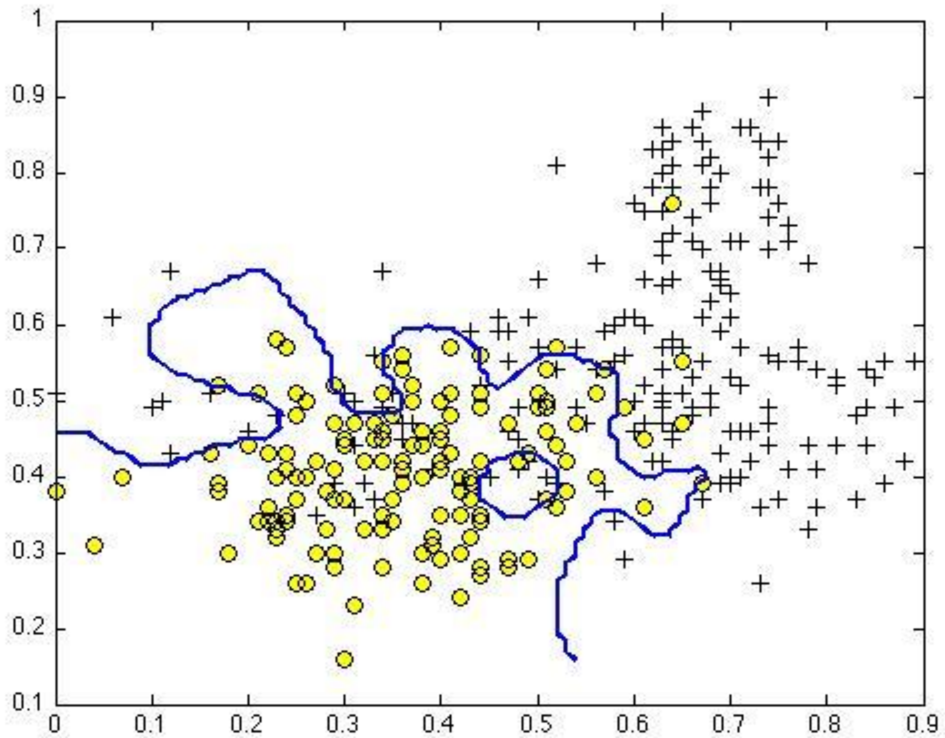
Support Vector Machines

Latest Submission Grade
80%
1.
Question 1

Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



When you measure the SVM's performance on a cross validation set, it does poorly. Should you try increasing or decreasing CCC? Increasing or decreasing σ2\sigma^2σ2?

1 / 1 point

It would be reasonable to try **increasing** CCC. It would also be reasonable to try **decreasing** σ2\sigma^2σ2.

It would be reasonable to try **decreasing** CCC. It would also be reasonable to try **increasing** σ2\sigma^2σ2.

It would be reasonable to try **increasing** CCC. It would also be reasonable to try **increasing** σ2\sigma^2σ2.

It would be reasonable to try **decreasing** CCC. It would also be reasonable to try **decreasing** σ2\sigma^2σ2.

