Foundation of Machine Learning

(CSE4032)

Lecture 07: Separating Hyperplane

Dr. Kundan Kumar
Associate Professor
Department of ECE

Faculty of Engineering (ITER)
S‘O’A Deemed to be University, Bhubaneswar, India-751030

Outline

1. Introduction
2. Support Vector Machine
3. Support Vector Classifier
4. Kernel Functions
5. Multiclass SVM
6. References

Separating Hyperplane

Introduction

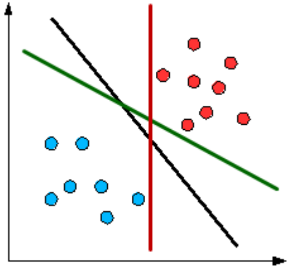
- Support vector machines are a class of statistical models first developed in the mid-1960s by Vladimir Vapnik.
- In later years, the model has evolved considerably into one of the most flexible and effective machine learning tools available.
- It is a supervised learning algorithm which can be used to solve both classification and regression problem, even though the current focus is on classification only.
- This algorithm looks for a linearly separable hyperplane, a decision boundary separating members of one class from the other.
- If such a hyperplane exists, the work is done! If such a hyperplane does not exist, SVM uses a nonlinear mapping to transform the training data into a higher dimension.

Introduction

- Then it searches for the linear optimal separating hyperplane.
- With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.
- The SVM algorithm finds this hyperplane using support vectors and margins.
- As a training algorithm, SVM may not be very fast compared to some other classification methods, but owing to its ability to model complex nonlinear boundaries, SVM has high accuracy.
- SVM is comparatively less prone to overfitting.
- SVM has successfully been applied to handwritten digit recognition, text classification, speaker identification etc.

Support Vector Machine

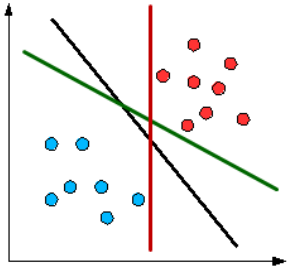
- Let us start with a simple two-class problem when data is clearly linearly separable as shown in the diagram below.
- Let the ith data point be represented by (X_i, y_i) where X_i represents the feature vector and y_i is the associated class label, taking two possible values $+1$ or -1.



- In the diagram, the balls having red color has class label $+1$ and the blue balls have class label -1, say.
- A straight line can be drawn to separate all the members belonging to class $+1$ from all the members belonging to the class -1. The two dimensional data are clearly linearly separable.

Support Vector Machine

- In fact, an infinite number of straight lines can be drawn to separate the blue balls from the red balls.
- The problem therefore is which among the infinite straight lines is optimal, in the sense that it is expected to have minimum classification error on a new observation.



- The straight line is based on the training sample and is expected to classify one or more test samples correctly.
- As an illustration, if we consider the black, red and green lines in the diagram above, is any one of them better than the other two? Or are all three of them equally well suited to classify? How is optimality defined here?

Support Vector Machine

- Intuitively it is clear that if a line passes too close to any of the points, that line will be more sensitive to small changes in one or more points.
- The green line is close to a red ball. The red line is close to a blue ball.
- If the red ball changes its position slightly, it may fall on the other side of the green line.
- Similarly, if the blue ball changes its position slightly, it may be misclassified.
- Both the green and red lines are more sensitive to small changes in the observations.
- The black line on the other hand is less sensitive and less susceptible to model variance.

Support Vector Machine

- In an n-dimensional space, a hyperplane is a flat subspace of dimension $n - 1$.
- For example, in two dimensions a straight line is a one-dimensional hyperplane.
- In three dimensions, a hyperplane is a flat two-dimensional subspace, i.e. a plane.
- Mathematically, in n dimensions, a separating hyperplane is a linear combination of all dimensions equated to 0; i.e.

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n = 0$$

- The scalar θ_0 is often referred to as a bias. If $\theta_0 = 0$, then the hyperplane goes through the origin.
- A hyperplane acts as a separator. The points lying on two different sides of the hyperplane will make up two different groups.

