CMU 10-715: Homework 2

Soft Support Vector Machine Theory and Implementation

Abishek Sridhar (Andrew Id: abisheks).

1 Soft Support Vector Machine Theory

Consider the primal problem for the soft support vector machine (soft SVM). Where $y_i \in \{-1, 1\}$ are the labels, $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ are the features (features already include the bias term), $\xi_i \in \mathbb{R}^+$ are the slack variables.

$$\begin{aligned} \underset{\mathbf{w}, \xi_i}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1 - \xi_i \quad i = 1, \ldots, n \\ & \xi_i \geq 0 \qquad\qquad\quad\;\; i = 1, \ldots, n \end{aligned} \tag{1}$$

(a) Show that the soft SVM problem can be written as a regularized Hinge Loss problem:

$$\underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\right) \tag{2}$$

Solution:
Consider the original minimization objective:

$$\begin{aligned} & \underset{\mathbf{w}, \xi_i}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} \xi_i \\ \iff \quad & \underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + \underset{\xi_1, \ldots, \xi_n}{\text{minimize}} \ C\sum_{i=1}^{n} \xi_i \qquad \textit{(first term doesn't depend on } \xi_i\textit{)} \\ \iff \quad & \underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} \underset{\xi_i}{\text{minimize}} \ \xi_i \quad \textit{(no cross dependency of terms in summation} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{and } C > 0\textit{)} \end{aligned} \tag{3}$$

Now, we can re-write the constraints in the below form:

$$\begin{aligned} & y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \quad i = 1, \ldots, n \\ \iff \quad & \xi_i \geq 1 - y_i(\mathbf{w}^T\mathbf{x}_i) \text{ and } \xi_i \geq 0 \quad i = 1, \ldots, n \\ \iff \quad & \xi_i \geq \max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\right) \qquad i = 1, \ldots, n \end{aligned} \tag{4}$$

Equation 3 implies we can individually minimize ξ_i's subject to the constraints on ξ_i's given by equation 4 alone. Given the nature of constraint, it is clear that the minimum possible value of ξ_i is $\max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\right)$ for $i = 1, \ldots, n$. Following this, equation 3 becomes:

$$
\begin{aligned}
& \underset{\mathbf{w},\xi_i}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\xi_i \\
\iff \quad & \underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\underset{\xi_\mathbf{i}}{\text{minimize}} \ \xi_i \\
\iff \quad & \underset{\mathbf{w}}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\max\big(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\big)
\end{aligned} \tag{5}
$$

From equation 5, it is clear that the Soft SVM objective is equivalent to the regularized Hinge Loss problem.

(b) Find an expression for the subgradient of the regularized Hinge Loss problem from equation (2).

Solution:
The first term of equation **??** is convex and differentiable throughout the domain of $\mathbf{w}$ ($\mathbb{R}^p$). Hence, its subgradient will be the same as gradient, that is:

$$
\nabla_\mathbf{w}\left(\frac{1}{2}||\mathbf{w}||_2^2\right) = \mathbf{w}
$$

The second term is a summation, where the terms in the summation are convex but not differentiable at $\mathbf{w}$ such that $y_i(\mathbf{w}^T\mathbf{x}_i) = 1$. In the other regions, the terms are differentiable and the subgradients (equivalent to the gradients) are given as:
Case 1: $y_i(\mathbf{w}^T\mathbf{x}_i) > 1$

$$
\begin{aligned}
& \max\big(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\big) = 0 \\
\implies \ & \nabla_\mathbf{w}\Big(\max\big(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\big)\Big) = 0
\end{aligned}
$$

Case 2: $y_i(\mathbf{w}^T\mathbf{x}_i) < 1$

$$
\begin{aligned}
& \max\big(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\big) = y_i(\mathbf{w}^T\mathbf{x}_i) \\
\implies \ & \nabla_\mathbf{w}\Big(\max\big(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\big)\Big) = -y_i\mathbf{x}_i
\end{aligned}
$$

Now, let us consider the the non-differentiable point (where left hand limit does not equal the right hand limit).

Case 2: $y_i(\mathbf{w}^T\mathbf{x}_i) = 1$
Subgradients will belong to the closed interval $[LHL, RHL]$, which is $[\mathbf{0}, -y_i\mathbf{x}_i]$. We define the sub-gradient to be $\mathbf{0}$ and show that it satisfies the subgradient property:

$$
f(\mathbf{w}) - f(\mathbf{w}_0) \geq \mathbf{0}\cdot(\mathbf{w} - \mathbf{w}_0) \quad \forall \ \mathbf{w} \in \mathbf{dom}f
$$

That is,

$$
f(\mathbf{w}) \geq f(\mathbf{w}_0) \quad \forall \ \mathbf{w} \in \mathbf{dom}f
$$

where $f(\mathbf{w})$ is $\max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i)\right)$ and $\mathbf{w}_0$ is the non-differentiable point with $f(\mathbf{w}_0) = 0$.
For $\mathbf{w}$ such that $1 - y_i(\mathbf{w}^T\mathbf{x}_i) < 0$, $f(\mathbf{w}) = 0 = f(\mathbf{w}_0)$, so the inequality $f(\mathbf{w}) \geq f(\mathbf{w}_0)$ holds.
For $\mathbf{w}$ such that $1 - y_i(\mathbf{w}^T\mathbf{x}_i) > 0$, $f(\mathbf{w}) > 0 = f(\mathbf{w}_0)$, so again the inequality holds.
Hence, our choice of subgradient as $\mathbf{0}$ is valid.

