Support Vector Machines

Falak Shah (201311024)

MTech, DA-IICT

December 16, 2014

Outline

Introduction to SVM

- Arguably the most successful machine learning tool for classification.
- Optimization packages available for solving the classification problem of SVM.(LibSVM, SVMTorch)
- Intended application: Binary or multiclass classification
- Linearly Separable data- selecting the best margin out of possible margins
- Bigger margin better as even in case of noisy data crossover probability is less.

Finding w with large margin

- Task is to find w and b for the separating hyperplane $w^T x + b = 0$.[1] [3]
- Here, b denotes bias, $x \in R^d$ is the input feature vector then $w = \{w_1, w_2, w_3 ... w_d\}$.
- For linearly separable points, the plane will not touch any points, i.e. $|w^T x + b| > 0$.
- Scaling of w and b by same amount result in the same plane. So, we select w and b such that $|w^T x_n + b| = 1$, x_n being the point closest to the hyperplane.
- Here, we'll use euclidean distance as the yardstick for measurement.

Distance Computation

- Distance between the plane $w^T x + b = 0$ and the nearest point x_n is the margin. Given that, $|w^T x_n + b| = 1$.
- Result 1:Vector w is $\perp$ to the plane.
- Proof: For x' and x'' on the plane $w^T x' + b = 0$ and $w^T x'' + b = 0$. So, $w^T(x' - x'') = 0$
- Projection of $x_n - x$ for any x on the plane onto the normal w.
- Distance d is given by d$=\frac{1}{||w||}|w^T(x_n - x)|=\frac{1}{||w||}|(w^T x_n + b) - (w^T x + b)| = \frac{1}{||w||}$

Optimization problem

- Maximize $\frac{1}{||w||}$ subject to $\min_{n=1,2..N} |w^T x_n + b| = 1$.
- Also, $|w^T x_n + b| = y_n(w^T x_n + b)$, since we only consider the correctly classified data.
- In the alternative representation, we can write the problem as Minimize $\frac{1}{2}(w^T w)$ subject to $y_n(w^T x_n + b) >= 1$ for $n = 1, 2...N$
- This statement is equivalent to the above one as minimum value for will only be achieved when $y_n(w^T x_n + b) = 1$ because till then w and b can still be proportionately scaled down.

Constrained Optimization problem

- Minimize $\frac{1}{2}(w^T w)$ subject $y_n(w^T x_n + b) - 1 >= 0$
- Solution to this will yield the separating hyperplane with the largest margin.
- Constrained optimization problem converted to unconstrained optimization by lagrangian.
- KKT conditions needed for solution of lagrangian under inequality constraint.

Lagrangian and KKT conditions

- KKT approach generalizes the method of Lagrange multipliers, which allows only equality constraints.
- Given a problem as $\min_x f(x)$ subject to $g(x) <= 0$.
- Define the lagrangian as
  $L(x, \lambda) = f(x) + \lambda g(x)$ Then,
  x^* a local minimum $\iff$ there exists a unique λ^* s.t.
  1. $\nabla_x L(x^*, \lambda^*) = 0$
  2. $\lambda^* >= 0$
  3. $\lambda^* g(x^*) = 0$
  4. $g(x^*) <= 0$

Lagrangian formulation of the problem

- Minimize $\frac{1}{2}(w^T w)$ subject to $y_n(w^T x_n + b) - 1 >= 0$ for $n = 1, 2 ... N$
- Minimize $L(w, b, \alpha) = \frac{1}{2}(w^T w) - \sum_{n=1}^{N} \alpha_n (y_n(w^T x_n + b) - 1)$ w.r.t w and b and maximize w.r.t. each $\alpha_n >= 0$
- $\nabla_w L = w - \sum_{n=1}^{N} \alpha_n y_n x_n = 0$
- $\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0$
- KKT condition 3: $\alpha_n (y_n(w^T x_n + b) - 1) = 0$
- Substituting these values in the original equation results in the dual representation of the problem.

Lagrangian formulation of the problem

- Substituting $w = \sum_{n=1}^{N} \alpha_n y_n x_n$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$ in the lagrangian
- $L(w, b, \alpha) = \frac{1}{2}(w^T w) - \sum_{n=1}^{N} \alpha_n (y_n (w^T x_n + b) - 1)$ we get,
- $L(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m$
- Maximise w.r.t. α and subject to $\alpha_n >= 0$ for n=1,2,...N and $\sum_{n=1}^{N} \alpha_n y_n = 0$

Quadratic Programming

$$\max_{\alpha} L(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m$$

Alternatively,

$$\min_{\alpha} L(\alpha) = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^T x_m - \sum_{n=1}^{N} \alpha_n$$

subject to $\alpha_n >= 0$ for n=1,2,...N and $\sum_{n=1}^{N} \alpha_n y_n = 0$

Quadratic programming formulation

$$\min_{\alpha} \frac{1}{2}\alpha^T \begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \dots & y_1 y_N x_1^T x_N \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \dots & y_2 y_N x_2^T x_N \\ .... & .... & ..... & .... \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \dots & y_2 y_N x_2^T x_N \end{bmatrix} \alpha + (-1)^T \alpha$$

subject to a linear constraint $y^T \alpha = 0$ and range of $0 <= \alpha <= \infty$ and

The size of the matrix depends on the size of the training dataset.

