Name: Connor Raymond Stewart
ID: 20673233

CS 886 Homework Two:

① Simplify Equation (1):

We train a linear model adversarially on the Soft-SVM loss

↳ Minimax objective:

$$(1)\quad \min_{W} E_{x,y} \max_{\|\delta\|_p \le \varepsilon} \left[\max(0, 1-yw^T(x+\delta))\right],$$ where x is the instance & $y \in \{-1,+1\}$ is the label

$\boxed{1.1}$ let $p=\infty$, let w be a fixed weight vector & let x be a data point

Show that the optimal perturbation $\delta^*(X)$ that maximizes the Soft-SVM loss has the Closed-form Solution:

$$\delta^* = -y\varepsilon\,\mathrm{Sign}(w)$$

Where

$$\mathrm{Sign}(a) = \begin{cases} 1, & \text{if } a>0 \\ 0, & \text{if } a=0 \\ -1, & \text{if } a<0 \end{cases}$$

and $\mathrm{Sign}(w)$ means running the Sign operation element wise on the vector W.

We know that $\max(0, 1-yw^T(x+\delta))$ is the loss function where $f(x) = w^T(x+\delta)$ meaning the function is in the term $\max(0, 1-yf(x))$. The SVM Partions data as follows:

We see that we can define ε as a point between margin & the Correct side of the hyperplane as: $0 \le \varepsilon \le 1$ for $y f(x) \ge 1-\varepsilon$

$\therefore \varepsilon = \max(0, 1-yf(x))$

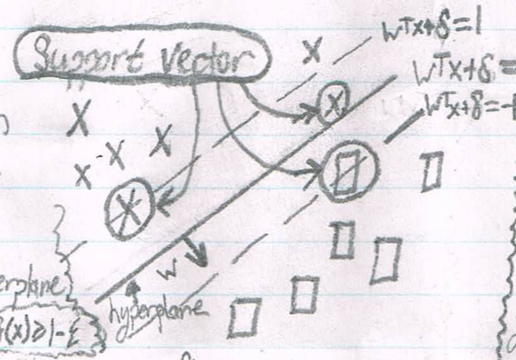


figure 1: SVM Margin

For the three points:

- $yf(x) > 1$: Point outside margin. No loss Contribution
- $yf(x) = 0$: Point on margin. No loss Contribution
- $yf(x) < 1$: Point violates margin Const. Contributes to loss

We note that SVM uses a hinge loss of the form $(0, 1-yf(x))$ to approx. 0-1 loss:

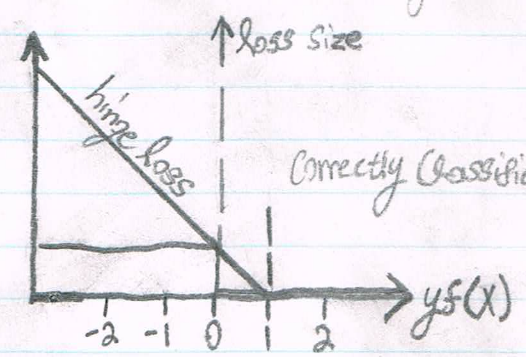
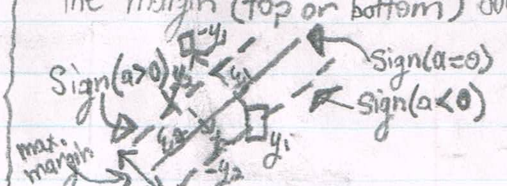


figure 2: Hinge-loss functions

Soft-SVM:

We use the value of ε to push Support vectors past the hyperplane & use $\mathrm{Sign}(w)$ to place Support vectors along the correct end of the margin (top or bottom) for its class:



□ is $y=-1$
X is $y=1$

figure 3: Soft-margin Movement

Since the following occurs in δ:

① Sign(w) sets a vector along its correct class of plane for a point

② $-y$ reverses the label of a point

↳ if $y=-1$ then $-y=1$

↳ if $y=1$ then $-y=-1$

③ ε is the point between the margin & the correct side of the plane

with ① we take a vector (w) & label points (X) to their correct category of 1 or -1, or 0 if on the plane itself.

we also get ②, $-y$, & flip point x's given label to flip its side to its mirror opposide side from $W^Tx+\delta \leq 0$ (use mirror symmetry)

we then take ③ ε, which is the point between the magin & the correct side of the plane (the other end of the hyperplane)

thus, we can see that the perturbation pushes the max. number of points over the hyperplane such that $yf(X)<1$ to make a misclassification

