RICE UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE

Assignment 4, COMP 540

Chen Zeng(cz39), Zhihui Xie(zx18)

March 5, 2018

1 Intuitions about support vector machines

Answer1 The margin is the distance of the separation line to the closest points of each class. Intuitively, bigger margin makes it to be more robust to noise and it can give a safer decision boundary. It means a little error in measurement of the input won't lead to misclassification. Thus, intuitively, a bigger margin will result in a model that will generalize better, or perform better in practice.

Answer2 No, it won't. The hinge loss function is defined as $J(\theta,\theta_0) = C\sum_{i=1}^{m} max\left(0, 1 - y^{(i)} h_\theta\left(x^{(i)}\right)\right)$. As for points that are not support vectors, the second term of max function is negative and moving them to be further away from the decision boundary will make it to be more negative. Therefore, no matter how you remove them, the max function will return 0. It proves that the loss function is entirely determined by the support vectors.

2 Fitting an SVM classifier by hand

Answer1 According to the feature vector $\phi(x) = \left(1, \sqrt{x}, x^2\right)$, we can map the point $\left(x^{(1)}, y^{(1)}\right)$ and $\left(x^{(2)}, y^{(2)}\right)$ which are $(0,-1)$ and $(2,+1)$ to $\left(\phi\left(x^{(1)}\right), y^{(1)}\right)$ and $\left(\phi\left(x^{(2)}\right), y^{(2)}\right)$ which are $((1,0,0),-1)$ and $((1,2,2),+1)$.
Let's suppose that the line cross these two points in three-dimension has the quation of $ax + by + cz = d$. Because $\phi\left(x^{(1)}\right) = (1,0,0)$ and $\phi\left(x^{(2)}\right) = (1,2,2)$ are on the line, we can get:

$$\begin{cases} a = d \\ a + 2b + 2c = d \end{cases} \Rightarrow \begin{cases} a = d \\ b + c = 0 \end{cases}$$

Therefore, the line cross these two points in three-dimension meets the following equation:

$$\begin{cases} x = 1 \\ y = z \end{cases}$$

Because the optima vector θ should be on this line, this question is just to find a vector which is parallel to the above line. For example, the following scaled vector can meet with this requirement:

$$\theta' = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T$$

Answer2 The decision boundary is vertical to θ. Because the value of margin is the distance from each support vector to the decision boundary, which is half the distance between two points. And these two points are symmetry to the decision boundary. Thus, the value of margin should be the distance between $\phi\left(x^{(1)}\right) = (1,0,0)$ and $\phi\left(x^{(2)}\right) = (1,2,2)$. Here we can calculate it:

$$margin = \frac{1}{2} dis\left(\phi\left(x^{(1)}\right), \phi\left(x^{(2)}\right)\right) = \frac{1}{2} dis\left((1,0,0),(1,2,2)\right) = \frac{1}{2}\sqrt{(1-1)^2 + (0-2)^2 + (0-2)^2} = \sqrt{2}$$

Therefore, the value of margin is $\sqrt{2}$.

Answer3 From the statement of this question, we have:

$$margin = \frac{1}{\|\theta\|} = \sqrt{2}$$

Therefore, we can solve out the value of $\|\theta\|$ as below:

$$\|\theta\| = \frac{\sqrt{2}}{2}$$

As we've got that vector θ is parallel to the vector $\theta' = \left(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^T$ in answer 1, we can set vector θ as below:

$$\theta = \left(0, \frac{\sqrt{2}}{2}t, \frac{\sqrt{2}}{2}t\right)^T$$

Here t is to be solved, and we can get:

$$\|\theta\| = \left\| \left(0, \frac{\sqrt{2}}{2}t, \frac{\sqrt{2}}{2}t\right)^T \right\| = \sqrt{0^2 + \left(\frac{\sqrt{2}}{2}t\right)^2 + \left(\frac{\sqrt{2}}{2}t\right)^2} = |t|$$

At the same time, we've got:

$$\|\theta\| = \frac{\sqrt{2}}{2}$$

Therefore, we can get the value of t as below:

$$t = \pm\frac{\sqrt{2}}{2}$$

It means there are two possible for vector θ:

$$\theta = \pm\left(0, \frac{1}{2}, \frac{1}{2}\right)^T$$

However, the following requirements need to be fulfilled:

$$y^{(1)}\left(\theta^T\phi\left(x^{(1)}\right) + \theta_0\right) \geq 1$$
$$y^{(2)}\left(\theta^T\phi\left(x^{(2)}\right) + \theta_0\right) \geq 1$$

