HMC CS 158, Fall 2017
Problem Set 5 Exercises: SVMs, Kernels

Goals:

- To investigate how variations of the SVM optimization problem affect the hyperplane and margin.
- To practice identifying and constructing kernels. To explore how kernel properties affect classification.

Submission

You should submit any answers to the exercises in a single file `writeup.pdf`. This writeup should include your name and the assignment number at the top of the first page, and it should clearly label all problems. Additionally, cite any collaborators and sources of help you received (excluding course staff), and if you are using slip days, please also indicate this at the top of your document.

1 SVM Primal [6 pts]

Suppose we are looking for a maximum-margin linear classifier *through the origin*, i.e. $b = 0$. This classifier should also use a hard margin, that is, no slack variables. In other words, we minimize $\frac{1}{2}||\boldsymbol{\theta}||^2$ subject to $y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \geq 1, i = 1, \ldots, n$.

(a) **(1 pts)** Given a single training vector $\boldsymbol{x} = [a, e]^T$ with label $y = -1$, what is the $\boldsymbol{\theta}^*$ that satisfies the above constrained minimization? (As usual with SVMs, you may find it useful to sketch a picture.)

(b) **(2 pts)** Suppose we have two training examples, $\boldsymbol{x}^{(1)} = [1, 1]^T$ and $\boldsymbol{x}^{(2)} = [1, 0]^T$ with labels $y^{(1)} = 1$ and $y^{(2)} = -1$. What is $\boldsymbol{\theta}^*$ in this case, and what is the margin γ?

(c) **(3 pts)** Suppose we now allow the offset parameter b to be non-zero. How would the classifier and the margin change in the previous question? What are $(\boldsymbol{\theta}^*, b^*)$ and γ? Compare your solutions with and without offset.

2 Kernels [9 pts]

(a) **(3 pts)** Assume that the function k is a kernel, i.e., $k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ is the inner product of vectors $\phi(\boldsymbol{x}^{(i)})$ and $\phi(\boldsymbol{x}^{(j)})$ for some function ϕ. Show that given any n vectors (where n can be any positive integer), the $n \times n$ Gram matrix $\boldsymbol{K}$ with $K_{ij} = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ should be positive semi-definite.

(b) **(2 pts)** For any two documents $\boldsymbol{x}$ and $\boldsymbol{z}$, define $k(\boldsymbol{x}, \boldsymbol{z})$ to equal the number of unique words that occur in both $\boldsymbol{x}$ and $\boldsymbol{z}$ (i.e., the size of the intersection of the sets of words in the two documents). Is this function a kernel? Give justification for your answer.

Parts of this assignment are adapted from course material by Tommi Jaakola (MIT) and Jenna Wiens (UMich).

(c) **(2 pts)** One way to construct kernels is to build them from simpler ones. We have seen various "construction rules", including the following: Assuming $k_1(\boldsymbol{x}, \boldsymbol{z})$ and $k_2(\boldsymbol{x}, \boldsymbol{z})$ are kernels, then so are

- (scaling) $f(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{z})f(\boldsymbol{z})$ for any function $f(\boldsymbol{x}) \in \mathbb{R}$
- (sum) $k(\boldsymbol{x}, \boldsymbol{z}) = k_1(\boldsymbol{x}, \boldsymbol{z}) + k_2(\boldsymbol{x}, \boldsymbol{z})$
- (product) $k(\boldsymbol{x}, \boldsymbol{z}) = k_1(\boldsymbol{x}, \boldsymbol{z})k_2(\boldsymbol{x}, \boldsymbol{z})$

Using the above rules and the fact that $k(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{z}$ is (clearly) a kernel, show that the following is also a kernel:

$$\left(1 + \left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||}\right) \cdot \left(\frac{\boldsymbol{z}}{||\boldsymbol{z}||}\right)\right)^3$$

(d) **(2 pts)** Suppose that you are given a dataset with 10 binary-labeled two-dimensional examples $\boldsymbol{x} \in \mathbb{R}^2$. You experiment with the following kernels in a soft-margin SVM framework:

- $k_1(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x} \cdot \boldsymbol{x}'$
- $k_2(\boldsymbol{x}, \boldsymbol{x}') = 1 - 3(\boldsymbol{x} \cdot \boldsymbol{x}')^3$
- $k_3(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}' + 2)^5$
- $k_4(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}') + 2(\boldsymbol{x} \cdot \boldsymbol{x}')^2$

Select one kernel for each of the following findings, and provide your rationale. You should use each kernel exactly once.

i. The optimization routine for the dual SVM problem complained about one of the kernels (it is not valid), and it could not be used further. Which one (1-4)? ____

ii. Only one of the kernels resulted in zero training error.
Which one (1-4)? ____

iii. As far as the training error is concerned, one of the kernels was by far the worst.
Which one (1-4)? ____

iv. We apply the three resulting classifiers to held-out test data. Which one would you expect to obtain the lowest test error (1-4)? ____