QP hands back α

- $\alpha = (\alpha_1, \alpha_2, ..., \alpha_N)$
- $w = \sum_{n=1}^{N} \alpha_n y_n x_n$
- α is a sparse vector. Sice we've the KKT condition $\alpha_n(y_n(w^T x_n + b) - 1) = 0$.
- Either $\alpha_n = 0$ for the interior points or $y_n(w^T x_n + b) = 1$ for the suppport vectors- the only ones where $\alpha_n > 0$. The x_n for which $\alpha_n > 0$ are called the support vectors as they only contribute to the solution. So now,
- $w = \sum\limits_{x_n \in S.V.} \alpha_n y_n x_n$
- Solve for any b using $y_n(w^T x_n + b) = 1$. This will give same b for any S.V. This is also verification that the task is correctly accomplished.

Linearly non-separable data

- Cover's theorem: The probability that classes are linearly separable increases when the features are nonlinearly mapped to a higher dimensional feature space. [2]
- Proof:For N samples in l-dimensional feature space, the number of dichotomies (linearly separable groupings) is [1]

$$O(N, l) = 2\sum_{i=0}^{l} \binom{N-1}{i}$$

- The total number of groupings is 2^N. Thus, the probability that the samples are linearly separable is the ratio

$$P_N^l = \frac{O(N, l)}{2^N}$$

Kernel Function

- Given 2 points x and x', we need z and z'. Where, $z = \phi(x)$
- Let $z^T z = K(x, x')$- the kernel function
- The trick is computing $K(x, x')$ without transforming x and x'.
- The function K can be arbitrarily chosen as long as the existence of $\phi(.)$ is guarnateed.
- A kernel K:$X \times X \to R$ is positive definite symmetric.

Kernel Function

- Mercer's condition: There exists a mapping $\phi(.)$ if and only if, for any g(x) such that

$$\int g(x)^2 dx$$

is finite then

$$K(x,y)g(x)g(y)dxdy \geq 0.$$

- Any kernel which can be expressed as $K(x,y) = \sum_{p=0}^{\infty} c_p (x.y)^p$, where the c_p are positive real coefficients and the series is convergent, satisfies the condition
- RBF kernel: $K(x,x') = exp(-\gamma||x-x'||^2)$
- Infinite dimensional z: with $\gamma = 1$ $K(x,x') = exp(-x^2)exp(-x'^2)\sum_{k=0}^{\infty} \frac{2^k (x)^k (x')^k}{k!}$

Linear Separability and Cover's Theorem

- SVM and several other pattern recognition techniques use the 'kernel' trick for projection in higher dimensional vector space
- Theorem by T.M Cover showing rise in linear separation probability in the higher dimensional space
- The probability that classes are linearly separable increases when the features are nonlinearly mapped to a higher dimensional feature space [3].
- Need- data not being linearly separable in lower dimensional space

Terminology

- Dichotomy: A dichotomy is any splitting of a whole into exactly two non-overlapping parts.
- Pattern: A vector in N-dimensional feature space
- Homogeneous linear thresholding function $f : E^N \to \{-1, 0, 1\}$ is defined in terms of a parameter w for every vector x in this space as:

$$f(x; w) = \begin{cases} 1, & w \cdot x > 0 \\ -1, & w \cdot x < 0 \\ 0, & w \cdot x = 0 \end{cases}$$

- A set of P vectors is in general position in N-space if every N element subset of vectors is linearly independent.

Linear Thresholding Device

- A d-input device that takes the sign of the sum of the products of pattern inputs with corresponding weights.

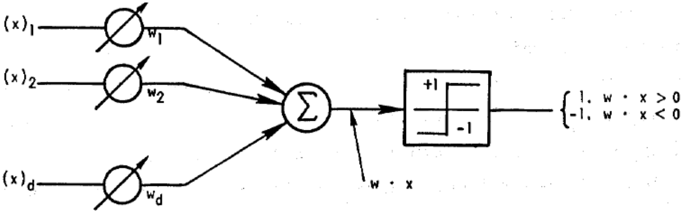

Figure 1: Linear Thresholding Device [1]

- Geometrically, such a device divides the space into two regions by a d-1 dimensional hyperplane through origin of the space.