Let's take the value of $y^{(1)} = -1$, $y^{(2)} = 1$, $\phi\left(x^{(1)}\right) = (1,0,0)$, and $\phi\left(x^{(2)}\right) = (1,2,2)$ into it, we can get the following restrictions:

$$-(\theta_0 + \theta_1) \geq 1$$
$$\theta_0 + \theta_1 + 2\theta_2 + 2\theta_3 \geq 1$$

Therefore, we can get the transformation:

$$\theta_2 + \theta_3 \geq \frac{1}{2}(-(\theta_0 + \theta_1) + 1) \geq \frac{1}{2}(1+1) = 1$$

Because the final result must meet the above restriction which is $\theta_2 + \theta_3 \geq 1$, here we have only one positive final result of θ as below:

$$\theta = \left(0, \frac{1}{2}, \frac{1}{2}\right)^T$$

Answer4 Let's take the value of $y^{(1)} = -1$ and $y^{(2)} = 1$ into following equations:

$$y^{(1)}\left(\theta^T\phi\left(x^{(1)}\right) + \theta_0\right) = 1$$
$$y^{(2)}\left(\theta^T\phi\left(x^{(2)}\right) + \theta_0\right) = 1$$

We can get the following equations:

$$\theta^T\phi\left(x^{(1)}\right) + \theta_0 = -1$$
$$\theta^T\phi\left(x^{(2)}\right) + \theta_0 = 1$$

From the value we get in the above solutions, we have:

$$\theta^T\phi\left(x^{(1)}\right) = \left(0, \frac{1}{2}, \frac{1}{2}\right)^T (1,0,0) = 0$$

$$\theta^T\phi\left(x^{(2)}\right) = \left(0, \frac{1}{2}, \frac{1}{2}\right)^T (1,2,2) = 2$$

Therefore, we can get:

$$0 + \theta_0 = -1$$
$$2 + \theta_0 = 1$$

Thus, the final result is that:

$$\theta_0 = -1$$

Answer5 Let the following $\phi(x)$ be equal to 0:

$$\begin{aligned}\phi(x) =& \theta^T x + \theta_0 = \left(0, \frac{1}{2}, \frac{1}{2}\right)\left(1, \sqrt{2}x, x^2\right) - 1 \\ =& \frac{1}{2}x^2 + \frac{\sqrt{2}}{2}x - 1 \\ =& 0\end{aligned}$$

Here we can get the equation for the decision boundary in terms of θ, θ_0, x:

$$x^2 + \sqrt{2}x - 2 = 0$$

3 Support vector machines for binary classification

Answer3.1B Figure 3.1 shows the decision boundary for C = 100. It shows as below.