To summarize, the subgradient expression for the whole regularized hinge loss is:

$$\nabla_{\mathbf{w}}\left(f(\mathbf{w})\right) = \mathbf{w} - C\sum_i y_i\mathbf{x}_i \quad \text{where } 1 \leq i \leq n \text{ and } y_i(\mathbf{w}^T\mathbf{x}_i) < 1$$

2 Report

(e) Results and Plots for different choices of C (with random seed = 1, total training iterations = 10,000 and learning rate = 1e-5) - **Note:** The losses mentioned are averaged over the samples for easy comparison irrespective of dataset size.

(i)
- **C = 0.1:** Final Train Loss = 2.071 | Final Test Loss = 2.072
- **C = 1:** Final Train Loss = 4.193 | Final Test Loss = 4.206
- **C = 50:** Final Train Loss = 61.952 | Final Test Loss = 63.65

(ii)
- **C = 0.1:** Final Train Accuracy = 97.43% | Final Test Accuracy = 96.8%
- **C = 1:** Final Train Accuracy = 97.48% | Final Test Accuracy = 96.7%
- **C = 50:** Final Train Accuracy = 97.37% | Final Test Accuracy = 96.6%

(iii)
- **C = 0.1:** 1
- **C = 1:** 2
- **C = 50:** 3

(f) From figures 1, 2 and 3, it can be observed that the least number of support vectors are present for **C = 50** in the sampled subset (sampled with same random seed for all C). The intuitive reasoning is that as C becomes large, the weightage given to error terms ξ_i's increase, and the learning algorithm focuses more on reducing ξ_i's. As a consequence, the algorithm tries to avoid misclassified points more for large C (so much that it might start to overfit on the training set than maximizing margin for a generalizable setting). Infact, we can observe that in the limit $C \to \infty$, the Soft SVM objective given reduces to Hard SVM problem.

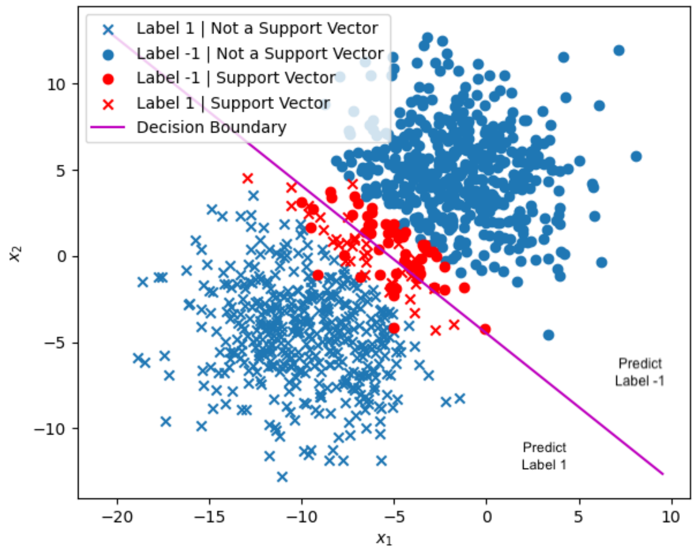

Figure 1: Scatter Plot of 10% training samples with decision boundary for $C = 0.1$

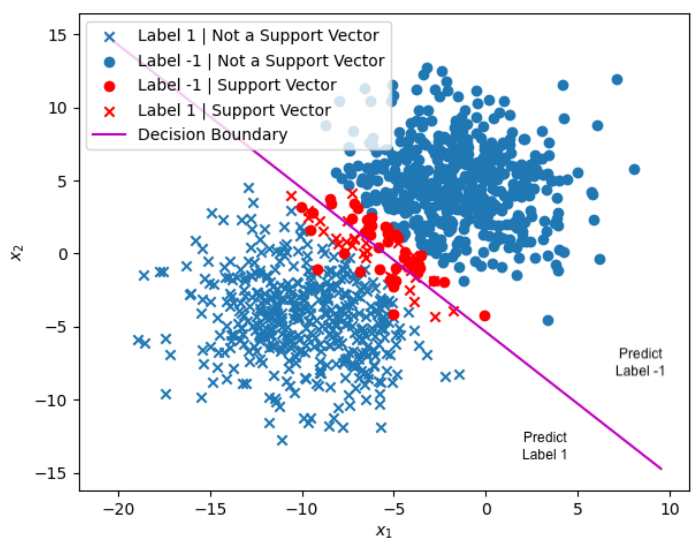

Figure 2: Scatter Plot of 10% training samples with decision boundary for $C = 1$

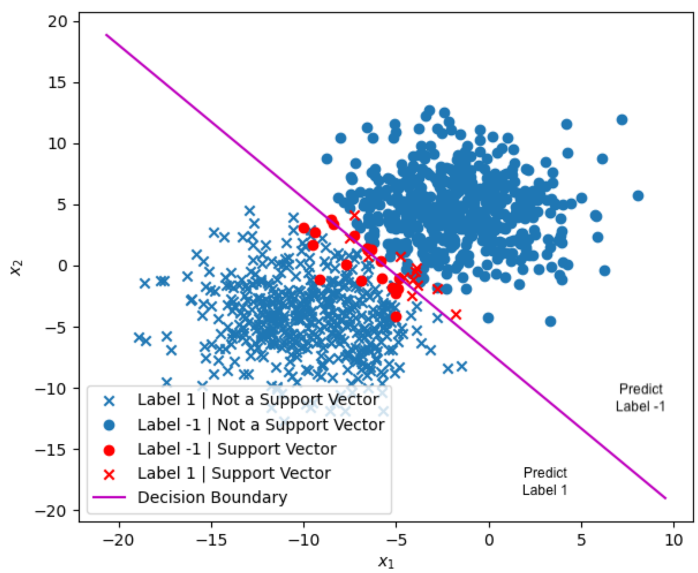

Figure 3: Scatter Plot of 10% training samples with decision boundary for $C = 50$