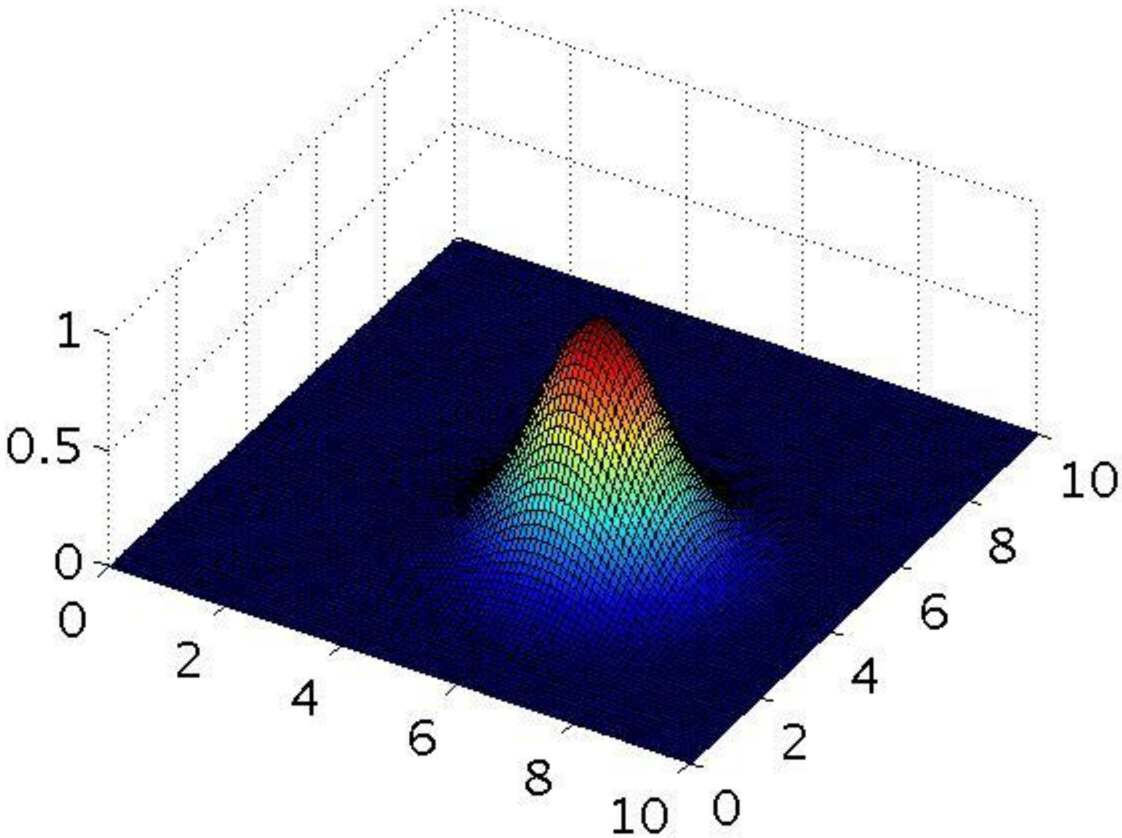
Correct

The figure shows a decision boundary that is overfit to the training set, so we'd like to increase the bias / lower the variance of the SVM. We can do so by either decreasing the parameter CCC or increasing σ2\sigma^2σ2.

2.
Question 2

The formula for the Gaussian kernel is given by similarity(x,l(1))=exp(−||x−l(1)||22σ2)\text{similarity}(x,l^{(1)}) = \exp{(-\frac{||x-l^{(1)}||^2}{2\sigma^2})}similarity(x,l(1))=exp(−2σ2||x−l(1)||2) .

The figure below shows a plot of f1=similarity(x,l(1))f_1 = \text{similarity}(x,l^{(1)})f1=similarity(x,l(1)) when σ2=1\sigma^2 = 1σ2=1.



Which of the following is a plot of f1f_1f1 when σ2=0.25\sigma^2 = 0.25σ2=0.25?

