The Coders' Club

Machine Learning: G1

Week 7: Assignment

Topics:
Support Vector Machines

Some Additional Courses:

- Machine Learning Onramp (MathWorks)
  https://www.mathworks.com/learn/tutorials/machine-learning-onramp.html

- Deep Learning Onramp (MathWorks)
  https://www.mathworks.com/learn/tutorials/deep-learning-onramp.html

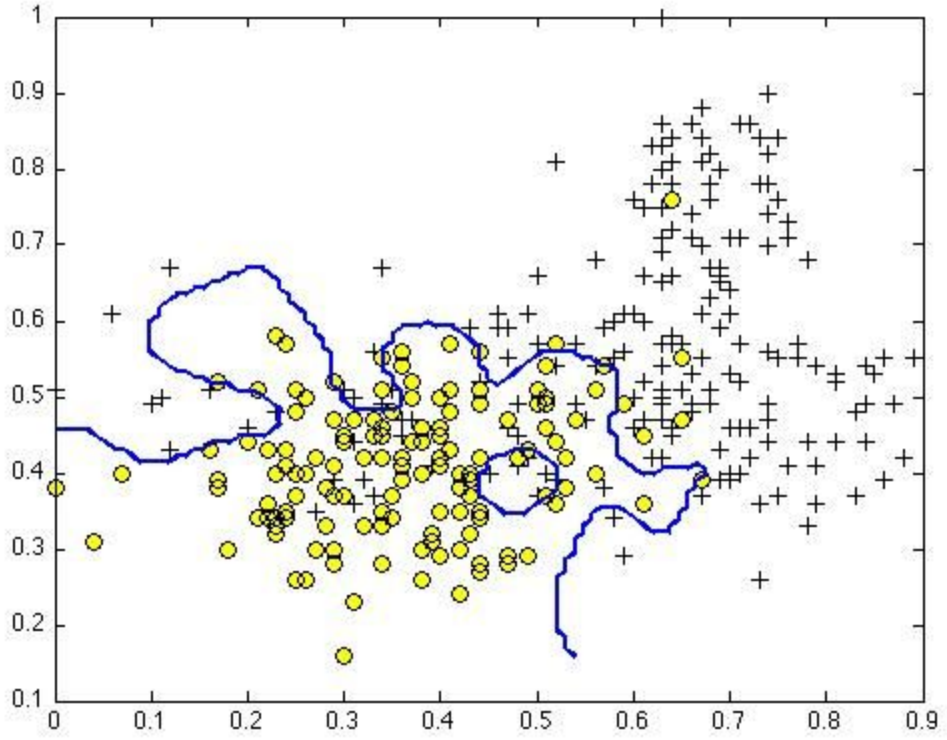
- AI From the Data Center to the Edge – An Optimized Path Using Intel® Architecture (Intel AI)
  https://software.intel.com/en-us/ai/courses/data-center-to-edge

- Machine Learning (Intel)
  https://software.intel.com/en-us/ai/courses/machine-learning

- Deep Learning (Intel)
  https://software.intel.com/en-us/ai/courses/deep-learning

Support Vector Machines

Q.1. Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



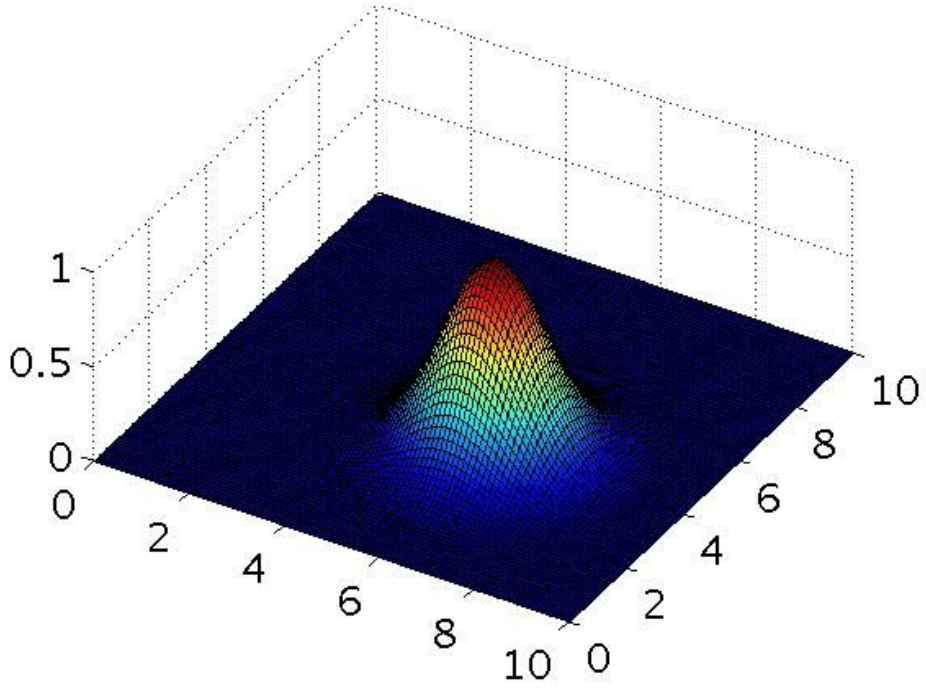
When you measure the SVM's performance on a cross validation set, it does poorly. Should you try increasing or decreasing C? Increasing or decreasing σ^2?

