Support Vector Machines

Nirdosh Bhatnagar

1. Introduction

- Support vector machine is a supervised learning scheme, which was initially designed for classifying data points into two classes.
- Some of its applications are: text and Web page classification, and bioinformatics.
- The basic idea behind the use of support vector machine (SVM) is as follows.
  - A nonlinear mapping is initially used to transform the original training data set into a higher-dimensional space. In this higher-dimensional space, the SVM technique determines optimal decision boundary to classify data points.
  - This technique is due to the work of Vladimir Vapnik and his colleagues.
- In general, there are two types of SVMs: linear and nonlinear. Only linear SVMs are discussed in this note.
- In linear SVMs, the optimal decision boundaries are hyperplanes. In this case, it is sometimes possible to classify data points properly via a hyperplane. This is called the *linear separable case.*
- In some cases the hyperplane may classify the data points reasonably well, but for some other cases it may misclassify. As the hyperplane sometimes may not cleanly separate the data points. This later case is called the *linear nonseparable case.*

2. Hyperplanes in a Vector Space

A vector is represented as a row matrix.

– Hyperplane in a vector space, and normal vector to a hyperplane are defined. Let $z = (z_1, z_2, \ldots, z_t)$ be a vector in $\mathbb{R}^t$, where $t \in \mathbb{P} \backslash \{1\}$.

– Let $w = (w_1, w_2, \ldots, w_t) \in \mathbb{R}^t$, where $w \neq 0$. Also let $b \in \mathbb{R}$. The set of vectors

$$\widehat{Z} = \left\{ z \mid wz^T = \sum_{i=1}^{t} w_i z_i = b \right\}$$

is called the hyperplane in $\mathbb{R}^t$. This is a t-dimensional plane.

– A vector $\eta = (\eta_1, \eta_2, \ldots, \eta_t) \in \mathbb{R}^t$ is normal (perpendicular) to the hyperplane $\widehat{Z}$, if $\sum_{i=1}^{t} \eta_i z_i = 0$, $\forall\ z \in \widehat{Z}$.

– Let $wz^T = b$ be a hyperplane $\widehat{Z}$ in $\mathbb{R}^t$, where $z \in \widehat{Z}$. The vector w is normal to the hyperplane $\widehat{Z}$.

– Note that the hyperplanes $wz^T = b_1$ and $wz^T = b_2$, where $b_1, b_2 \in \mathbb{R}$, are parallel to each other.

– Let $\|\cdot\|$ be the Euclidean norm defined on the vector space $\mathbb{R}^t$. Thus for $z \in \mathbb{R}^t$, $\|z\|$ is the length of the vector z.

3. Shortest Distance Between a Point and a Hyperplane

– The shortest distance between a hyperplane and a point (not on the hyperplane) is determined in this section.

– Let $q \in \mathbb{R}^t$ be a point which does not lie on the hyperplane $\widehat{Z}$. Assume that $w \neq 0$. This implies $\|w\| \neq 0$. The shortest distance between the point q and the hyperplane $\widehat{Z}$ is $\left|wq^T - b\right| / \|w\|$. Thus, the perpendicular distance from the origin to the hyperplane is $|b| / \|w\|$. This result is next established.

– *Step 0*: Let $w, z \in \mathbb{R}^t$, where $w = (w_1, w_2, \ldots, w_t)$, and $z = (z_1, z_2, \ldots, z_t)$, and also let $b \in \mathbb{R}$. Further, $w \neq 0$. The set of vectors

$$\widehat{Z} = \left\{ z \mid wz^T = \sum_{i=1}^{t} w_i z_i = b \right\}$$

is called the hyperplane in $\mathbb{R}^t$.

Let $q \in \mathbb{R}^t$, where $q \notin \widehat{Z}$. It is established that the shortest distance between the point q and the hyperplane $\widehat{Z}$ is

$$r = \frac{\left|wq^T - b\right|}{\|w\|}$$

– *Step 1:* In this step, it is shown that, the vector w is normal (perpendicular) to the hyperplane $\widehat{Z}$. Let $z_1, z_2 \in \widehat{Z}$. Then

$$wz_1^T = b, \quad \text{and} \quad wz_2^T = b$$

This implies

$$w\left(z_1 - z_2\right)^T = 0$$

As the vector $(z_1 - z_2)$ is parallel to the hyperplane $\widehat{Z}$. The above relationship implies that the vector w is perpendicular to the hyperplane $\widehat{Z}$.