↳ $-y$ sign(w) reverses symmetry w.r.t. hyperplane

↳ ε moves point x outside to opposite end of the margin w.r.t. hyperplane

$\therefore$ $\delta^* = -y\varepsilon \, \text{Sign}(w)$ is the optimal perterbation that maximizes the soft-SVM loss

$\boxed{1.2}$ let $p=\infty$. Simplify adverserial training Equation (1):

We can rewrite Equation (1) as an optimization problem with Constraints

$\xi_i = \max(0, 1 - y_i w^T(x_i + \delta))$ for each $i \in \{1, \dots, n\}$

ξ_i is the Smallest Positive number Satisfying $y_i w^T(x_i+\delta) \geq 1 - \xi_i$

We can write the following optimization problem:

Solution

$$\text{(LP)} \quad \underset{w, \delta, \xi_i}{\text{minimize}} \; E_{x,y} \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|_1$$

$$\text{s.t.} \quad y_i w^T(\bar{x}_i + \delta) - S_i \|w\|_1 \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \; i = 1, \dots, n$$

- For the hyperplane, w & b could be normalized so that $w^T x + b = -1$ or $+1$ goes through Support vectors of Class -1 or $+1$ respectively.
- We can convert Eq. (1) to a LP Problem with the objective $\frac{1}{2}\|w\|_1 + E_{x,y} \sum_{i=1}^{m} \xi_i$

Notes:

Data with perturbations is expressed as $x_i = \bar{x}_i + \delta_i^*$ where the mean vector $\bar{x}_i$ plus perturbation δ^* is bounded by by the L_p norm as $\|\delta_i\|_p \leq S_i$ for all $i = 1, \dots, n$

using Hölder's inequality, we see that for duel norms L_p & L_q, with $p, q \in [1, \infty]$ and $1/p + 1/q = 1$, the following inequality holds:

$\|fg\|_1 \leq \|f\|_p \|g\|_q$ ∴ the lower bound is $-S_i\|w\|_q$

The distance between two hyperplanes is $2/\|w\|_1$ So max. margin is the Same as min. of $1/2\|w\|_1$ Subject to Separation constraints

The objective function penalizes slack variables so that optimization is a balance between a large margin and a Small error penalty

$E_{x,y}$ is a tradeoff parameter in this Context

We know that in general the Quadratic formulation is:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 \quad \text{s.t.} \; y_i w^T(x_i + \delta) \geq 1, \; i = 1, \dots, n$$

the above Solution is a LP formulation that occurs with the L_∞-norm only

$\boxed{1.3}$ Simplify adversarial training Equation (1) for general $p \geq 1$:

Simplifying & expanding upon $\boxed{1.2}$ we see that δ_i^* is bounded by the L_p-norm with $\|\delta_i\|_p \leq S_i$, $i=1,\dots,n$. So for the worst case perturbation:

$$\min_{\|\delta_i\| \leq S_i} y_i w^T(x_i+\delta) + \delta_i^* w^T y_i \geq 1-\varepsilon_i, \; i=1,\dots,n$$

thus, we attempt to minimize $\delta_i^* w^T y_i$ under the norm p

As was shown in $\boxed{1.2}$ with Hölders inequality, we see that

Perterbation $\leq \|\delta_i\|_p \|w\|_q \leq S_i \|w\|_q$ for $p,q \in [1,\infty]$ & $1/p + 1/q = 1$

We can substitute the lower bound into the original formulations from $\boxed{1.2}$

$$\min_{w,b,\varepsilon_i} \frac{1}{2}\|w\|_2^2 + E_{x,y}\sum_{i=1}^{n} \varepsilon_i$$

$$\text{s.t. } y_i w^T(\bar{x}_i+\delta) - S_i\|w\|_p \geq 1-\varepsilon_i, \; \varepsilon_i \geq 0, \; i=1,\dots,n$$

Note:

We use the quadratic formulation $\frac{1}{2}\|w\|_2^2$ s.t. $y_i w^T(x_i+\delta) \geq 1$, $i=1,\dots,n$

$E_{x,y}$ is a tradeoff parameter

We penalize the slack summation in the objective function

$\bar{x}_i$ is the mean vector

S_i is the norm bound in $\|\delta_i\|_p \leq S_i$

$p \in [1,\infty]$

ε_i is the smallest positive number satisfying $y_i w^T(x_i+\delta) \geq 1-\varepsilon_i$ & is the $\max(0, 1-y_i w^T(x_i+\delta))$ for $i=1,\dots,n$ such that $\varepsilon_i \geq 0$