1 / 1 point

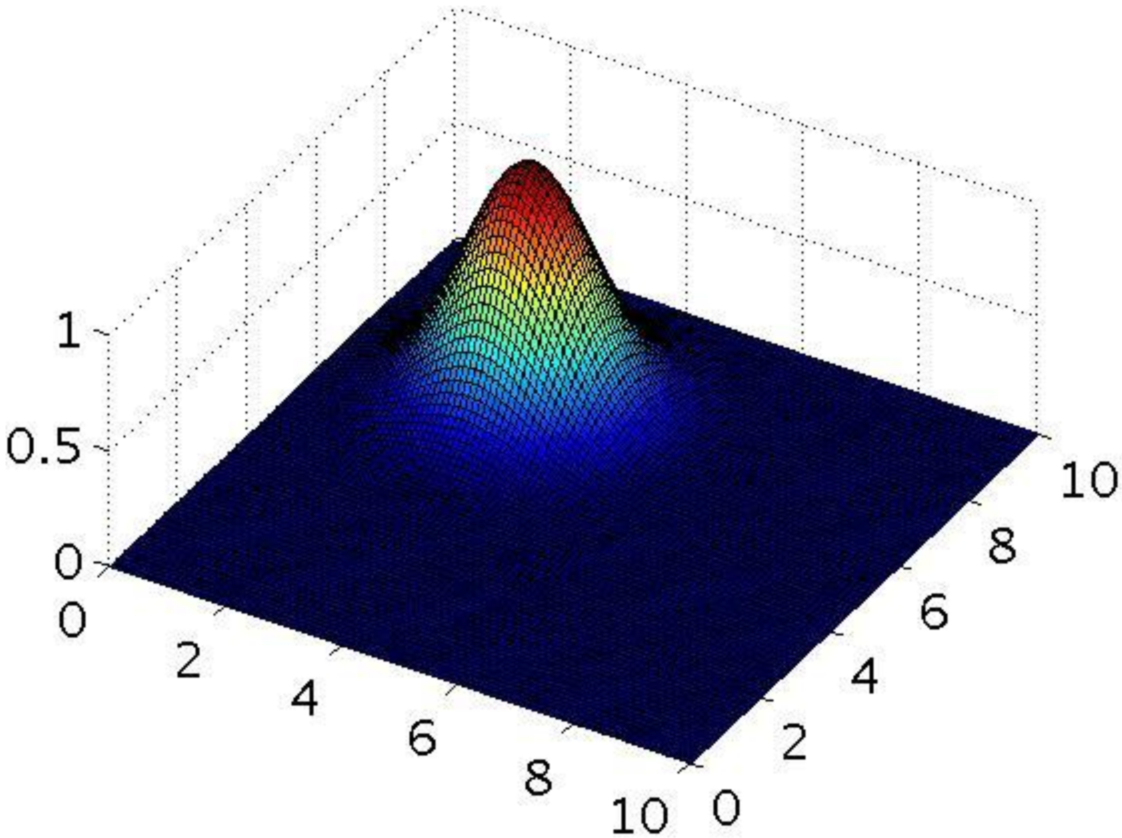


Figure 4.

