AIML - CS 337

Lecture 11: SVM : Non-separable Instances

12 August 2022

Lecturer: Abir De *Scribe: Anand Kumar, Kajal Malik, Landa Jitendra, Shubh Kumar*

1 Review

Recall that in previous lecture, We found that $\frac{2}{\|w\|}$ as the distance between the planes $w^T x + b = 1$ and $w^T x + b = -1$, and in order to maximize this distance, We minimized $\|w\|^2$. This approach tends to work when the data-points available to us are linearly separable.

Convex Hull

If there are positive points, you put a nail at top of it and extend a rubber band, so that it covers the entire region. The shape of the rubber band when we remove it from there, this gives us the Convex Hull.

The non-separable case

We have the convex hulls corresponding to the positive and negative points, and if they overlap the condition : $y(w^T x + b) > 1 \forall x, y$. No matter what toolcase we use, This will give some error in this case.

2 How to deal with it?

There are ways to deal with it :

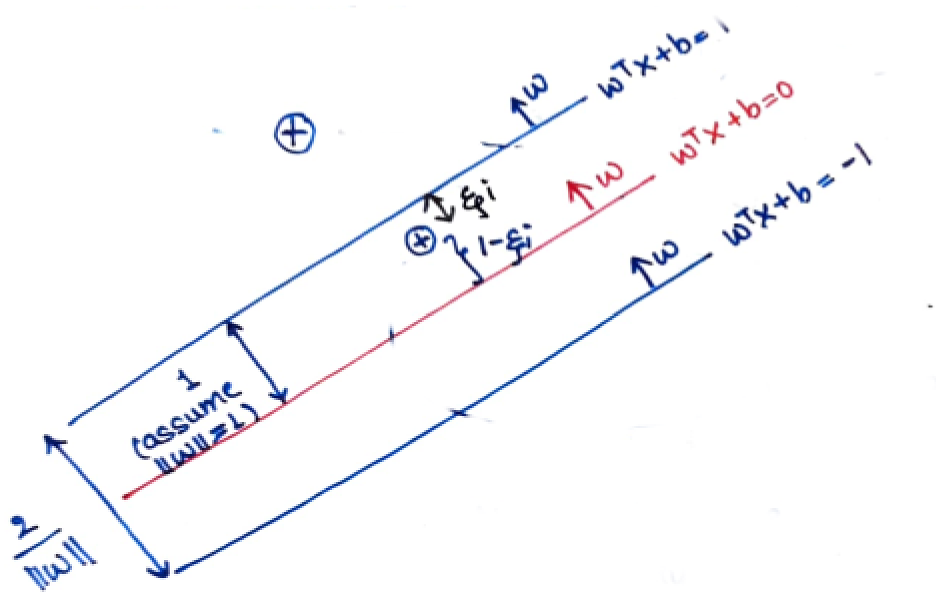
1. Remove the points in the overlap
2. Slightly Relax our constraints, So that the new constraints are feasible.
   - The Problem is, we can't simply handcraft these constraints.
   - Therefore, We'll also model these constraints and try to learn them.
     - For Example We may proceed like this: If the given expression isn't greater than 1 for all (x_k, y_k), there must exist some (x, y) for which it doesn't.
     - So we may, Replace the 1 in $y(w^T x + b) > 1 \forall x, y$ by $y(w^T x + b) > 1 - \zeta \ \forall x, y$.
     - In this case, We want $\zeta > 0$, to be as small as possible!

- This means that for some boundary case (x_k, y_k), We have that : $1-\zeta = y_k(w^T x_k + b)$
- In order to learn this, We may re-frame our problem by introducing slack variables ζ for a modified loss function.
- Simple Exercise : For all the points (x_i, y_i) that have been misclassified, Can you tell us about $y_i(w^T x_i + b) > 1 - \zeta_i$ after our optimization routine has returned.
  * Ans : If we define $\zeta_i \forall i$, Then we must have that $\zeta_i = max(0, 1 - y_i(w^T x_i + b))$
- Due to the solution of the above exercise, We may write down a separate loss function given by : $\mathcal{L}(w; X, Y) = \|w\|^2 + c \sum_{x_i, y_i} (1 - y_i(w^T x_i + b))_\dagger$

• Can be hyperparameterize the value of b in the separable case? in-separable case?
  - Ans : We can just go ahead and batch-normalize the Data, to do away with b.

A Geometric Perspective

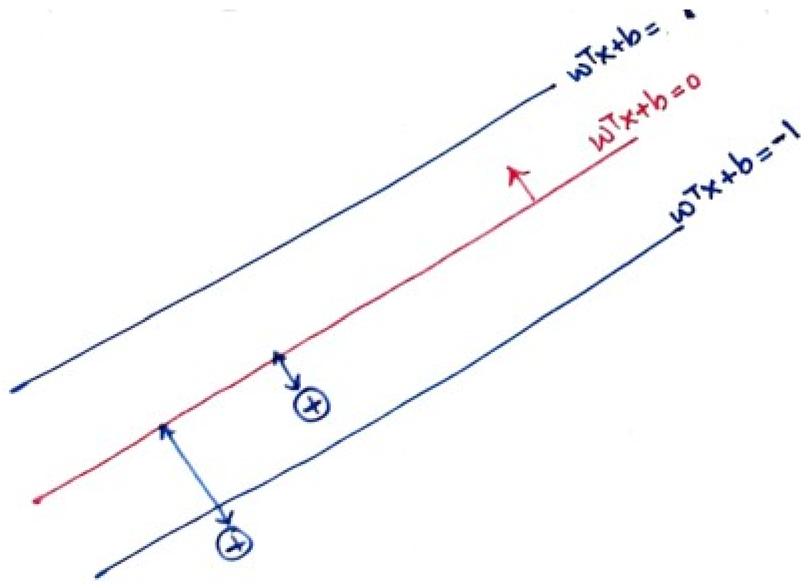
We have that $y_i(w^T x_i + b) > 1$ is not satisfied for all points of the dataset. If it were satisfied then we would be dealing with the separable case anyway. So, For the case, When $y_i(w^T x_i + b) > 1$ doesn't hold, We are trying to find the minimum ζ_i s.t. $y_i(w^T x_i + b) \geq 1 - \zeta_i$.



