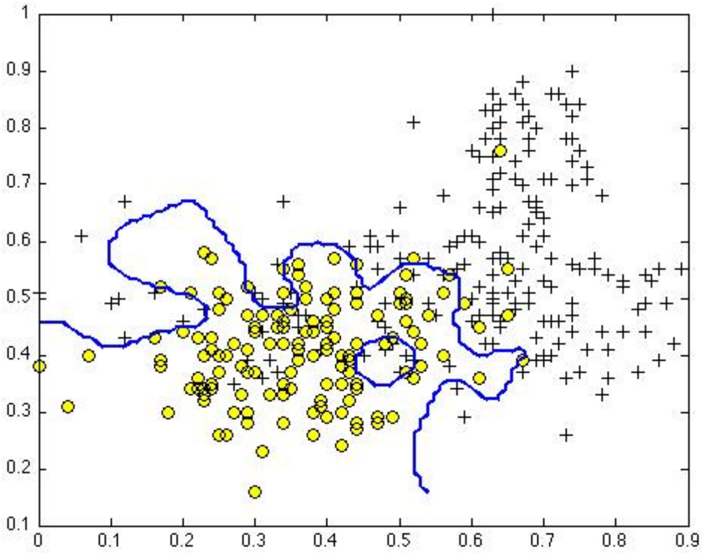
Support Vector Machines

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 C? Increasing or decreasing σ^2?

☑ **It would be reasonable to try decreasing CC. It would also be reasonable to try increasing σ^2.**

> 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 C or increasing σ^2.

□ It would be reasonable to try **increasing** CC. It would also be reasonable to try **increasing** σ^2.

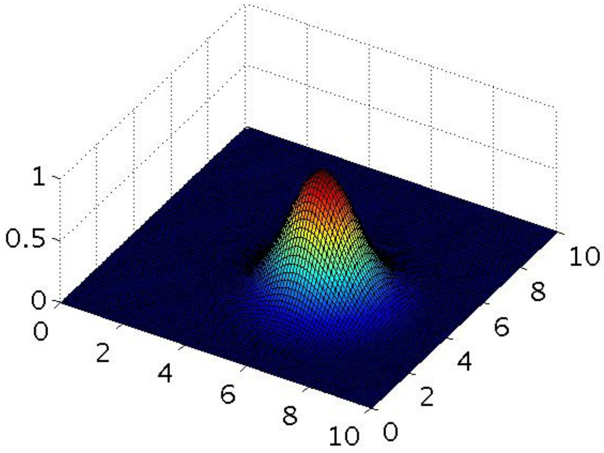
□ It would be reasonable to try **increasing** CC. It would also be reasonable to try **decreasing** σ^2.

□ It would be reasonable to try **decreasing** CC. It would also be reasonable to try **decreasing** σ^2.

Question 2

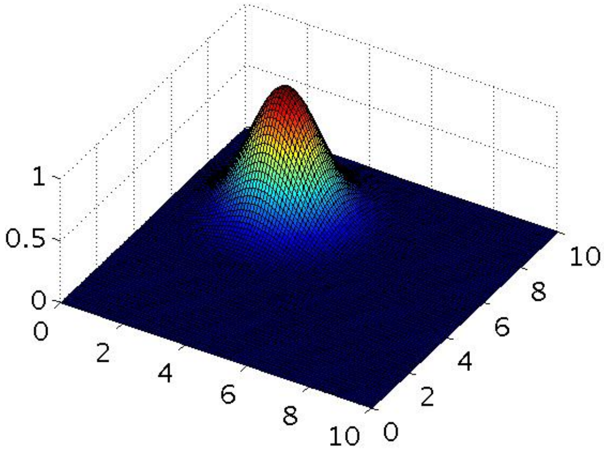
The formula for the Gaussian kernel is given by $\text{similarity}(x, l^{(1)}) = exp(-\frac{||x-l^{(1)}||^2}{2\sigma^2})$.

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

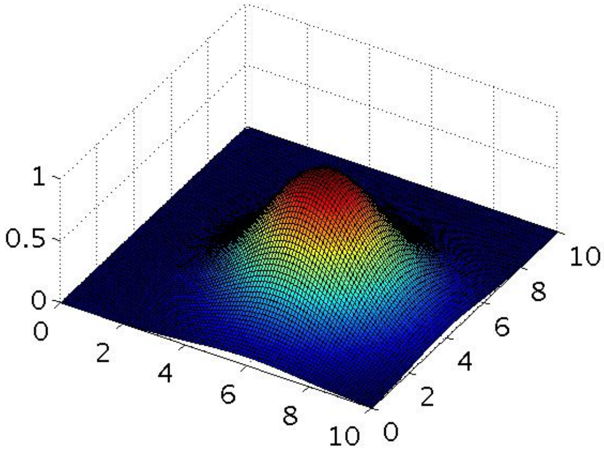


Which of the following is a plot of f_1 when $\sigma^2 = 0.25$?