Support Vector Machine

- Basic idea of support vector machines is to find out the optimal hyperplane for linearly separable patterns.
- A natural choice of separating hyperplane is optimal margin hyperplane (also known as optimal separating hyperplane) which is farthest from the observations.
- The perpendicular distance from each observation to a given separating hyperplane is computed.
- The smallest of all those distances is a measure of how close the hyperplane is to the group of observations.
- This minimum distance is known as the margin.

Support Vector Machine

- The operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples, i.e. to find the maximum margin. This is known as the maximal margin classifier.
- A separating hyperplane in two dimension can be expressed as

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

- Hence, any point that lies above the hyperplane, satisfies

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$$

and any point that lies below the hyperplane, satisfies

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 < 0$$

Support Vector Machine

- The coefficients or weights $\theta 1$ and $\theta 2$ can be adjusted so that the boundaries of the margin can be written as

$$\begin{aligned} H_1 &: \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} \geq 1, \text{ for } y_i = +1 \\ H_2 &: \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} \leq -1, \text{ for } y_i = -1 \end{aligned}$$

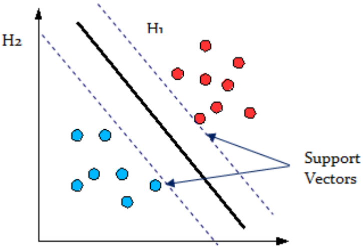
- This is to ascertain that any observation that falls on or above H_1 belongs to class $+1$ and any observation that falls on or below H_2, belongs to class -1.
- Alternatively, we may write

$$y_i \left(\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}\right) \leq \text{ for every observation}$$

- The boundaries of the margins, H_1 and H_2, are themselves hyperplanes too.

Support Vector Machine

- The training data that falls exactly on the boundaries of the margin are called the support vectors as they support the maximal margin hyperplane in the sense that if these points are shifted slightly, then the maximal margin hyperplane will also shift.



- Note that the maximal margin hyperplane depends directly only on these support vectors.
- If any of the other points change, the maximal margin hyperplane does not change, until the movement affects the boundary conditions or the support vectors.

Support Vector Machine

- The support vectors are the most difficult to classify and give the most information regarding classification.
- Since the support vectors lie on or closest to the decision boundary, they are the most essential or critical data points in the training set.
- For a general n-dimensional feature space, the defining equation becomes

$$y_i\left(\theta_0 + \theta_1 x_{2i} + \theta_2 x_{2i} + \ldots + \theta_n x_n i\right) \geq 1, \text{ for every observation}$$

- If the vector of the weights is denoted by Θ and $|\Theta|$ is the norm of this vector, then it is easy to see that the size of the maximal margin is $\frac{2}{|\Theta|}$.

Support Vector Machine

- Finding the maximal margin hyperplanes and support vectors is a problem of convex quadratic optimization.
- It is important to note that the complexity of SVM is characterized by the number of support vectors, rather than the dimension of the feature space.
- That is the reason SVM has a comparatively less tendency to overfit.
- If all data points other than the support vectors are removed from the training data set, and the training algorithm is repeated, the same separating hyperplane would be found.
- The number of support vectors provides an upper bound to the expected error rate of the SVM classifier, which happens to be independent of data dimensionality.
- An SVM with a small number of support vectors has good generalization, even when the data has high dimensionality.

Support Vector Classifier

- The maximal margin classifier is a very natural way to perform classification, is a separating hyperplane exists.
- However existence of such a hyperplane may not be guaranteed, or even if it exists, the data is noisy so that maximal margin classifier provides a poor solution.
- In such cases, the concept can be extended where a hyperplane exists which almost separates the classes, using what is known as a soft margin.
- The generalization of the maximal margin classifier to the non-separable case is known as the support vector classifier, where a small proportion of the training sample is allowed to cross the margins, or even the separating hyperplane.
- Rather than looking for the largest possible margin so that every observation is on the correct side of the margin.