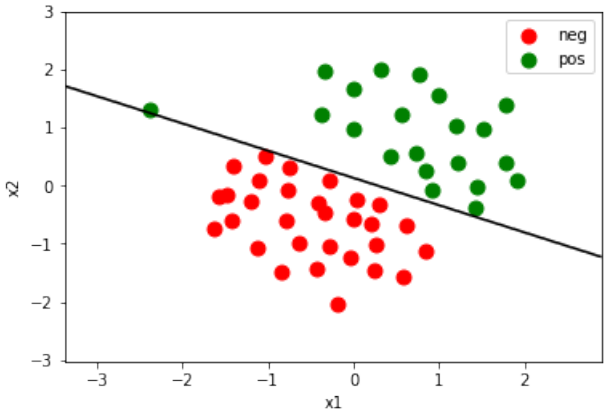

Figure 3.1: decision boundary for C = 100

Answer3.3 As for this task, I tried two models: one is Linear Model, and another one is Gaussian Kernel. As for Linear Model, I tried 8 models where C = 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30. As for Gaussian Kernel, I tried 8*8=64 models where C = 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30 and sigma = 0.01 ,0.03, 0.1, 0.3, 1, 3, 10, 30. I trained each of them for 2000 times iterations and test them on dataset. Here are the results:

```
1. Using Linear Model:
While C = 0.01 , accuracy on validation data = 0.68625
While C = 0.03 , accuracy on validation data = 0.68625
While C = 0.1 , accuracy on validation data = 0.68625
While C = 0.3 , accuracy on validation data = 0.68625
While C = 1 , accuracy on validation data = 0.69625
While C = 3 , accuracy on validation data = 0.89
```

```
While C = 10 , accuracy on validation data = 0.965
While C = 30 , accuracy on validation data = 0.97375
The best case: best_C = 30 , best_accuracy on training data = 0.97375
Accuracy on training data =  0.9709375
Accuracy on test data =  0.973

2. Using Gaussian Kernel:
While sigma = 0.01 , C = 0.01 , accuracy on validation data = 0.77625
While sigma = 0.01 , C = 0.03 , accuracy on validation data = 0.77625
While sigma = 0.01 , C = 0.1 , accuracy on validation data = 0.77625
While sigma = 0.01 , C = 0.3 , accuracy on validation data = 0.77625
While sigma = 0.01 , C = 1 , accuracy on validation data = 0.77625
While sigma = 0.01 , C = 3 , accuracy on validation data = 0.78125
While sigma = 0.01 , C = 10 , accuracy on validation data = 0.78125
While sigma = 0.01 , C = 30 , accuracy on validation data = 0.78125
While sigma = 0.03 , C = 0.01 , accuracy on validation data = 0.77625
While sigma = 0.03 , C = 0.03 , accuracy on validation data = 0.77625
While sigma = 0.03 , C = 0.1 , accuracy on validation data = 0.77625
While sigma = 0.03 , C = 0.3 , accuracy on validation data = 0.77625
While sigma = 0.03 , C = 1 , accuracy on validation data = 0.77625
While sigma = 0.03 , C = 3 , accuracy on validation data = 0.78125
While sigma = 0.03 , C = 10 , accuracy on validation data = 0.78125
While sigma = 0.03 , C = 30 , accuracy on validation data = 0.78125
While sigma = 0.1 , C = 0.01 , accuracy on validation data = 0.78375
While sigma = 0.1 , C = 0.03 , accuracy on validation data = 0.78375
While sigma = 0.1 , C = 0.1 , accuracy on validation data = 0.78375
While sigma = 0.1 , C = 0.3 , accuracy on validation data = 0.78375
While sigma = 0.1 , C = 1 , accuracy on validation data = 0.78375
While sigma = 0.1 , C = 3 , accuracy on validation data = 0.78375
While sigma = 0.1 , C = 10 , accuracy on validation data = 0.79125
While sigma = 0.1 , C = 30 , accuracy on validation data = 0.79125
While sigma = 0.3 , C = 0.01 , accuracy on validation data = 0.7925
While sigma = 0.3 , C = 0.03 , accuracy on validation data = 0.7925
While sigma = 0.3 , C = 0.1 , accuracy on validation data = 0.7925
While sigma = 0.3 , C = 0.3 , accuracy on validation data = 0.7925
While sigma = 0.3 , C = 1 , accuracy on validation data = 0.7925
While sigma = 0.3 , C = 3 , accuracy on validation data = 0.7925
While sigma = 0.3 , C = 10 , accuracy on validation data = 0.8
While sigma = 0.3 , C = 30 , accuracy on validation data = 0.8
While sigma = 1 , C = 0.01 , accuracy on validation data = 0.78875
While sigma = 1 , C = 0.03 , accuracy on validation data = 0.78875
While sigma = 1 , C = 0.1 , accuracy on validation data = 0.78875
While sigma = 1 , C = 0.3 , accuracy on validation data = 0.78875
While sigma = 1 , C = 1 , accuracy on validation data = 0.78875
```

```
While sigma = 1 , C = 3 , accuracy on validation data = 0.785
While sigma = 1 , C = 10 , accuracy on validation data = 0.785
While sigma = 1 , C = 30 , accuracy on validation data = 0.7875
While sigma = 3 , C = 0.01 , accuracy on validation data = 0.79
While sigma = 3 , C = 0.03 , accuracy on validation data = 0.79
While sigma = 3 , C = 0.1 , accuracy on validation data = 0.79
While sigma = 3 , C = 0.3 , accuracy on validation data = 0.79
While sigma = 3 , C = 1 , accuracy on validation data = 0.77625
While sigma = 3 , C = 3 , accuracy on validation data = 0.765
While sigma = 3 , C = 10 , accuracy on validation data = 0.755
While sigma = 3 , C = 30 , accuracy on validation data = 0.7475
While sigma = 10 , C = 0.01 , accuracy on validation data = 0.79375
While sigma = 10 , C = 0.03 , accuracy on validation data = 0.79375
While sigma = 10 , C = 0.1 , accuracy on validation data = 0.79375
While sigma = 10 , C = 0.3 , accuracy on validation data = 0.78875
While sigma = 10 , C = 1 , accuracy on validation data = 0.79
While sigma = 10 , C = 3 , accuracy on validation data = 0.7925
While sigma = 10 , C = 10 , accuracy on validation data = 0.78625
While sigma = 10 , C = 30 , accuracy on validation data = 0.77
While sigma = 30 , C = 0.01 , accuracy on validation data = 0.83
While sigma = 30 , C = 0.03 , accuracy on validation data = 0.83
While sigma = 30 , C = 0.1 , accuracy on validation data = 0.83
While sigma = 30 , C = 0.3 , accuracy on validation data = 0.8275
While sigma = 30 , C = 1 , accuracy on validation data = 0.8275
While sigma = 30 , C = 3 , accuracy on validation data = 0.83
While sigma = 30 , C = 10 , accuracy on validation data = 0.825
While sigma = 30 , C = 30 , accuracy on validation data = 0.8175
The best case: best_sigma = 30 , best_C = 0.01 , best_accuracy on training data = 0.83
Accuracy on training data =  0.9878125
Accuracy on test data =  0.831
```