○

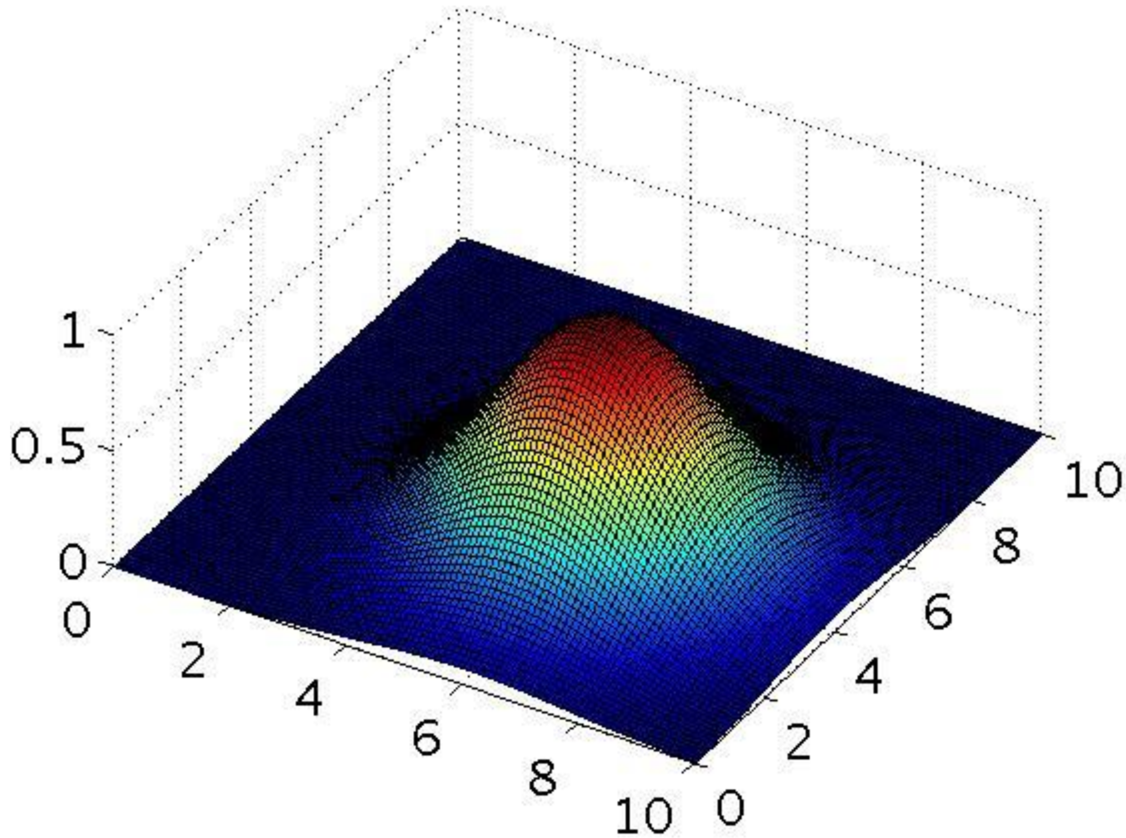
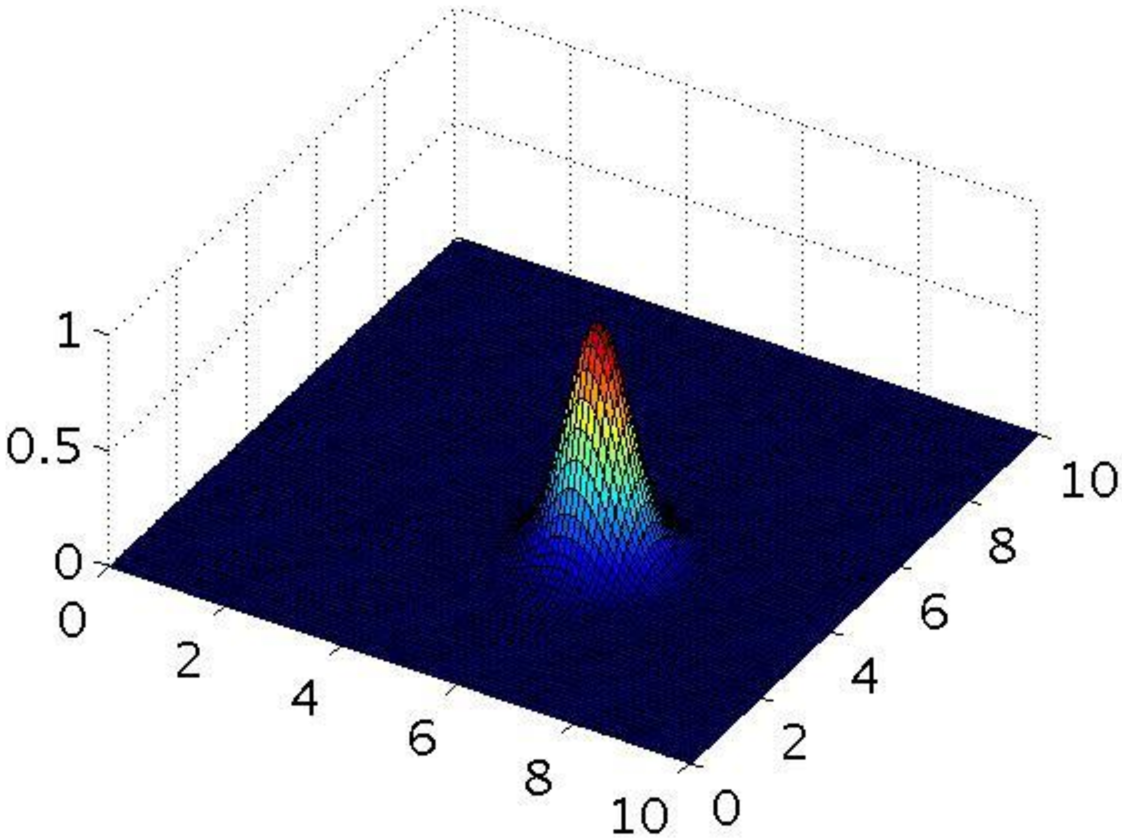


Figure 2.