Support Vector Classifier

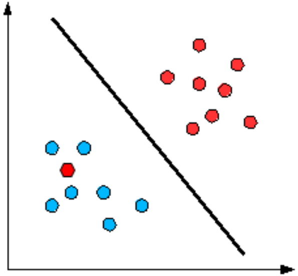
- Thereby making the margins very narrow or non-existent, some observations are allowed to be on the incorrect side of the margins.
- The margin is soft as a small number of observations violate the margin.
- The softness is controlled by slack variables which control the position of the observations relative to the margins and separating hyperplane.
- The support vector classifier maximizes a soft margin.
- The optimization problem can be modified as

$$y_i\left(\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_n x_{ni}\right) \geq 1 - \epsilon_i \text{ for every observation}$$

$$\text{Where } \epsilon_i \geq 0 \text{ and } \sum_{i=1}^{n} \epsilon_i \leq C$$

Support Vector Classifier

- The ϵ_i is the slack corresponding to ith observation and C is a regularization parameter set by user. Larger value of C leads to larger penalty for errors.
- However there will be situations when a linear boundary simply does not work.
- SVM is quite intuitive when the data is linearly separable. However, when they are not, as shown in the diagram below, SVM can be extended to perform well.



Support Vector Classifier

- There are two main steps for nonlinear generalization of SVM.
  - □ The first step involves transformation of the original training (input) data into a higher dimensional data using a nonlinear mapping.
  - □ Once the data is transformed into the new higher dimension, the second step involves finding a linear separating hyperplane in the new space.
- The maximal marginal hyperplane found in the new space corresponds to a nonlinear separating hypersurface in the original space.

Example: Feature Expansion

- Suppose the original feature space includes two variables X_1 and X_2. Using polynomial transformation the space is expanded to $(X_1, X_2, X_1^2, X_2^2, X_1X_2)$. Then the hyperplane would be of the form

$$\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_1^2 + \theta_4 X_2^2 + \theta_5 X_1 X_2 = 0$$

- This will lead to nonlinear decision boundaries in the original feature space. If upto second degree terms are considered, 2 features are expanded to 5 . If upto third degree terms are considered the same to features can be expanded to 9 features. The support vector classifier in the expanded space solves the problems in the lower dimension space.

Kernel Functions

- Handling nonlinear transformation of input data into higher dimension may not be easy. There may be many options available to begin with and the procedures may be computationally heavy also. To avoid some of those problems, the concept of Kernel functions is introduced.
- It so happens that in solving the quadratic optimization problem of the linear SVM, the training data points contribute through inner products of nonlinear transformations. The inner product of two n-dimensional vectors is defined as

$$\sum_{j=1}^{n} x_{1j} x_{2j}$$

Where $X_1 = (x_{11}, x_{12}, \cdots x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots x_{2n})$.

Kernel Functions

- Kernel function is a generalization of the inner product of nonlinear transformation and is denoted by $\mathrm{K(X1, X2)}$. Anywhere such an inner product appears, it is replaced by the kernel function.
- In this way, all calculations are made in the original input space, which is lower dimensionality. Some of the common kernels are polynomial kernel, sigmoid kernel and Gaussian radial basis function. Each of these will result in a different nonlinear classifier in the original input space.
- There are no golden rule to determine which kernel will provide the most accurate result in a given situation. In practice, accuracy of SVM does not depend on the choice of the kernel.

Multiclass SVM

- The SVM as defined so far works for binary classification. What happens if the number of classes is more than two?
  - □ One-versus-All: If the number of classes is $\mathrm{K} > 2$ then K different 2 -class SVM classifiers are fitted where one class is compared with the rest of the classes combined. A new observation is classified according to where the classifier value is the largest.
  - □ One-versus-One: All $\begin{pmatrix} K \\ 2 \end{pmatrix}$ pairwise classifiers are fitted and a test observation is classified in the class which wins in the majority of the cases. The latter method is preferable but if K is too large, the former is to be used.

References

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Hastie, Tibshirani, and Friedman, Springer.
- In Introduction to Statistical Learning with Application in R, Second Edition, James, Witten, Hastie, and Tibshirani, Springer.

Thank you!