CS3244 Machine Learning Homework 2 Essay Question

Dong Shaocong A0148008J

October 10, 2017

Abstract

This document is the essay question for CS3244 machine learning homework 2. This homework is about support vector machines. Throughout this homework, we examined several different kernels with different methods of implementation. In the essay questions, we examined different treatments for errors, a small example with 6 data points, and two different but valid kernel functions.

1 Question 1

Statement of the problem

Consider a binary classification problem. Let us denote the input set as $\{x_n, y_n\}$ where x_n is the data point and $y_n \in \{-1, +1\}$ is the corresponding class label. Further, misclassifying a positive class data point costs k times more than misclassifying a negative class data point. Rewrite the SVM optimization function to model this constraint.

1.1 Answer

Reformulation:

Maximize: $\frac{1}{2} \times w^T w + C \times \sum_{n=1}^{N} (k\varepsilon_n[Data\ i\ is\ false\ negative] + \varepsilon_n[Data\ i\ is\ false\ positive])$

subject to: $y_n(w^T x_n + b) \geq 1 - k\varepsilon_n$ for $n = 1, 2, , , N$
$\varepsilon_n \geq 0$ for $n = 1, 2, , , N$ and ε, w are real valued vectors, b is real value

where we have $[A] = 1$ when A is true (event A happens)

Explanation: We want to penalize on the error, or the points that are violating the margin. Given a false negative, we want to penalize k times heavier than the false positive points. As a result, we have to differentiate between the two types of error in the optimization function, and times k for the false negative errors. Thus, I changed the optimization function to incorporate this changes of modelling.

2 Question 2

2.1 a. Plot these six training points. Are the classes linearly separable?

x_1	x_2	y
1	1	+
2	2	+
2	0	+
0	1	-
1	0	-
0	0	-

For the above data, the points are scatter plotted using the following *Python* snippet.

```
import numpy as np
from matplotlib import pyplot as plt

x1 = np.vstack((1,2,2,0,1,0)).reshape(1,6)[0]
x2 = np.vstack((1,2,0,1,0,0)).reshape(1,6)[0]
y = np.vstack((1,1,1,-1,-1,-1)).reshape(1,6)[0]

pl.scatter(x1,x2,c=y)
```

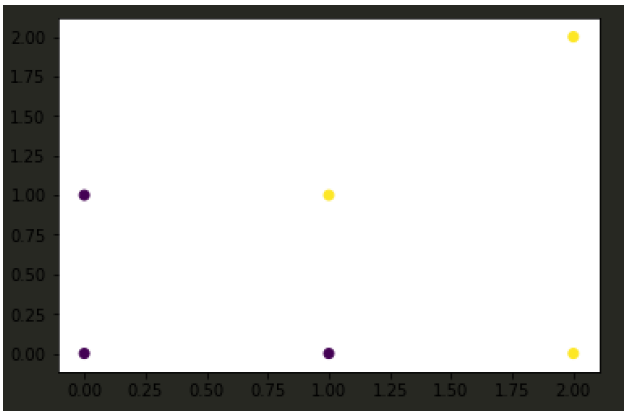

Figure 1: Scatter plot of the six points

From the scatter plot above, we can see the six training points are linearly separable. The purple points are located mainly at the left lower corner whereas the yellow points are more concentrated at the top right hand side.

2.2 b. Construct the weight vector of the maximum margin hyper-plane by inspection and identify the support vectors.

By inspection, we can see the clear boundary between the training data sets. There are four support vectors, which are $(0,1),(1,0),(1,1),(2,0)$. The maximum margin hyper plane follows this equation $x_2 = 1.5 - x_1$. Actually the weight vector is $(1,1)$ and the offset b is of value -1.5.

