Name – Sayem Lincoln

PID – A54207835

CSE 404 Homework 9

Note: (1) LFD refers to the textbook “Learning from Data”.

1. (25 points) Exercise 8.13 (e-Chap:8-31) in LFD.

KKT complementary slackness gives that if , then is on the boundary of the optimal fat-hyperplane and . Show that the reverse is not true. Namely, it is possible that and yet is on the boundary satisfying

[Hint: Consider a toy data set with two positive examples at (0,0) and (1, 0), and one negative example at (0, 1).]

[Note: You don't have to read through the KKT complementary slackness to work out this question.]

*Answer:*

Assuming we have , we need to show that it is possible for . We can do so by using the toy data set described in the hint. In that problem, the optimal hyperplane is where . For this case, the class labels would have to be and . It’s easy to see that the point is not a support vector. If we remove it, the optimal hyperplane wouldn’t change. Thus, for that data point. All that’s left to show is that for all three data points. If this is the case, we have all data points lying on the boundary, but one point exists with .

For , we have .

For , we have .

Lastly, for , we have

Thus, there exists a counterexample, so the statement cannot be true.

1. (25 points) Exercise 8.15 (e-Chap:8-38) in LFD.

Consider two finite-dimensional feature transforms and and their corresponding kernels and .

1. Define . Express the corresponding kernel of in terms of and .
2. Consider the matrix and let be the vector representation of the matrix (say, by concatenating all the rows). Express the corresponding kernel of in terms of and .
3. Hence, show that if and are kernels, then so are and .

The results above can be used to construct the general polynomial kernels and (when extended to the infinite-dimensional transforms) to construct the general Gaussian-RBF kernels.

*Answer:*

1. Firstly, it’s important to note that is also finite-dimensional since both and are finite-dimensional. Next, it’s a simple matter of linear algebra to derive the kernel.
2. First, we need to find exactly what is in this situation (i.e. what does it mean to concatenate rows). When we concatenate the rows, we simply put row 2 behind row 1, then row 3 behind row 2, and so on. We take the transpose of this vector so that it’s a column vector. Now, let and to simplify the notation a bit. Also, assume has dimensions, and has dimensions. We have:
3. Assume otherwise. This means that it’s not possible to find kernels and such that and are both separately kernels. However, we just did that in parts (a) and (b); thus, we have a contradiction. This means the original statement must be true.
4. (25 points) Experiment: A data set (data.mat) is given and you are asked to apply support vector machines on this data set. You can use the software package LIBSVM1 for this purpose. Note that LIBSVM has a Python interface2, so you can call the SVM functions in Python. You can use the first 150 samples for training and the rest for testing. In this task, you should try different values for the parameter , different kernel functions, etc. and write a short report summarizing your observations. Also, please report how the number of support vectors changes as the value for increases (while all other parameters remain the same).

[Note: You can load the mat dataset into Python using the function loadmat in Scipy.io.]

1http://www.csie.ntu.edu.tw/˜cjlin/libsvm/

2https://github.com/cjlin1/libsvm/tree/master/python

*Answer:*

The accuracy result and total support vector number of each experiment model using a dataset (datamat.mat) with 150 samples for training and the rest for testing.

|  |  |  |  |
| --- | --- | --- | --- |
| Cost Parameter | Kernel | Support Vectors | Test Accuracy |
| 0.01 | Default (Radial Basis Function) | 140 | 58.33 % |
| 0.1 | Default (Radial Basis Function) | 132 | 84.17 % |
| 1 | Default (Radial Basis Function) | 83 | 84.17 % |
| 2 | Default (Radial Basis Function) | 76 | 82.50 % |
| 3 | Default (Radial Basis Function) | 78 | 83.33 % |
| 5 | Default (Radial Basis Function) | 74 | 80 % |
|  |  |  |  |
| 1 | Linear | 58 | 85 % |
| 1 | Polynomial | 118 | 82.50 % |
| 1 | Default (Radial Basis Function) | 83 | 84.17 % |
| 1 | Sigmoid | 79 | 84.17 % |

***Varying******Cost***: As the cost increases, we care more about violating the margin. In this case, we see that the number of support vectors decreases as the cost increases (but not for . On the other hand, the test accuracies increase as the cost increases but the test accuracies decrease when . This intuitively makes sense because, as the cost increases, the soft-margin SVM increasingly resembles the hard-margin SVM. The hard-margin SVM should certainly be more accurate with linearly separable data because we emphasize or more highly value correctly classifying data points.

***Varying******Kernels***: As for varying the kernels, the three of them achieve roughly equivalent test accuracies. However, the polynomial kernel gets bad result and it requires 2x as many support vectors, which is computationally more expensive.

The accuracy result and total support vector number of each experiment model using a USPS dataset (USPS.mat) with 2500 samples for training and the rest for testing.

|  |  |  |  |
| --- | --- | --- | --- |
| Cost Parameter | Kernel | Support Vectors | Test Accuracy |
| 0.01 | Default (Radial Basis Function) | 2500 | 0 % |
| 0.1 | Default (Radial Basis Function) | 1775 | 19.8 % |
| 1 | Default (Radial Basis Function) | 1058 | 34.2 % |
| 2 | Default (Radial Basis Function) | 984 | 34.6 % |
| 3 | Default (Radial Basis Function) | 971 | 34.8 % |
| 5 | Default (Radial Basis Function) | 976 | 35% |
|  |  |  |  |
| 1 | Linear | 718 | 33.6 % |
| 1 | Polynomial | 1051 | 29.4 % |
| 1 | Default (Radial Basis Function) | 1058 | 34.2 % |
| 1 | Sigmoid | 1111 | 32.6 % |

***Varying******Cost***: As the cost increases, we care more about violating the margin. In this case, we see that the number of support vectors decreases as the cost increases (but not for . On the other hand, the test accuracies increase as the cost increases. This intuitively makes sense because, as the cost increases, the soft-margin SVM increasingly resembles the hard-margin SVM. The hard-margin SVM should certainly be more accurate with linearly separable data because we emphasize or more highly value correctly classifying data points.

***Varying******Kernels***: As for varying the kernels, the model achieve various test accuracies. The linear kernel gets the best result with smallest number of support vectors. However, the polynomial kernel gets bad result and sigmoid kernel gets the largest number of support vectors.