Note here that for the point that lies above the positive hyperplane, We have that $y_i(w^T x_i + b) > 1 \Rightarrow \zeta_i = 0$

For the other point, which is labelled as +ve but is below the said hyperplane, We'll have that $y_i(w^T x_i + b) = 1 - \zeta_i$, in any such case, We'll find that $0 \leq 1 - y_i(w^T x_i + b) \leq 1$

If the case were rather like this :



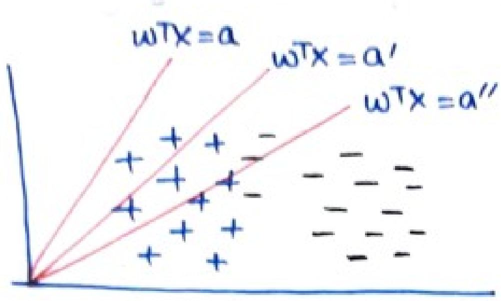
We'll instead have that $0 \leq 1 - y_i(w^T x_i + b) \geq 1 \Rightarrow \zeta_i \geq 1$

Thus, We may now conclude that the quantity $1 - y_i(w^T x_i + b)$ is negative, When the point is correctly classfied, and positive if its incorrectly specified.

3 Motivation for Batch Normalization

Tuning b

Even in the linearly separable case, If we choose $b = 0$, then we'll always have some error.



Therefore, Tuning b is an important step in almost all version of SVM(s). We may get over b and choose the pre-processing route instead in case of non-separable cases.

Another thing, We may want to do is to normalize all the features so that all the features are of comparable magnitude.

Batch Normalization

1. **Shift the Origin :** $(X_i)_{new} = X_i - E[X]$. $E[X]$ is the empirical mean vector in case of vectorized Datasets.

2. **Standard Deviation Adjustment :** $((X_i)'_{new})_j = \frac{((X_i)_{new})_j}{(\sigma_{new})_j}$. $(\sigma_{new})_j$ is the standard deviation coreesponding to the j^{th} feature.

Features of Batch-Normalization

1. Batch Normalization makes bias very small. (In Regression/Classification problems! Would apply to Neural Networks iff we add Batch Normalization after each layer.)

2. It also makes the training much more stable.

3. If we have lots of data, Batch Normalizing each batch separately, may cause issues, as each batch has different mean/ Standard Deviation.

4. This is why, we have trouble with less Data, and we have trouble with lots of Data. We just need a sweet spot between these two extremes.

4 Dual Formulation

All the discussion in this section is under the condition of convexity on the function which is to be minimized! Consider the Problem :

$$\min f(w)$$
$$\text{s.t. } g(w) \leq c$$

This is provably equivalent to :

$$\max_{\lambda \geq 0} \min_w f(w) + \lambda^T (g(w) - c)$$

Similarly, If you consider :

$$\min \lambda \|w\|^2 + \textstyle\sum_{i,j} \zeta_{i,j}$$
$$\text{s.t. } \forall i : y_i(w^T x_i + b) \geq 1 - \zeta_i$$
$$\zeta_i \geq 0$$

We may formulate the dual of this as :

$$\max_{\alpha \geq 0, \beta \geq 0} \min_{w,b} \|w\|^2 + \sum_{i,j} \zeta_{i,j} + \sum_i \alpha_i (1 - \zeta_i - y_i(w^T x_i + b)) - \beta_i \zeta_i$$

At the optimal,
For Points, which are misclassified, We have $1 - \zeta_i - y_i(w^T x_i + b) = 0$, $\zeta_i > 0$ and $\beta_i = 0$. For points, which are correctly classified $\alpha_i = 0$ and $\beta_i = 0$.

Therefore, We may generalize it to state : $\forall i : \beta_i \zeta_i = 0$ and $\forall i \alpha_i (1 - \zeta_i - y_i(w^T x_i + b)) = 0$

α_i : Penalizes the amount of misclassification
β_i : Adjusts the contribution for ζ_i for correctly classified points!

Differentiating the given expression with various variables, We get :

$$2\lambda w = \sum_i \alpha_i x_i y_i \Rightarrow w = \frac{\sum_i \alpha_i x_i y_i}{2\lambda}$$
$$\sum_i \alpha_i y_i = 0$$
$$\forall i : \alpha_i + \beta_i = 1$$

We might substitute these in the original expression for loss to get :

$$\max_{\alpha \geq 0} \sum_i \alpha_i - \sum_{i,j} \alpha_i \alpha_j (x_i^T x_j) y_i y_j$$
$$\text{Subject to : } \sum_i \alpha_i y_i = 0$$