[1] T M Cover, Geometrical and Statistical Properties of Linear Threshold Devices, PhD thesis, Stanford Electronics Laboratories, May 1964.

Homogeneous linear separability

- A dichotomy $\{X^+, X^-\}$ of X is **Linearly Separable** if and only if there exists a weight vector w in R^N and a scalar t such that
$$\begin{cases} w \cdot x > t & if x \in X^+ \\ w \cdot x < t & if x \in X^- \end{cases}$$
- The dichotomy is said to be homogeneously linearly separable if it is linearly separable with t=0
- Inhomogeneous is just a special case of homogeneous with one added variable
- A vector w satisfying the equations for homogeneous case is called the solution vector and the plane $\{x : w \cdot x = 0\}$ is the separating hyperplane for dichotomy $\{X^+, X^-\}$

Function Counting Theorem

- Function Counting Theorem [2]: There are C(P,N) homogeneously linearly separable dichotomies of P points in general position in Euclidean N-space where

$$C(P,N) = 2\sum_{k=0}^{N-1}\binom{P-1}{k}$$

Proof of Cover's Theorem

- Start with P points in general position. Assume that there are C(P,N) dichotomies possible on them, so how many dichotomies are possible if another point (in general position) is added C(P+1,N).
- Let $(b_1, ..., b_P)$ be a dichotomy realizable by a hyperplane over the set of P inputs, $b_i \in \{-1, +1\}\ \forall i = 1..P$, and there is a set of weights w so that for each of them $sign(w^T x_1), ..., sign(w^T x_P) = (b_1, ..., b_P)$.
- For every linearly realized dichotomy over P points there is at least one linearly realized dichotomy over P + 1 points

Proof of Cover's theorem

- There are some additional dichotomies possible which classify the newly added point P+1 into different classes

$$C(P+1, N) = C(P, N) + D$$

- D is the number of those dichotomies over P points that are realized by a hyperplane that passes through a certain fixed point x^{P+1}
- By forcing the hyperplane to pass through a certain fixed point, we are in fact moving the problem to one in N − 1 dimensions, instead of N

Recursive Formula for possible dichotomies

- So, D = C(P,N − 1), and the recursion formula is $C(P+1,N) = C(P,N) + C(P,N-1)$
- Further proof using mathematical induction for

$$C(P,N) = 2\sum_{k=0}^{N-1}\binom{P-1}{k}$$

$$C(P+1,N) = 2\sum_{k=0}^{N-1}\binom{P-1}{k} + 2\sum_{k=0}^{N-2}\binom{P-1}{k}$$

$$= 2\sum_{k=0}^{N-1}\binom{P-1}{k} + 2\sum_{k=0}^{N-1}\binom{P-1}{k-1}$$

$$= 2\sum_{k=0}^{N-1}\binom{P}{k}$$

Separability by arbitrary surfaces

- A family of surfaces each of which divide the given space into two regions and a colection of N points, each assigned either to X^+ or X^-
- On each pattern $x \in X$, a set of real valued measurement functions $\phi_1, \phi_2, ..., \phi_N$ comprises the vector of measurements $\phi(x) = (\phi_1(x), \phi_2(x)..., \phi_N(x)), \quad x \in X$
- A dichotomy is ϕ separable if there exists a vector w such that

$$\begin{cases} w \cdot \phi(x) > 0 & if x \in X^+ \\ w \cdot \phi(x) < 0 & if x \in X^- \end{cases}$$

Benifit: Nonlinear decision boundary

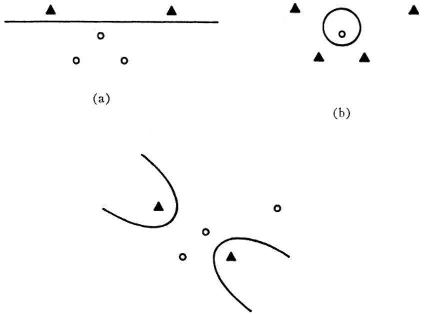

Figure 2: Examples of ϕ separable dichotomies of five points in two dimensions. (a) Linearly separable dichotomy. (b) Spherically separable dichotomy. (c) Quadrically separable dichotomy. [1]

[1] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. Electronic Computers, IEEE Transactions on, EC-14(3):326-334, June 1965

Christopher J. C. Burges.
A tutorial on support vector machines for pattern recognition.
Data Min. Knowl. Discov., 2(2):121–167, June 1998.

T.M. Cover.
Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.
Electronic Computers, IEEE Transactions on, EC-14(3):326–334, June 1965.

Vladimir N. Vapnik.
The Nature of Statistical Learning Theory.
Springer-Verlag New York, Inc., New York, NY, USA, 1995.