1
0.5
0
0
2
4
6
8
10
0
2
4
6
8
10

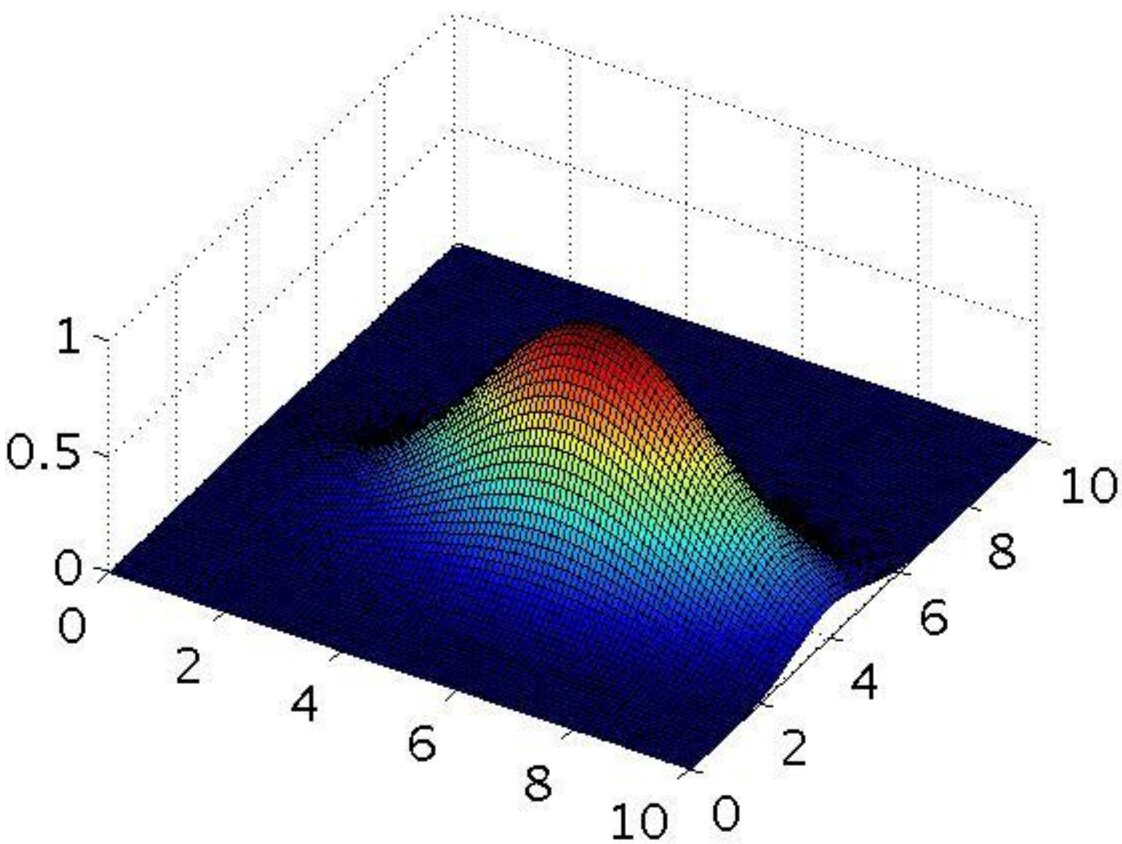

1
0.5
0
0
2
4
6
8
10
0
2
4
6
8
10

Figure 3.

Correct

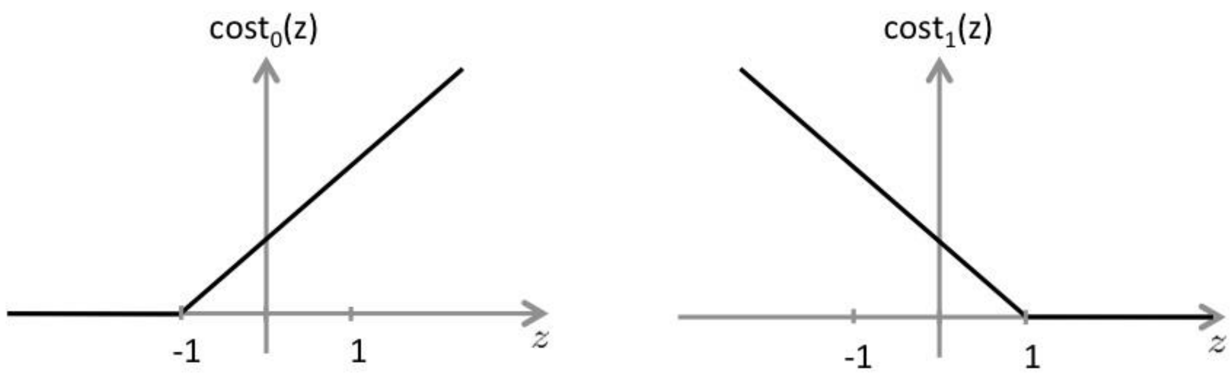
This figure shows a "narrower" Gaussian kernel centered at the same location which is the effect of decreasing σ2\sigma^2σ2.

