Challenge 5 – Support Vector Machines

Siddharth Ajit
Texas A&M University

I. Introduction

Support vector machines are supervised machine learning model capable for both classification and regression. In classification, the algorithm assigns each data point to a class in a non-probabilistic fashion [1]. The main objective is to divide the data cloud into binary classes with the help of a separating hyperplane. Further, the separation or “margin” between the two classes must be maximized to obtain a more generalizable model. However, in some cases, the data points are not linearly separable. For such scenarios, a hinge loss function is adopted with a pre-selected coefficient to penalize misclassifications. Minimization of the loss function yields the best possible separating hyperplane. “kernel trick” is employed to map the input vector to high dimensional feature spaces in the instance of nonlinear classification boundaries [1]

$$J(\mathbf{w}, b) = C\sum_{i=1}^{m} max\left(0, 1 - y^{(i)}(\mathbf{w}^t \cdot \mathbf{x}^{(i)} + b)\right) \quad + \quad \frac{1}{2}\mathbf{w}^t \cdot \mathbf{w}$$

- w is the model's feature weights and b is bias parameter.
- m is the number of training instances.
- C is the regularization hyperparameter [2].

$X^{(i)}$ is the i^{th} feature vector and $Y^{(i)}$ is the target class.

A. Dataset description

The training and test dataset has 200 data instances with two features

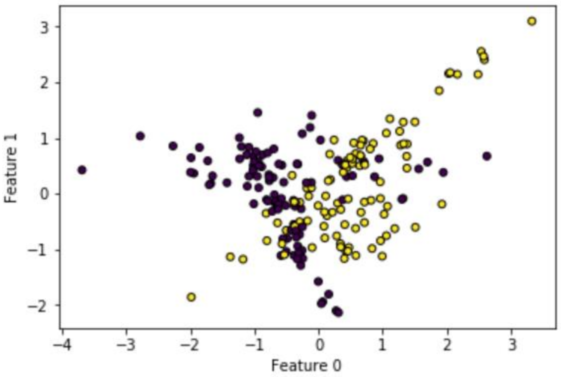

Figure 1. Feature space

From the training dataset, a linear classifier doesn’t seem suitable. Other nonlinear classifiers may provide better classification accuracy

B. Methodology

To identify the best kernel and the corresponding hyperparameters, cross validation misclassification rate was used. The kernels used are linear, polynomial (degree = 3) and Radial basis function. Further, hyperparameters C (Regularization parameter) and Gamma are optimized using Gridsearch CV to obtain the optimal hyper-parameters of the corresponding kernels. Both Gamma (γ) and C parameter was varied from 0.001 to 10 in 12 spaced intervals.

A smaller value of C will result in larger margin and a more generalizable model at the cost of training accuracy [3 (Wikipedia.org, n.d.)]. Similarly, gamma parameter influences the effect of single data point on the decision point. If the gamma value is less, the points farther from the decision boundary have an impact on the shape of the shape on the decision boundary leading to a model with higher variance.

RBF Kernel function

$$K(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$$

RBF kernel function on two sample points x, x'. The term $(1/2\sigma^2)$ is equivalent to γ parameter [4].

The following cross validation errors are obtained for the best models in respective kernel functions.

Kernel function	C	Gamma (γ)	Cross validation accuracy
Linear	5	0.001	0.805
Polynomial (degree =3)	0.1	1	0.77
RBF	10	0.75	0.875

Table1. Cross validation accuracy

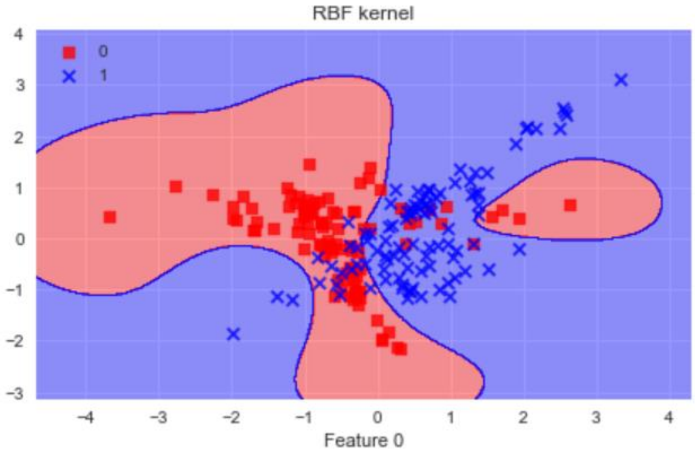

Figure 2. Training Decision boundary RBF

Figure 2 shows the decision boundary of RBF kernel (best model) with hyperparameters C = 10 and $\gamma = 0.75$. As expected, a nonlinear boundary performs better than a linear hyperplane in terms of misclassification rate. Further, a model trained with these parameters of RBF was used to predict the test data set.

C. References

1. Retrieved from Wikipedia.org: https://en.wikipedia.org/wiki/Support_vector_machine
2. Retrieved fromStackexchange.com: https://stats.stackexchange.com/questions/215524/is-gradient-descent-possible-for-kernelized-svms-if-so-why-do-people-use-quadr

3. Retrieved from scikit-learn.org https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html
4. Retrieved from wikipedia.org https://en.wikipedia.org/wiki/Radial_basis_function_kernel