SVM (Support Vector Machines)

Aaryan
CO21BTECH11001

Let number of features of dataset = n

Let number of training examples = m

Data consists of matrices X and y where i^{th} column of X represents the i^{th} feature of dataset and i^{th} element of y represents the value of variable dependent on set of features listed in i^{th} row of X.

Labels in y consists of {-1,1} only i.e., it is a binary classifier.

The classifier function $h(X)$ is written as

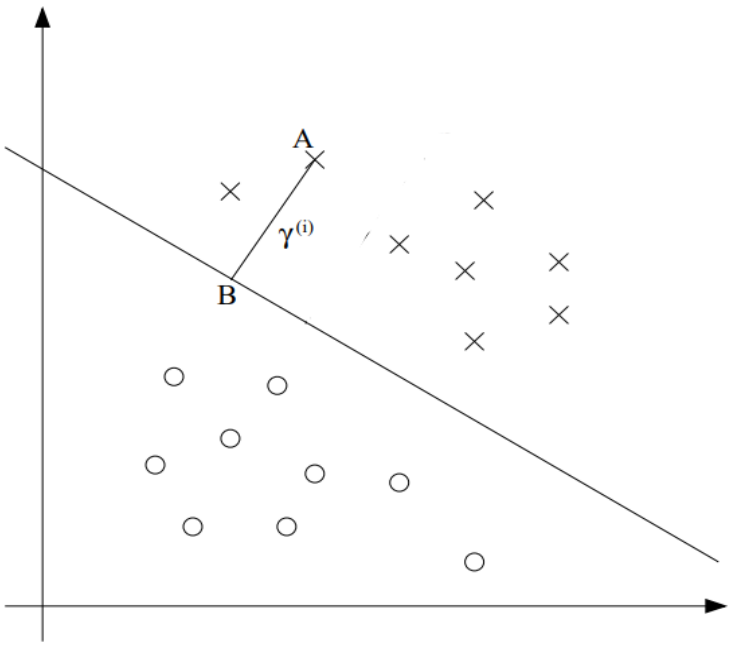
$$h_{w,b}(x) = g(w^T x + b)\ for\ some\ w \in R^n, b \in R, x \in R^n$$

Where

$$g(z) = \begin{Bmatrix} 1 & if\ z \geq 0 \\ -1 & otherwise \end{Bmatrix}$$

Geometric margin w.r.t $(X^{(i)}, y_i)$ –

It represents the distance of $(X^{(i)}, y_i)$ from decision surface. It is represented by $\gamma^{(i)}$



$$\gamma^{(i)} = \frac{y_i\left(w^T X^{(i)} + b\right)}{||w||}$$

Geometric margin (γ) w.r.t training set –

$$\gamma = \min_i \gamma^{(i)}$$

Optimal Margin Classifier –

Objective – Choose w,b to maximize γ.

i.e. $\max_{\gamma,w,b} \gamma$ s.t. $\frac{y_i\left(w^T X^{(i)}+b\right)}{||w||} \geq \gamma$ ... (1)

This problem can be reduced to this problem –

$\min_{w,b} \frac{1}{2}||w||^2$ s.t. $y_i\left(w^T X^{(i)} + b\right) \geq 1$ ... (2)

Let's suppose $w = \sum_{i=1}^{m} \alpha_i y_i X^{(i)}$ ... (3)

Substituting (3) in (2), we can further reduce the problem to –

$\max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j < X^{(i)}, X^{(j)} >\right)$

$< X^{(i)}, X^{(j)} >$ is the inner product of $X^{(i)}$ and $X^{(j)}$

s.t. $\alpha_i \geq 0 \ \forall\, i \in [1, m]$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$

This is a dual problem which can be solved with SMO Algorithm.

To make non-linear decision boundaries, we will introduce the Kernel trick in this algorithm.

We are going to replace $< X^{(i)}, X^{(j)} >$ with a Gaussian Kernel

$$K\left(X^{(i)}, X^{(j)}\right) = \exp\left(\frac{-\left|\left|X^{(i)} - X^{(j)}\right|\right|^2}{2\sigma^2}\right)$$

The visualization for function of SVMs can be found here.

Questions –

1. What type of algorithm is SVM?
   Ans. It is a supervised machine learning algorithm that can be used for both classification and regression problems. However, it is mostly used in classification problems.

2. What is the objective of using SVMs?
   Ans. The objective of the SVMs is to extend the data to a higher dimensional space and find a hyperplane separating the data in that space and then project that hyperplane to original space which results in a non-linear decision surface separating the datasets with different labels.

3. What is the advantage of using Kernels in SVM?
   Ans. The data can be transformed into a very high dimensional (even infinite dimensional) with very little increase in computational cost.
   In case of Gaussian Kernel, the feature's vector is being mapped to an infinite dimensional space.

4. What is the disadvantage of using SVM?
   Ans. SVM optimizes for worst-case margin, therefore it is very sensitive to outliers.

5. How to avoid outliers from affecting the optimal decision boundary?
   Ans. We can use the L_1 norm soft margin SVM.
   In this we have to solve
   $\min_{w,b,\zeta} \left(\frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\zeta_i\right)$ s.t. $y^{(i)}\left(w^T X^{(i)} + b\right) \geq 1 - \zeta_i \; \forall \; i \in [1, m]$
   Which reduces to $\max_{\alpha} \left(\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j < X^{(i)}, X^{(j)} >\right)$
   s.t. $\sum_{i=1}^{m}\alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C \; \forall \; i \in [1, m]$
   Now we can adjust the value of C and ζ_i s.t. the outliers doesn't affect the decision boundaries.