– *Step 2*: The length of the shortest distance r between the point $q \notin \widehat{Z}$ and the plane $\widehat{Z}$, is the absolute value of

$$(q - z) \cdot \frac{w}{\|w\|}$$

where '$\cdot$' is the dot product operator. Note that $z \cdot w = wz^T = b$. Therefore

$$\begin{aligned} r &= \frac{|q \cdot w - b|}{\|w\|} \\ &= \frac{\left|wq^T - b\right|}{\|w\|}, \quad \|w\| \neq 0 \end{aligned}$$

4. **Linear Separable Classes**

- Assume that the set of data-points belong to two different classes. The data points in the two different classes are said to be *linearly separable*, if they can be split into two different regions in the t-dimensional space via a hyperplane.
- Let $\mathcal{X} \subseteq \mathbb{R}^t$ be the set of data points, where each data point has $t \in \mathbb{P}$ coordinates.
- Let $\Omega = \{-1, +1\}$ be the set of labels for the two classes of data points.
- Classifier is a mapping $\Psi_{cl} : \mathcal{X} \to \Omega$.
  - The training data is the set $\mathcal{D}$ of n observations,

$$\mathcal{D} = \left\{(x_i, y_i) \mid x_i \in \mathcal{X} \subseteq \mathbb{R}^t, y_i \in \{-1, +1\}, 1 \leq i \leq n\right\}$$

    The members of this set are called training data points.
  - This is so, because the data points in this set will be used to determine the *best possible* separating hyperplane.
- If the data is 2-dimensional, that is $t = 2$, then the classifier divides the data into two classes separated via a straight line. In this case the data is said to be linearly separable; and the classes are specified by -1, and $+1$.
- In principle, it might be possible to draw an infinite number of lines which separate the data points into two non-overlapping parts. SVM attempts to find the *best* straight line which will give the smallest possible classification error on new data points.
- If the dimension t is greater than 2, then it is hoped that the classification scheme bifurcates the data points via hyperplanes.
  - SVM attempts to find the best hyperplane which separates the data points into two classes.
- The set of data points $\mathcal{D}$ are linearly separable, if there exists a nonzero weight vector $w \in \mathbb{R}^t$ and a scalar $\beta \in \mathbb{R}$ such that

$$\begin{aligned} wx_i^T + \beta &\geq +1, \quad \textit{if } y_i = +1 \\ wx_i^T + \beta &\leq -1, \quad \textit{if } y_i = -1 \end{aligned}$$

  or equivalently $y_i \left(wx_i^T + \beta\right) \geq 1$ for each $i = 1, 2, \ldots, n$.

5. **Analysis**

- The primary purpose of the SVM is to determine a classifier, which is a linear function of the form
$$f(x) = wx^T + \beta, \quad w, x \in \mathbb{R}^t$$
where, $w = (w_1, w_2, \ldots, w_t)$ is the nonzero weight vector. The parameter $\beta \in \mathbb{R}$ is the *bias*.
- Let $x_i \in \mathbb{R}^t$. If $f(x_i) \geq 1$, then x_i is classified as $y_i = +1$. On the other hand, if $f(x_i) \leq -1$, then x_i is classified as $y_i = -1$. The corresponding x_i's are called positive and negative points respectively.
- SVM determines the hyperplane $f(x) = \left(wx^T + \beta\right) = 0$, by using the training data set $\mathcal{D}$. SVM selects a hyperplane which maximizes the gap between the two (positive and negative) sets of training data points.
- Assume that the d_+ is the shortest distance from the separating hyperplane $f(x) = 0$ to the closest positive training data point $x_+ \in \mathcal{X}$.
- Similarly, assume that d_- is the shortest distance from the separating hyperplane $f(x) = 0$ to the closest negative training data point $x_- \in \mathcal{X}$.
- Thus, the net margin of the separating hyperplane is $(d_+ + d_-)$. The SVM technique maximizes this margin.
- Let H_+ and H_- be two hyperplanes, which are parallel to the hyperplane $f(x) = 0$, and pass through data points x_+ and x_- respectively. That is, no training data point falls in between these two hyperplanes. The hyperplanes H_+ and H_- are termed support planes.
- Thus the weight vector w and the bias β can be rescaled to obtain:
  - The hyperplane H_+ as $\left(wx^T + \beta\right) = +1$, where $x_+ \in H_+$.
  - The hyperplane H_- as $\left(wx^T + \beta\right) = -1$, where $x_- \in H_-$.