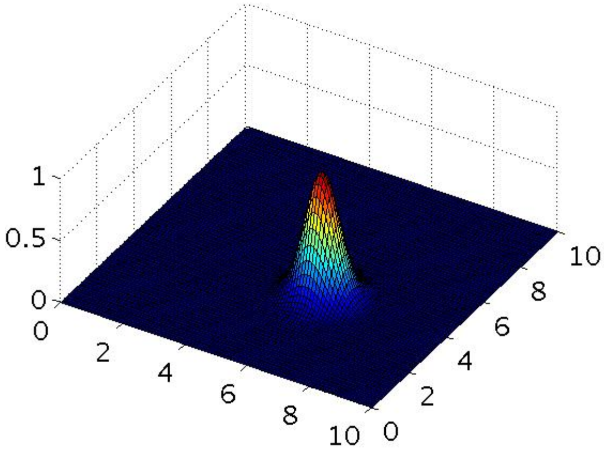
□



□

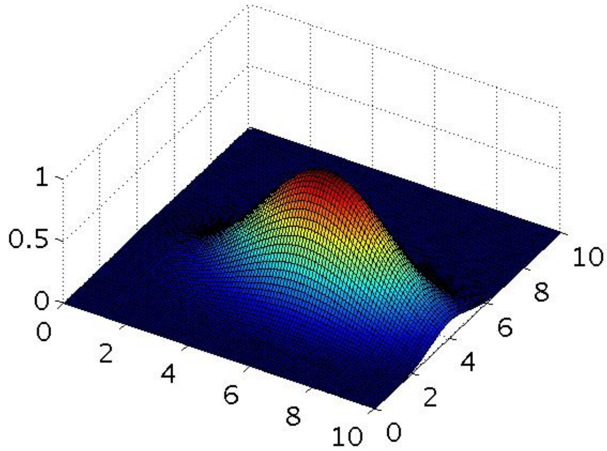


☑



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

□



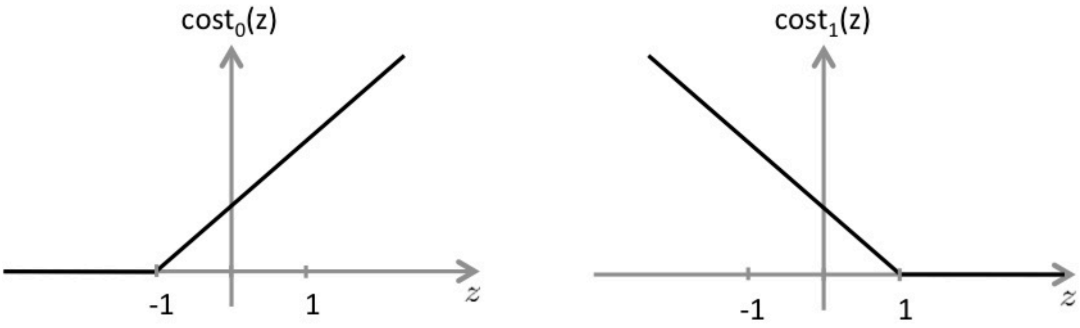
Question 3

The SVM solves

$\min_\theta \ C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) + \sum_{j=1}^n \theta_j^2$

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

his:



The first term in the objective is:

$C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)})$.

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?

☑ **For every example with $y^{(i)} = 1$, we have have that $\theta^T x^{(i)} \geq 1$.**

> For examples with $y^{(i)} = 1$, only the $\text{cost}_1(\theta^T x^{(i)})$ 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 $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq -1$.**

> For examples with $y^{(i)} = 0$, only the $\text{cost}_0(\theta^T x^{(i)})$ 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 $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 0$.

□ For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq 0$.

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.

☑ **Create / add new polynomial features.**

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

□ Reduce the number of examples in the training set.

☑ **Try using a neural network with a large number of hidden units.**

> A neural network with many hidden units is a more complex (higher variance) model than logistic regression, so it is less likely to underfit the data.

□ Use a different optimization method since using gradient descent to train logistic regression might result in a local minimum.

Question 5

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

☑ **It is important to perform feature normalization before using the Gaussian kernel.**

> The similarity measure used by the Gaussian kernel expects that the data lie in approximately the same range.

□ If the data are linearly separable, an SVM using a linear kernel will return the same parameters θ regardless of the chosen value of C (i.e., the resulting value of θ does not depend on C).

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

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

□ Suppose you are using SVMs to do multi-class classification and would like to use the one-vs-all approach. If you have K differentclasses, you will train K - 1 different SVMs.