Lecture 14

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

`Assignment Questions Overview`

SVMs

Continuing from the previous lecture,
We saw multiple (convex) optimization objective functions in Lecture 13

For mathematical convenience,

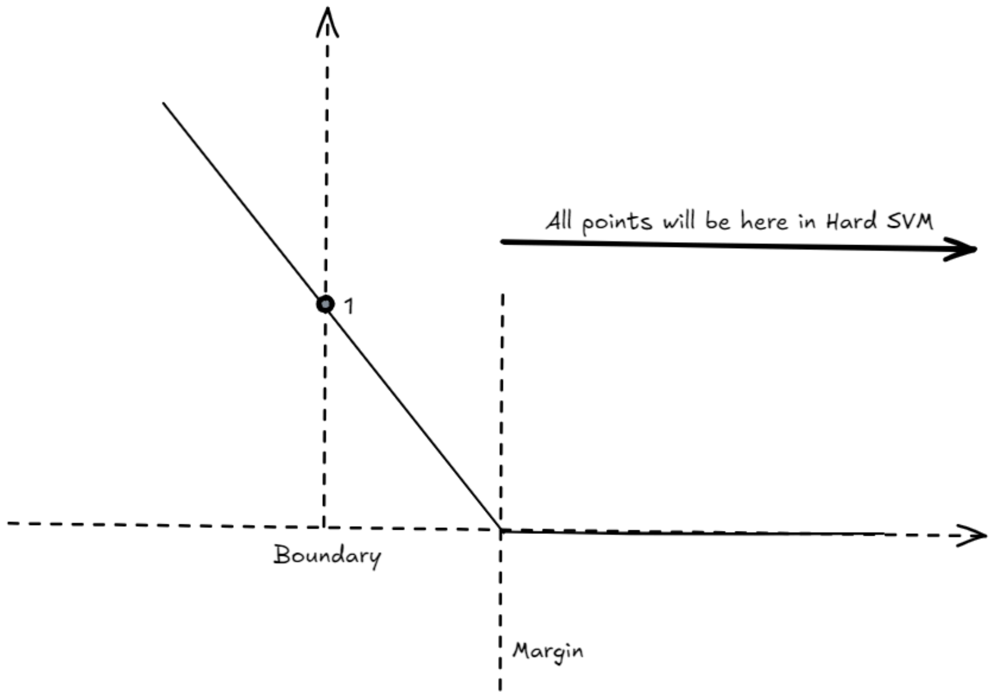
$$\max[0, y] \equiv [y]_+$$

and

$$max[0, 1-y] \equiv l_{hinge}[y]$$

Up until now, we have discussed `Hard SVM` where, we do not allow points between the margins/support vectors (and the loss between the margins increases linearly up to the boundary, where the loss would already be 1, beyond boundary would be > 1). If the classification point lies beyond the margins, then the loss is definitely zero.

But, if we add newer points which lies between the current support vectors, then the new points closest to the boundary would end up becoming the new margins (`support vector`).



y-axis in this plot is the l_{hinge}. And from this plot, it can be seen that the loss function is convex, using `Jensen's Inequality`

Soft-Margin SVM

Formulation:

$$\min\left(\|w\|_2^2 + C\sum_i \zeta_i\right)$$

where, ζ_i are the slack variables subject to -

- $\forall i \; t_i[w^T x_i + b] \geq 1 - \zeta_i$
- $\forall i \; \zeta_i \geq 0$

Equivalently,

$$\min_{w,b} \sum_i l_{hinge}(t_i[w^T x_i + b]) + Regularization$$

where we can choose the regularization of our choice, $L1$ or $L2$.
`As can be noted again, this optimization is also a convex problem`

For Logistic Regression: $\min \text{CE} + \frac{\lambda}{2}\|w\|_2^2$

Values of ζ_i:

- Correct side of margin = 0
- On the margin = 0
- Inside the margin > 0

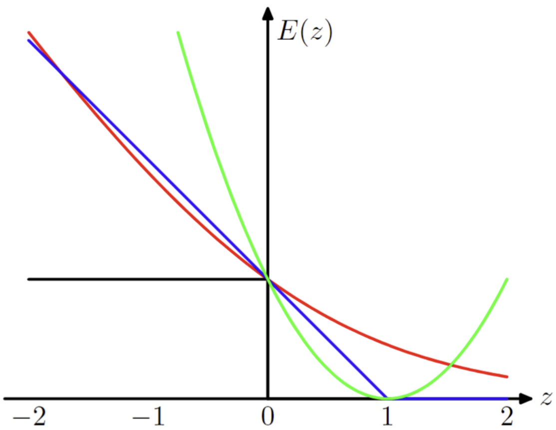
- On the boundary = 1
- Beyond the boundary > 1

Comparison with Hard SVM:

In Hard SVMs, support vectors define the hyperplane and are closest to the hyperplane, which is usually a sparse set of points. The remaining points do not matter since they cannot affect the values of w and b.

In Soft SVMs, all the misclassified points are support vectors too, and consequently all points with $\zeta_i > 0$ or just at the cusp (read `margin`).

Plot of the 'hinge' error function used in support vector machines, shown in blue, along with the error function for logistic regression, rescaled by a factor of $1/\ln(2)$ so that it passes through the point $(0, 1)$, shown in red. Also shown are the misclassification error in black and the squared error in green.



The red curve is the cross entropy over sigmoid of $w^T x_i + b$. CE(Sigmoid())
`A small caveat to keep in mind is that we considered targets as {-1, +1} but for cross entropy over sigmoid would have to relabel to {0, +1} for usage`

SV Regression

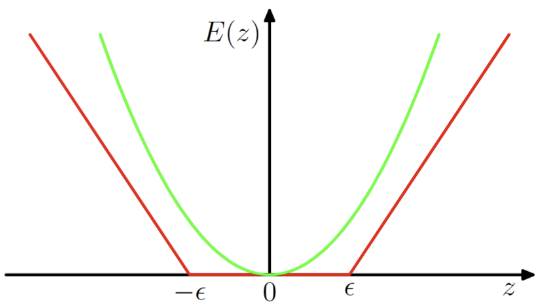
We change the mean squared error function to a ϵ-insensitive (fault-tolerant) error function.

$$C \sum_{n=1}^{N} E_\epsilon(y(x_n) - t_n) + \frac{1}{2}||w||^2$$

where,

$$E_\epsilon(y(x) - t) = \begin{cases} 0 & \text{if } |y(x) - t| < \epsilon \\ |y(x) - t| - \epsilon & \text{otherwise} \end{cases}$$

Plot of an ϵ-insensitive error function (in red) in which the error increases linearly with distance beyond the insensitive region. Also shown for comparison is the quadratic error function (in green).



Now, defining the slack variables (for fixed ϵ) -
$t_i \leq y_i + \epsilon + \xi_i$ for $\xi \geq 0$
$t_i \geq y_i - \epsilon - \hat{\xi}_i$ for $\hat{\xi}_i \geq 0$

Then, the cost function can be written as -

$$\min_{w,b} \mathrm{C} \sum_i \left(\xi_i + \hat{\xi}_i\right) + \frac{1}{2}\|w\|_2^2$$