Neural Networks

Assignment 7

Ravikiran Bhat
Rubanraj Ravichandran
Ramesh Kumar

November 22, 2017

1 Support Vector Machines(SVMs)

1.1 Introduction about SVMs

- Pioneered by Vapnik
- Like multi-layer perceptrons and radial-basis function network, SVMs are also used for classification and nonlinear regression
- Main idea of SVMs is to draw a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized
- The machine achieves this desirable property by following method of structural risk minimization
- This principle is based on fact that error rate of learning machine on test data is bounded by sum of training-error rate and a term that depends on Vapnik-Chervonenkis(VC) dimension
- SVM learning algorithm is the inner-product kernel b/w a "support vector" x_i and the vector x drawn from the input space.
- Support vectors consist of a small subset of training data extracted by the algorithm
- Different ways to compute inner-product of kernel to characterize non-linear decision surface:
  - Polynomial learning machines
  - Radial-basis function networks
  - Two-layer perceptrons(i.e with a single hidden layer)

1.2 Optimal Hyperplane for Linearly Separable Patterns

- If patterns are linearly separable then hyperplane equation is:

  $$w_x^T + b = 0$$

  where x is an input vector, w is an adjustable weight vector and b is a bias

- Goal of SVMs is to determine particular hyperplane for which margin of separation is maximized, that is considered as optimal hyperplane
- support vectors are those data points that lie closest to decision surface and are difficult to classify
- Margin of separation between two classes is given as

  $$p = \frac{2}{\|w_o\|}$$

  Above optimal hyperplane is unique in sense that optimum weight vector w_o provides the maximum possible separation between positive and negative examples

1.2.1 Quadratic Optimization for Finding Optimal Hyperplane

- Goal is to develop computationally efficient procedure for using the training samples to find optimal hyperplane
- Well-studied class of optimization algorithms to maximize a quadratic function of some real-valued variables to linear constraints
- Duality theorem:
  - If primal problem has an optimal solution, dual problem also has an optimal solution and corresponding optimal values are equal
  - In order for w_o to be an optimal primal solution and α_o to be an optimal dual solution, it is necessary and sufficient that w_o is feasible for primal problem

1.3 Optimal Hyperplane for Non-separable Patterns

- Margin of separation between classes is soft if a data point (x_i, d_i) violates following condition

  $$d_i(w^T x_i + b) \geq +1$$

  i = 1, 2, .. N
- For formal treatment of non-separable data points, we introduce ξ_i called slack variables, so above equation becomes

  $$d_i(w^T x_i + b) \geq +1 - \xi_i$$

i = 1, 2, .. N

1.4 Build a SVM for Pattern Recognition

- In order to find optimal hyperplane for non-separable patterns, Cover's theorem states that:
    - A multi-dimensional space may be transformed into a new feature space where the patterns are linearly separable with high probability, provided two conditions are satisfied:
        - ∗ Transformation is nonlinear
        - ∗ Dimensionality of feature space is high enough

1.4.1 Examples of Support Vector Machine

- Requirement of kernel K(x,x_i) is to satisfy Mercer's theorem
- Inner-product kernels for three common types of support vector machines are:
    - Polynomial learning machine given as:
    $$(x^T x_i + 1)^T$$
    - Radial-basis function given as
    $$exp(\frac{-1}{2\sigma^2}\|x - x_i\|^2)$$
    - Two-layer perceptron given as
    $$tanh(\beta_0 x^T x_i + \beta_1)$$
- SVM differs from MLP in a fundamental way.
- In conventional approach, model complexity is controlled by keeping number of features(hidden neurons) small.
- On the other hand, SVM offers a solution to the design of a learning machine by controlling model complexity independently of dimensionality

1.5 ϵ-Insensitive Loss Function

When performing a non-linear regression task, support vector learning algorithm minimizes an ϵ-insensitive loss function that is an extension of the mean absolute error criterion of minimax theory, this makes algorithm robust.