Using the following code snippet, we can get the maximum margin hyper plane

```
x_lin = np.arange(0, 2, 0.1)
boundary_function = lambda x1: 1.5-x1
y_lin = boundary_function(x_lin)
plt.plot(x_lin, y_lin)
```

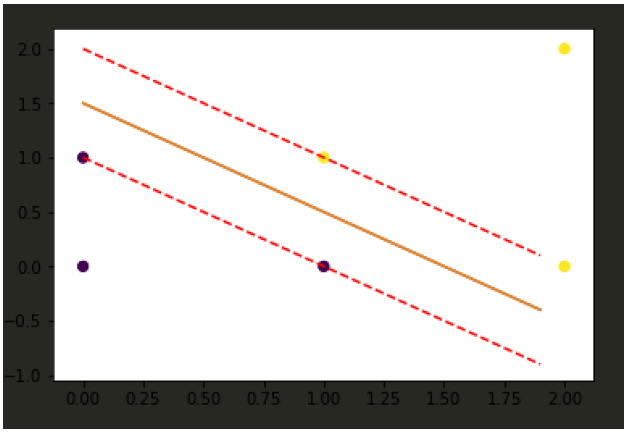

Figure 2: Scatter plot of the six points with maximum margin hyper plane

2.3 c. If you remove one of the support vectors, does the size of the optimal margin decrease, stay the same, or increase? Explain.

It depends. As discussed previously, these four are the support vectors: $(0,1),(1,0),(1,1),(2,0)$. If you remove $(1,0)$ or $(1,1)$, the size will increase from $\frac{1}{\sqrt{2}}$ to $\frac{1}{2}$. If you remove $(0,1)$ or $(2,0)$, the size of the optimal margin will stay the same.

2.4 d. Is your answer to (c) also true for any dataset? Provide a counterexample or give a short proof.

True, the size of the optimal margin will be increasing or staying the same with less number of support vectors given (some of them (removed).

Proof: Suppose we have a set of all the support vectors for a particular support vector machine problem. These support vectors become the real effective constraints in deriving the weight vectors. For SVM problem, the optimization function is to maximize the margin. Given a subset of the support vectors, we have a subset of the original constraints, thus the yielded weight vector will give a margin that is at least as the same as the original margin. There's also good chance that the margin will become larger. (subset of constraints will not decrease the maximized value we obtained)

3 Question 3

Statement of the problem

A kernel is valid if it corresponds to a scalar product in some (perhaps infinite dimensional) feature space. Remember a necessary and sufficient condition for a function $K(x,x)$ to be a valid kernel is that associated Gram matrix, whose elements are given by $K(x_n,x_m)$, should be positive semi-definite for all possible choices of the set x. Show whether the following are also kernels:

3.1 $K(x,x') = \mathbf{c} < x,x' >$

It is a kernel.
From tutorial, we know a property of the kernel function as follows:

" If $K(x,x')$ is a valid kernel, $q[K(x,x')]$ is also a valid kernel, where q[.] is a polynomial function with a non-negative coefficients. "

Because c is just a constant, and $< x,x' >$ is a valid kernel by itself, we can infer that $K(x,x') =$ c $< x,x' >$ is a valid kernel.

3.2 $K(x,x') = < x,x' >^2 + e^{(-||x||^2)} \times e^{(-||x'||^2)}$

It is a kernel.
I will use the following properties of kernel to prove this:

1. $K(x,z) = g(x)g(z)$ is a valid kernel for function $g : X \to R$.
2. $e^{(-||x||^2)}$ is a real valued function from X to R.
3. given two valid kernels $K1$ and $K2$: $K(x,x') = K1(x,x') \times K2(x,x')$ is also a valid kernel.
4. given two valid kernels $K1$ and $K2$: $K(x,x') = K1(x,x') + K2(x,x')$ is also a valid kernel.

Follows from property 1 and 2, we can have $e^{(-||x||^2)} \times e^{(-||x'||^2)}$ is indeed a valid kernel.
Follows from property 3, $< x,x' >^2$ is also a valid kernel.
Follows then from property 4, it is a valid kernel function.