Application of SVM on MNIST dataset

Arkonil Dhar Saumyadip Bhowmick Shreya Pramanik
Shubha Sankar Banerjee Souvik Bhattacharya

Summer Project 2021
Indian Institute of Technology Kanpur

July 22, 2021

Data Description The training data consists of 50000 observations on the label (response y) which has $K = 10$ classes viz. $\{0, 1, 2, \ldots, 9\}$ and 784 features (the independent variables $X = (X_1, \ldots, X_{784})$).

In some typical learning algorithm such as Logistic Regression, LDA or QDA we try to directly model the conditional probability of y given x i.e. $p(y|x)$. But in case of non-probabilistic approaches like Support Vector Machines, we are interested to find a hyperplane in an n-dimensional space(where, n is the number of features) which classifies the data points into two classes.To separate the data points into two classes, there are many possible choice of hyperplanes. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes.

Maximal Margin Classifier

The support vector machine is a generalization of a simple and intuitive classifier called the maximal margin classifier.suppose that we have a $n \times p$ data matrix X that consists of n training observations in p-dimensional space,

$$x_1 = \begin{pmatrix} x_{11} \\ . \\ . \\ . \\ x_{1p} \end{pmatrix}, ..., x_n = \begin{pmatrix} x_{n1} \\ . \\ . \\ . \\ x_{np} \end{pmatrix} \tag{1}$$

and that these observations fall into two classes that is, $y_1, ..., y_n \epsilon \{-1, 1\}$ where -1 represents one class and 1 the other class. We also have a test observation, a p-vector of observed features $x^* = (x_1^* ... x_p^*)^T$ Our goal is to develop a classifier based on the training data that will correctly classify the test observation using its feature measurements.**Separating Hyperplanes** are boundaries that help to classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. In general, if our data can be perfectly separated using a hyperplane according to their class labels, then there will in fact exist an infinite number of such hyperplanes, given by the equation, $w^T x + b = 0$.**Margin** is the distance between nearest data point (either class) and the hyper-plane.The maximal margin hyperplane is the separating hyperplane for which the margin is largest—that is, it is the hyperplane that has the farthest minimum distance to the training observations. We can then classify a test observation based on which side of the maximal margin hyperplane it lies. This is known as the maximal margin classifier. **Support vectors** are data points that are closest to the hyperplane, i.e. thhey are margin away from the Separating Hyperplane. Our objective is to find the value of w and b, such that the hyperplane is separating enough data points to corresponding classes correctly.

Decision Rule for Binary Classification

Let us denote, {1,-1} to denote the class labels. Let, our classifier be,

$$h_{w,b}(x) = g(w^T x + b) \tag{2}$$

where,

$$g(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ -1 & , otherwise \end{cases} \tag{3}$$

From the definition of g above, our classsifier will direectly compute the class label of the given data point.

Decision Rule for Multi-Class Classification

In general we may have more than one classes, there are several approaches to handle such multi-class classification. We have used **One vs One** approach here.In this method,for a single data point we classify the point into all possible binary classes, and take the majority occurance of the predicted class.

Support Vector Classifiers

The maximal margin classifier is a very natural way to perform classification, if a separating hyperplane exists. However, in many cases no separating hyperplane exists, and so there is no maximal margin classifier, i.e. it is impossible to find the value of w and b, for which all points are classified to correct classes.In this case, we might be willing to consider a classifier based on a hyperplane that does not perfectly separate the two classes, in the interest of greater robustness to individual observations, and better classification of most of the training observations. The support vector classifier, sometimes called a soft margin classifier, support vector classifier soft margin classifier does exactly this. Rather than seeking the largest possible margin so that every observation is not only on the correct side of the hyperplane but also on the correct side of the margin, we instead allow some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane. The tuning parameter C is the budget for the amount that the margin can be violated by the n observations.When C is small, we seek narrow margins that are rarely violated; this amounts to a classifier that is highly fit to the data, which may have low bias but high variance. On the other hand, when C is larger, the margin is wider and we allow more violations to it; this amounts to fitting the data less hard and obtaining a classifier that is potentially more biased but may have lower variance.

Support Vector Machines

The support vector classifier is a natural approach for classification in the two-class setting, if the boundary between the two classes is linear. However, in practice we are sometimes faced with non-linear class boundaries.The support vector machine (SVM) is an extension of the support vector classifier that results from enlarging the feature space in a specific way, using kernels.