3.
Question 3

The SVM solves

minθ C∑i=1my(i)cost1(θTx(i))+(1−y(i))cost0(θTx(i))+∑j=1nθj2\min_\theta \space C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^Tx^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^Tx^{(i)}) + \sum_{j=1}^n \theta_j^2minθ C∑i=1my(i)cost1(θTx(i))+(1−y(i))cost0(θTx(i))+∑j=1nθj2

where the functions cost0(z)\text{cost}_0(z)cost0(z) and cost1(z)\text{cost}_1(z)cost1(z) look like this:



The first term in the objective is:

C∑i=1my(i)cost1(θTx(i))+(1−y(i))cost0(θTx(i)).C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^Tx^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^Tx^{(i)}).C∑i=1my(i)cost1(θTx(i))+(1−y(i))cost0(θTx(i)).

This first term will be zero if two of the following four conditions hold true. Which are the two conditions that would guarantee that this term equals zero?

1 / 1 point

☑

For every example with y(i)=1y^{(i)} = 1y(i)=1, we have that θTx(i)≥1\theta^Tx^{(i)} \geq 1θTx(i)≥1.

Correct

For examples with y(i)=1y^{(i)} = 1y(i)=1, only the cost1(θTx(i))\text{cost}_1(\theta^Tx^{(i)})cost1(θTx(i)) term is present. As you can see in the graph, this will be zero for all inputs greater than or equal to 1.

☑

For every example with y(i)=0y^{(i)} = 0y(i)=0, we have that θTx(i)≤−1\theta^Tx^{(i)} \leq -1θTx(i)≤−1.

Correct

For examples with y(i)=0y^{(i)} = 0y(i)=0, only the cost0(θTx(i))\text{cost}_0(\theta^Tx^{(i)})cost0(θTx(i)) term is present. As you can see in the graph, this will be zero for all inputs less than or equal to -1.

☐

For every example with y(i)=0y^{(i)} = 0y(i)=0, we have that θTx(i)≤0\theta^Tx^{(i)} \leq 0θTx(i)≤0.

☐

For every example with y(i)=1y^{(i)} = 1y(i)=1, we have that θTx(i)≥0\theta^Tx^{(i)} \geq 0θTx(i)≥0.

4.
Question 4

Suppose you have a dataset with n = 10 features and m = 5000 examples.

After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets.

Which of the following might be promising steps to take? Check all that apply.

0 / 1 point

☐

Use an SVM with a Gaussian Kernel.

☐

Increase the regularization parameter λ\lambdaλ.

☑

Create / add new polynomial features.

Correct

When you add more features, you increase the variance of your model, reducing the chances of underfitting.

☑

Use an SVM with a linear kernel, without introducing new features.

This should not be selected

An SVM with only the linear kernel is comparable to logistic regression, so it will likely underfit the data as well.

5.
Question 5

Which of the following statements are true? Check all that apply.

1 / 1 point
☐

If the data are linearly separable, an SVM using a linear kernel will

return the same parameters θ\thetaθ regardless of the chosen value of

CCC (i.e., the resulting value of θ\thetaθ does not depend on CCC).

☑

The maximum value of the Gaussian kernel (i.e., sim(x,l(1))sim(x, l^{(1)})sim(x,l(1))) is 1.

Correct

When x=l(1)x = l^{(1)}x=l(1), the Gaussian kernel has value exp(0)=1\exp{(0)} = 1exp(0)=1, and it is less than 1 otherwise.

☐

If you are training multi-class SVMs with the one-vs-all method, it is

not possible to use a kernel.

Suppose you have 2D input examples (ie, x(i)∈R2x^{(i)} \in \mathbb{R}^2x(i)∈R2). The decision boundary of the SVM (with the linear kernel) is a straight line.

Correct

The SVM without any kernel (ie, the linear kernel) predicts output based only on θTx\theta^TxθTx, so it gives a linear / straight-line decision boundary, just as logistic regression does.