x^* - \bar{x})}{\|w\|_2^2} = d - \underbrace{\frac{w^T \bar{x}}{\|w\|_2^2}} = d - \frac{b}{\|w\|_2^2}$

(b) the new classification margin is

$$> 1 - \frac{b}{\|w\|_2^2}$$

(since we proved in 2(c) that the old classification margin is > 1)