  Consequently for $x_i \in \mathcal{X}$, we have:
  - $\left(wx_i^T + \beta\right) \geq +1$, if $y_i = +1$.
  - $\left(wx_i^T + \beta\right) \leq -1$, if $y_i = -1$.
- The margin $(d_+ + d_-)$ is next determined.
  - The distance d_+ between x_+ and the hyperplane $f(x) = 0$ is
$$d_+ = \frac{\left|wx_+^T + \beta\right|}{\|w\|} = \frac{1}{\|w\|}$$
  - Similarly
$$d_- = \frac{\left|wx_-^T + \beta\right|}{\|w\|} = \frac{1}{\|w\|}$$

- Consequently, the margin
$$(d_+ + d_-) = \frac{2}{\|w\|}$$
- The SVM algorithm maximizes $1/\|w\|$. This is equivalent to minimizing $\|w\|$.
- This is identical to minimizing $\|w\|^2/2$.
- The SVM problem, can be described completely as follows.
  - *Input*: Training data points $\mathcal{D}$, where:
  - $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{X} \subseteq \mathbb{R}^t, y_i \in \{-1, +1\}, 1 \leq i \leq n\}$.
  - *Output*: Nonzero weight vector $w = (w_1, w_2, \ldots, w_t) \in \mathbb{R}^t$, and $\beta \in \mathbb{R}$.
  - *Objective*: Minimize $\|w\|^2/2$ by varying w and β.
  - *Subject to*: $y_i \left\{ wx_i^T + \beta \right\} \geq 1$, where $1 \leq i \leq n$.

□

- Notice that the constraint $y_i \left\{ wx_i^T + \beta \right\} \geq 1$, where $1 \leq i \leq n$; is equivalent to: $\left(wx_i^T + \beta \right) \geq +1$, if $y_i = +1$; and $\left(wx_i^T + \beta \right) \leq -1$, if $y_i = -1$.

6. Solution of the SVM Problem

- Lagrange multipliers are used to solve the SVM optimization problem.
- The objective function in this problem is strictly convex, and the constraints are linear. Consequently, the local minima is also global and unique, and the optimal hyperplane classifier of a SVM is unique.
  - The Lagrangian of this optimization problem is:

$$\mathcal{L}(w,\beta,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i \left\{ y_i \left(w x_i^T + \beta \right) - 1 \right\}$$

    where

$$\lambda_i \in \mathbb{R}, \quad 1 \le i \le n, \quad \text{and} \quad \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$$

  - The corresponding Karush-Kuhn-Tucker conditions of optimality are:

$$\begin{aligned} \frac{\partial \mathcal{L}(w,\beta,\lambda)}{\partial w_j} &= 0, \quad 1 \le j \le t \\ \frac{\partial \mathcal{L}(w,\beta,\lambda)}{\partial \beta} &= 0 \\ \lambda_i \left\{ y_i \left(w x_i^T + \beta \right) - 1 \right\} &= 0, \quad 1 \le i \le n \\ \lambda_i &\in \mathbb{R}_0^+, \quad 1 \le i \le n \end{aligned}$$

  - The conditions $\partial \mathcal{L}(w,\beta,\lambda)/\partial w_j = 0$, $1 \le j \le t$; and $\partial \mathcal{L}(w,\beta,\lambda)/\partial \beta = 0$ lead to

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i, \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i y_i = 0$$