- It would be reasonable to try **decreasing** C. It would also be reasonable to try **increasing** σ^2
- It would be reasonable to try **decreasing** C. It would also be reasonable to try **decreasing** σ^2
- It would be reasonable to try **increasing** C. It would also be reasonable to try **increasing** σ^2
- It would be reasonable to try **increasing** C. It would also be reasonable to try **decreasing** σ^2

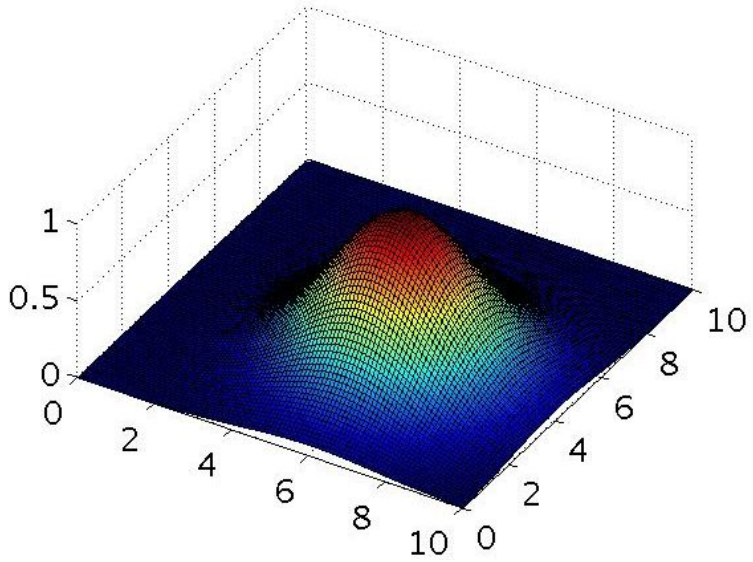
Q.2. The formula for the Gaussian kernel is given by:

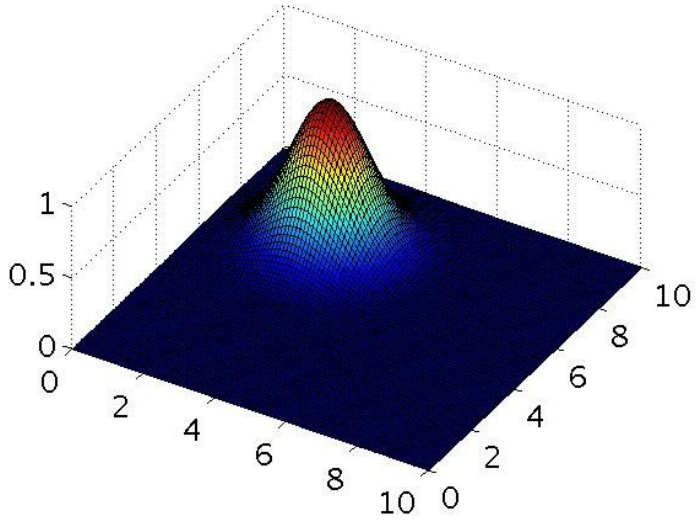
$$\text{similarity}(x, l^{(i)}) = \exp\left(-\frac{||x - l^{(i)}||^2}{2\sigma^2}\right)$$

