# 

# EE5904 Neural Network

# SVM for Classification of Spam Email Messages

**LI XINRUI**

**A0279734M**

**e1143547@u.nus.edu**

## **Section 1: Data Pre-processing**

Data was loaded and separated into input data and labels, both for training set and test set.

The training set contains 2000 input data with 57 features each, while test set contains 1563 input data with 57 dimensions. Labels are vectors correspond to the input data, elements are either 1 or -1. the training set is a matrix with 57 rows and 2000 columns, test set has 57 rows and 1536 columns. Both training set and test set were first checked if there are null values. Then the standard normalization was performed on both training set and test set, subtracting the mean of the feature from each data point in training set and then dividing by the standard deviation of the feature, transforming the value to have a mean of zero and a standard deviation of one.

## **Section 2: Discriminant Function Computation**

Discriminant function is defined as below formula:

where and are the optimal value of the algorithm.

The calculation for discriminant function is different between types of kernel and margin. To obtain this function is to get optimal weights and bias first. The algorithm is mainly solving a quadratic programming problem with equality constraints, finding support vectors and calculate and . the quadratic problem is in the following form:

subject to for

To satisfy this requirement, the SVM problem should be transformed into the the minimization problem:

for

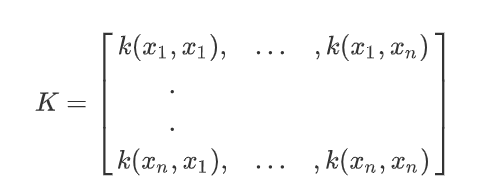
It it important to generate the Hessian matrix H and find the coefficient of the linear term . For linear kernel with hard margin, Hessian matrix is a symmetric matrix with element in the following form:

and the coefficient for linear term is a column vector with N rows:

After solving the quadratic problem, optimal gets and then

where indicates the support vector with .

For polynomial kernel, there is a trick when computing Hessian matrix. the gram matrix is defined for kernel as



The optimal weights is hard to compute, so it does not be computed directly, and is determined by solving the equation:

, where represents the support vector. In practice, the optimal bias can be computed as the average value from all support vector input.

The discriminant function is defined as:

## **Section 3: Modeling Result and Admissibility of Kernel**

There are three main types of SVM model, linear kernel with hard margin, polynomial kernel with hard margin and polynomial kernel with soft margin.

The models training and test accuracy are computed and listed in the following table:

| **SVM type** | **Training accuracy** | | | | **Test accuracy** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hard margin &linear kernel** | 0.7875 | | | | 0.7728 | | | |
| **Hard margin & poly kernel** | **P=2** | **3** | **4** | **5** | **2** | **3** | 4 | **5** |
| 0.9995 | 0.857 | 0.843 | 0.813 | 0.9271 | 0.8333 | 0.8151 | 0.7897 |
| **Soft margin & poly kernel** | **C=0.1** | **0.6** | **1.1** | **2.1** | **0.1** | **0.6** | **1.1** | **2.1** |
| **P=1** | 0.9370 | 0.9395 | 0.9395 | 0.939 | 0.9290 | 0.9290 | 0.9303 | 0.9297 |
| **2** | 0.9730 | 0.9795 | 0.9780 | 0.9835 | 0.9284 | 0.9310 | 0.9303 | 0.9310 |
| **3** | 0.8880 | 0.8840 | 0.8830 | 0.9100 | 0.8424 | 0.8444 | 0.8444 | 0.8678 |
| **4** | 0.6065 | 0.6065 | 0.6155 | 0.6215 | 0.6003 | 0.6016 | 0.6120 | 0.6159 |
| **5** | 0.6485 | 0.6270 | 0.6685 | 0.6555 | 0.6328 | 0.6204 | 0.6589 | 0.6484 |

It could be told from this table that for linear kernel with hard margin, both training and test accuracy is good but not that perfect, with over 70% accuracy. The model is proper fitting because the values of accuracy between training and test are almost the same. The linear kernel is admissible.

For polynomial kernel with different degrees, the accuracy results indicates different admissibility of kernels. For degree of 1, 2, 3, the model perform well with both soft and hard margin, with similar values between training set and test set and high accuracy, indicating the admissible kernels. For degree of 4 and 5, models perform well with hard margin, but the accuracy is obviously smaller, around 0.6, with soft margin. It indicates that polynomial kernel with excessively high degrees may result in worse performance for SVM model. Since the value of accuracy between training set and test set are similar, it does not indicate an overfitting problem. The possible reason for lower accuracy is that when data are mapped to higher dimensional feature space by high degree kernel, data points may become more sparse, making it difficult to find a hyperplane that separates the data well. It indicates that high degrees of polynomial kernel with soft margin is not admissible.

## **Section 4 : Existence of Optimal Hyperplane**

Determining the existence of a hyperplane essentially involves assessing whether the given dataset is linearly separable. If the dataset is linearly separable, there exists a hyperplane that can completely separate the positive and negative instances.

The hyperplane can be represented by the equation of discriminant function:

.

If the dataset is linearly separable, then there exists a hyperplane satisfying the following conditions:

For all, if then ;

For all, if then ;

Practically speaking, this problem can be transformed into the optimization problem. If the optimal exists, that is the minimum value of the quadratic problem can be found, the hyperplane exists.

In the project, the minimum value of the quadratic function cannot be found only in the case of linear kernel with hard margin, indicating that SVM of linear kernel with hard margin have no hyperplane. Other types of SVM all have hyperplanes.

## **Section 5: Design Decision**

For model evaluation, the extra data set 'eval.mat' is used, which contains evaluation data and labels. The polynomial kernel with degree of 2 and soft margin with C = 0.6 are chosen as the parameters of SVM. The reason for the design is that the hyperplane exists and the training and test accuracy are both relatively high among all types, indicating a good model fitting result and no overfitting problem.