To conclude, we find the best model is to use Linear Model with C = 30. Therefore, I use this model to train the final model to get spam words with 20000 times iterations. And here are the result of testing accuracy and Top 15 spam and ham words:

```
Accuracy on training data =  0.993125
Accuracy on test data =  0.988

Top 15 words predicted to be spam are:
click
remov
our
nbsp
basenumb
```

```
free
your
will
guarante
pleas
you
here
most
visit
offer

Top 15 words predicted to be ham are:
wrote
date
the
httpaddr
url
spamassassin
re
numbertnumb
it
thei
user
list
my
author
prefer
```

4 Support vector machines for multi-class classification

Answer 4A The zero loss function is not differentiable. Thus, the numerical cannot successfully check. This can be caused by the discrepancy stated in the question.
However, it's not a reason for concern. Because this rarely happens and we usually have a large amount of data.

Answer 4D The best case is:

```
lr 5.000000e-08 reg 5.000000e+04 train accuracy: 0.371776 val accuracy: 0.382000
```

More detailed results can be found in the notebook submitted.

Answer 4E Firstly, let's compare the run time and performance of softmax and multi-class SVM regression:

The run time of multi-class SVM regression is shorter than softmax.
The total prediction accuracy of softmax regression in the last assignment:

```
softmax on raw pixels final test set accuracy: 0.398700
```

The total prediction accuracy of multi-class SVM in this assignment:

```
linear SVM on raw pixels final test set accuracy: 0.368700
```

Also, we can compare the prediction accuracy on different classes as below:

Classes	plane	car	bird	cat	deer	dog	frog	horse	ship	truck
Softmax	0.465	0.486	0.242	0.266	0.299	0.347	0.502	0.425	0.621	0.459
SVM	0.407	0.469	0.297	0.298	0.336	0.343	0.315	0.487	0.381	0.393

From the above comparing on prediction accuracy, we can find on CIFAR-10 task, softmax regression can achieve slightly higher performance. In general, their performance don't different too much.

Secondly, let's compare the visualizations of the θ parameters. 4.1 is the figure showing visualizations of θ parameters in softmax regression and 4.2 is the figure showing visualizations of θ parameters in multi-class SVM. From my perspective of view, we cannot clearly figure out which result of the training model is better from viewing the visualizations of θ parameters. They seems to be very close.

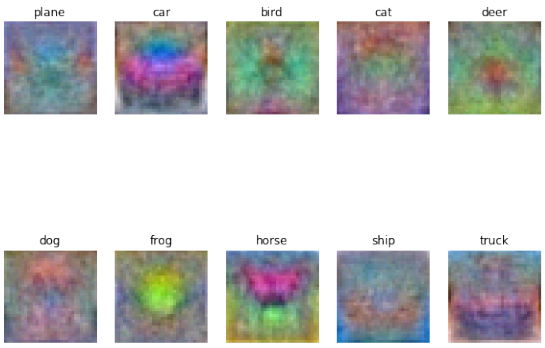

Figure 4.1: visualizations of theta parameters in softmax regression

Figure 4.2: visualizations of theta parameters in multi-class SVM

Finally, to compare their parameters selection process, we can find both of them need to consider the parameter $learning\_rate$ and $regularization\_strength$. And sometimes both of them need to be considered with $batch\_size$ and $iteration$ in the training process. Therefore, the parameters need to be considered in sofatmax regression and multi-class SVM are same and their parameters selection process are very similar.