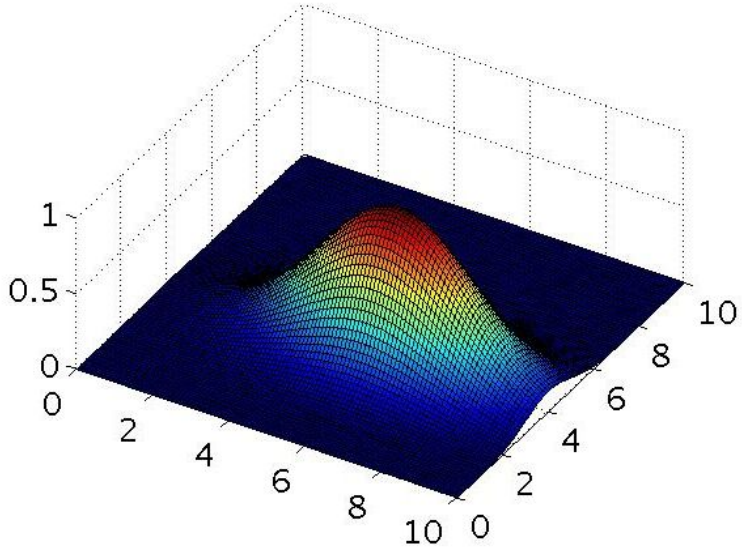
The figure below shows a plot of f_1 = similarity(x, $l^{(1)}$) when $\sigma^2 = 1$.

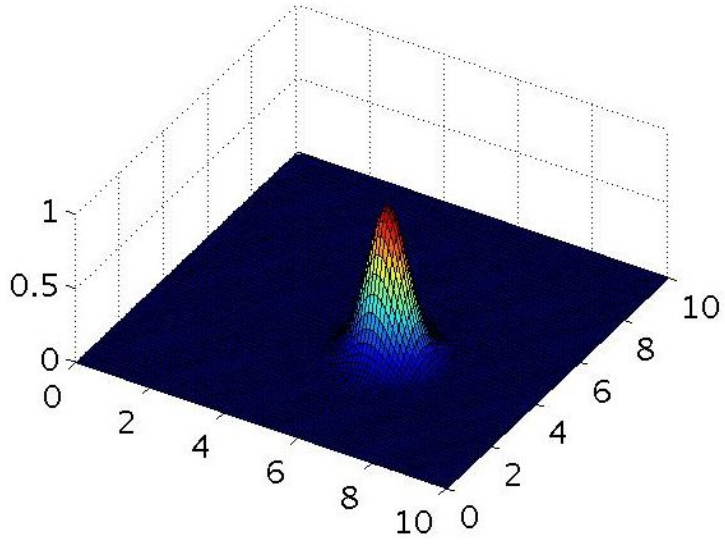


Which of the following is a plot of f_1 when $\sigma^2 = 0.25$?




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

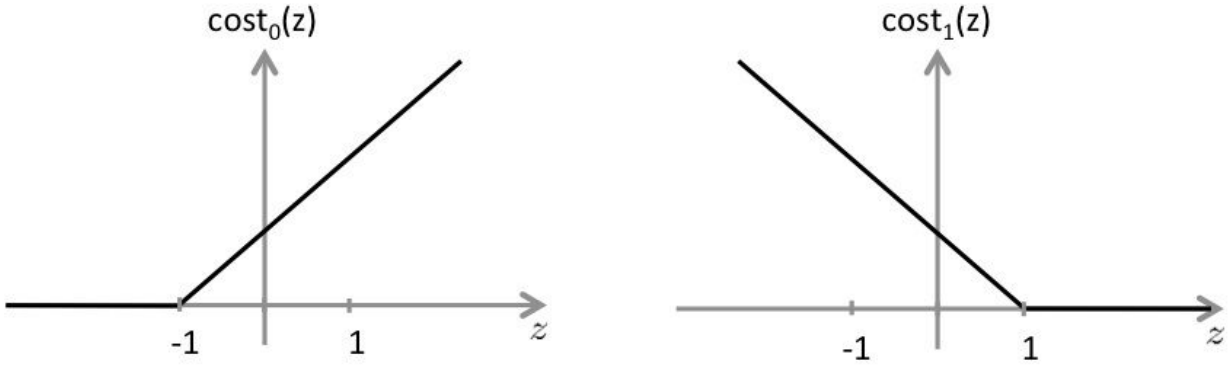

1
0.5
0
0
2
4
6
8
10
0
2
4
6
8
10


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

Q.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:



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 that $\theta^T.x^{(i)} \geq 1$
- For every example with $y^{(i)} = 0$, we have that $\theta^T.x^{(i)} \leq -1$
- For every example with $y^{(i)} = 0$, we have that $\theta^T.x^{(i)} \leq 0$
- For every example with $y^{(i)} = 1$, we have that $\theta^T.x^{(i)} \geq 0$

Q.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.

- Use an SVM with a linear kernel, without introducing new features.
- Create/add new polynomial features.
- Use an SVM with a Gaussian Kernel.
- Increase the regularization parameter λ

Q.5. Which of the following statements are true? Check all that apply.

- 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).

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

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

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