  - The Lagrange multipliers (λ_i's), can be either equal to zero or greater than zero. If $\lambda_i > 0$, the corresponding vector x_i lies either on H_+ or H_- hyperplane.
  - Inequality constraints make the solution of the optimization problem nontrivial.
  - Lagrangian duality theory simplifies the problem to a certain extent. The corresponding dual problem can be stated as follows:
  - The dual optimization problem is:
    - *Input*: Set of training data points $\mathcal{D}$, where:
      $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{X} \subseteq \mathbb{R}^t, y_i \in \{-1, +1\}, 1 \le i \le n\}$.
    - *Output*: The Lagrangian vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$.
    - *Objective*: $\max_\lambda \left\{ \lambda e - \frac{1}{2} \lambda H \lambda^T \right\}$,

      where e is an all-1 vector of size n. That is, $e = (1, 1, \ldots, 1)^T$;
      and $H = [h_{ij}]$ is an $n \times n$ matrix, $h_{ij} = y_i y_j x_i x_j^T$, for $1 \le i, j \le n$.
    - *Subject to*: $\sum_{i=1}^{n} \lambda_i y_i = 0$, and $\lambda \ge 0$.

– The objective function in the corresponding dual problem has a quadratic form. Consequently, the corresponding optimization problem, is also referred to as a quadratic programming problem. Problems of this type are typically solved numerically.

– After determining the vector λ, the weight vector is determined from the equation: $w = \sum_{i=1}^{n} \lambda_i y_i x_i$. Assume that S is the set of indices of the support vectors. That is, $S = \{i \mid \lambda_i \in \mathbb{R}^+, 1 \leq i \leq n\}$. Thus, $w = \sum_{i \in S} \lambda_i y_i x_i$.

– The bias β, in principle, can be determined in principle, from anyone of the following equations, where $\lambda_i > 0$, and $\lambda_i \left\{ y_i \left(w x_i^T + \beta \right) - 1 \right\} = 0$, for $1 \leq i \leq n$. Nevertheless, β is determined by using all support vectors, and then taking their average value. These calculations are required for numerical stability.

– Further

$$f(x) \triangleq w x^T + \beta = \sum_{i \in S} \lambda_i y_i x_i x^T + \beta$$

– Note that, the decision boundary, or the maximum margin hyperplane is $f(x) = 0$.

– After determining the maximum margin hyperplane, a test data point $u \in \mathbb{R}^t \backslash \mathcal{X}$ is be classified by computing $f(u) = \left(w u^T + \beta \right)$. If $f(u) \geq 0$, implies u has classification which is positive, otherwise its classification is negative.

7. Algorithm for Linear Support Vector Machine: Separable Case

Algorithm Linear SVM: *Separable Case Algorithm*

Input: $\mathcal{D}$ is the set of training data points, where:

$$\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{X} \subseteq \mathbb{R}^t, y_i \in \{-1, +1\}, 1 \leq i \leq n\}.$$

Test data point is $u \in \mathbb{R}^t \backslash \mathcal{X}$.

Output: Compute the maximum margin hyperplane.

This decision boundary is optimal.

Classify the test data point $u \in \mathbb{R}^t \backslash \mathcal{X}$.

begin

Step 1: Compute $H = [h_{ij}]$, where $h_{ij} = y_i y_j x_i x_j^T$, for $1 \leq i, j \leq n$

Step 2: Determine λ so that $\max_\lambda \left\{ \lambda e - \frac{1}{2} \lambda H \lambda^T \right\}$;

subject to $\sum_{i=1}^n \lambda_i y_i = 0$, and $\lambda \geq 0$.

Note that, e is an all-1 vector of size n.

Step 3: Determine S, the set of indices of the support vectors.

$S = \{i \mid \lambda_i \in \mathbb{R}^+, 1 \leq i \leq n\}$

Step 4: Determine the weight vector $w = \sum_{i \in S} \lambda_i y_i x_i$

Step 5: Determine β from $\left\{ y_i \left(w x_i^T + \beta \right) - 1 \right\} = 0$, for each $i \in S$.

Determine the average value of all such computations of the β's.

Step 6: Let $f(x) \triangleq w x^T + \beta = \sum_{i \in S} \lambda_i y_i x_i x^T + \beta$

The maximum margin hyperplane is $f(x) = 0$

Step 7: Determine $f(u) = \left(w u^T + \beta \right)$. If $f(u) \geq 0$, then u is classifed

as positive, otherwise its classification is indeed negative.

end