#15EC10044 #Rajeswari Mahapatra #Assignment 5

Assignment-5

Problem Statement:

- Spam email classification using Support Vector Machine: In this assignment you will use a SVM to classify emails into spam or non-spam categories.
- And report the classification accuracy for various SVM parameters and kernel functions.
- You have to submit the report file in pdf format. No programs need to be submitted.

Data Set Description:

- An email is represented by various features like frequency of occurrences of certain keywords, length of capitalized words etc. A data set containing about 4601 instances are available in this link (data folder): https://archive.ics.uci.edu/ml/datasets/Spambase
- The data format is also described in the above link. You have to randomly pick 70% of the data set as training and the remaining as test data.

Methodology:

- SVM package used: sklearn.svm
- Default parameters: *class* sklearn.svm.SVC(*C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='auto_deprecated'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*, *verbose=False*, *max_iter=-1*, *decision_function_shape='ovr'*, *random_state=None*)

Given training vectors $x_i \in \mathbb{R}^p$, i=1,..., n, in two classes, and a vector $y \in \{1, -1\}^n$, SVC solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$
$$\zeta_i \geq 0, i = 1, \ldots, n$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$
$$\text{subject to } y^T \alpha = 0$$
$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$

where e is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The decision function is:

$$\text{sgn}(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho)$$

Experimental Results:

- C is iterated from 0 to 100 and the best value is noted.

Kernel Type	Training Accuracy	Test Accuracy	Value of C
Linear	0.8618012422360248	0.8587979724837075	11
Quadratic	0.9031055900621118	0.8906589427950761	4
RBF	0.9903726708074534	0.8435916002896452	6