NAME: DUONG DUY CHIEN

Support Vector Machines

I. Theory:

1. Gaussian Kernel

$$K_{gaussian}(x^{(i)}, x^{(j)}) = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{k=1}^{n}(x_k^{(i)} - x_k^{(j)})^2}{2\sigma^2}\right)$$

II. Results:

1. Example Dataset 1

In this section, I will adjust C parameter to see how this controls the penalty for misclassified.

- When C is large, the role of B become less important, so that the more opportunity over-fit problem can happen. Intuitively, the decision boundary become more sensitive with data. (Figure 1,2)
- When C is small, the role of B become more important (loss SVM: CA+B), so that the less opportunity over-fit problem can happen. Intuitively, the decision boundary become less sensitive with data. (Figure 3)

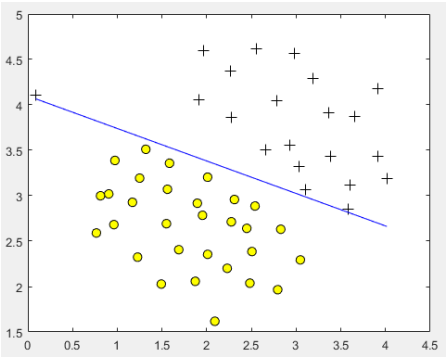

Figure 1 SVM Decision Boundary with C=1000

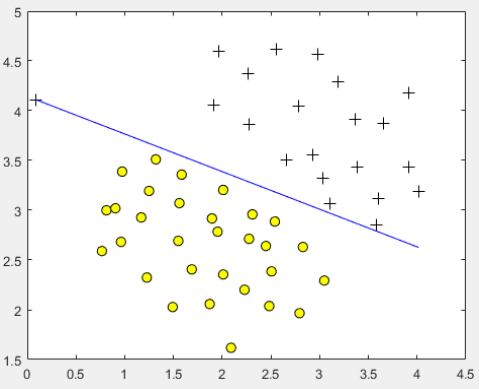

Figure 2 SVM Decision Boundary with C=100

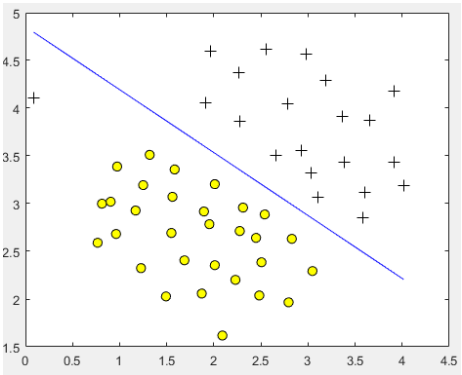

Figure 3 SVM Decision Boundary with C=1

2. SVM with Gaussian Kernels

2.1 Gaussian Kernels

Gaussian Kernels can be think as a similarity function that measures the "distance" between a pair of examples.

2.2 Example dataset 2

After finished Gaussian Kernels, I learn a non-linear decision boundary (using off the shell package) that can perform reasonably well for dataset2.

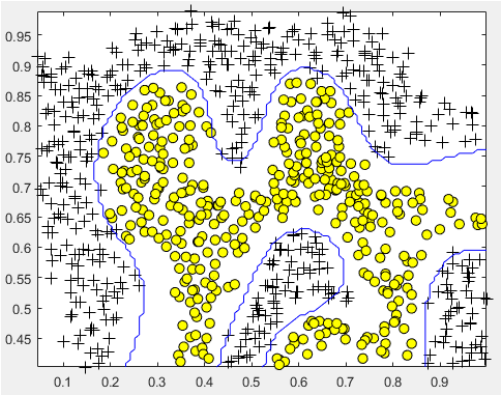

Figure 4 SVM (Gaussian Kernel) Decision Boundary (Example dataset2)

2.3 Example Dataset 3

In this part, I will choose the best C and sigma base on the Cross validation set. Then choose the best C, sigma which produce the smallest error on CV set and use it to draw Decision boundary for dataset3 (Figure 5)

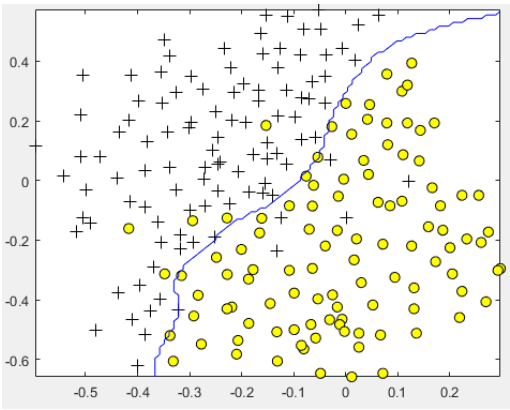

Figure 5 SVM (Gaussian Kernel) Decision Boundary (Example Dataset 3)

3. Spam Classification

3.1 Prepared data:

1. Before to train network, the email data should be repaired (for example Lower-casing, stripping HTML…), the result is shown in Figure 6. Then from the modified email, we will create the vocabulary list which store the most 1899 words happen in dataset.

2. We translate every words in dataset email into the index list of vocabulary. (Figure 7)

```
==== Processed Email ====

anyon know how much it cost to host a web portal well it depend on how mani
visitor you re expect thi can be anywher from less than number buck a month
to a coupl of dollarnumb you should checkout httpaddr or perhap amazon ecnumb
if your run someth big to unsubscrib yourself from thi mail list send an
email to emailaddr
```

Figure 6 Preprocessed Sample email

```
Word Indices:
 86 916 794 1077 883 370 1699 790 1822 1831 883 431 1171 794 1002 1893 1364 592 1676 238 162 89 688 945 1663 1120 1062 1699


375 1162 479 1893 1510 799 1182 1237 810 1895 1440 1547 181 1699 1758 1896 688 1676 992 961 1477 71 530 1699 531
```

Figure 7 Word Indices for Sample Email

3. We need to generate a feature vector for each email, given the word indices. Form matrix nxm (m is number of sample; n is number of words in vocabulary list)

3.2 Training SVM for Spam Classification

- SVM training accuracy is: `Training Accuracy: 99.825000`

- SVM attesting accuracy is: `Test Accuracy: 98.900000`

Top predict spam:

```
Top predictors of spam:
 our              (0.499863)
 click            (0.468342)
 remov            (0.416699)
 guarante         (0.384170)
 visit            (0.368466)
 basenumb         (0.349256)
 dollar           (0.323117)
 price            (0.267879)
 will             (0.265102)
 pleas            (0.260421)
 most             (0.260219)
 lo               (0.256582)
 nbsp             (0.252671)
 ga               (0.243873)
 al               (0.240227)
```

Optional exercise: I make my own spam email in spamSample3.txt to test the learned SVM. The prediction is exact.

My Spam email:

```
==== Processed Email ====

hi ta machin learn i am chien and i want to make a spam email to test svm
which have some word like our or click and guarant you can contact me through
thi number haha number number number make the call and get the fact invest
number minut in yourself now number number number look forward to your call
and i will introduc you to peopl like yourself who ar current make dollarnumb
number plu per week number number number numberljgvnumb
numberleannumberlrmsnumb numberwxhonumberqiytnumb numberrjuvnumberhqcfnumb
numbereidbnumberdmtvlnumb
```

Figure 8 My spam email

Result of trained SVM:

```
Processed spamSample3.txt

Spam Classification: 1
(1 indicates spam, 0 indicates not spam)
```