CS 4780/5780: Machine Learning for Intelligent Systems **(Due: 10/08/19)**

Assignment #3: SVMs, Kernels, Duality, and Leave-One-Out Errors

Instructor: Nika Haghtalab, Thorsten Joachims *Name:* Student name(s), *Netid:* NetId(s)

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- Please include your name and NetIDs on the first page. We recommend typesetting your submission in LaTeX, and an Overleaf template is linked here.
- Assignments are due by noon on the due date in PDF form on Gradescope.
- Late assignments can be submitted on Gradescope up to noon Sunday, Oct 13. This is also when the solutions will be released.
- You can do this assignment in groups of 2-3. Please submit no more than one submission per group. Collaboration across groups is not permitted.
- All sources of material outside the course must be cited. The University Academic Code of Conduct will be strictly enforced.

Submission Instructions: All group members must be added to the Gradescope submission. If you're the one submitting, add your group members on Gradescope. Otherwise, make sure you are added to the submission. Put your names on the PDF, this helps us track the groups in case there are errors on Gradescope. See this Piazza post for more information.

Problem 1: Hard-Margin SVMs

(10 + 13 + 10 = 33 points)

In this problem we will be working exclusively with a small toy dataset which is visualized in Figure 1. Consider all red training examples as negative instances and all blue training examples as positive instances. Each of the coordinates is an integer value. In this problem, we will explore some properties of hard-margin SVMs.

(a) Let us first think about how a hard-margin SVM would do on this problem. Draw a hyperplane with latex or by hand that achieves the largest hard margin on S without using an SVM package. Clearly mark all support vectors in your diagram. See Overleaf template for drawing assistance.

(b) Now, compute by hand the weight vector w_{opt} and bias b_{opt} of the maximum margin hyperplane of the SVM in (a). The weight vector must satisfy $\|w_{opt}\|_2 = 1$. What is the geometric margin γ_{opt} over the sample of data points in Figure 1? Also indicate which dual variables α_i are non-zero. Show all work leading to your answer.

(c) You decide to implement and train a dual batch perceptron, like the one seen in lecture. You also want to make your implementation memory efficient by allocating the smallest possible integer datatype (e.g., byte vs int32) for storing the dual variables, but for that you need to know the largest possible value that any dual variable can take without explicitly limiting the number of iterations. For the provided toy dataset, what is the tightest bound that you can obtain on any α_i, assuming that the perceptron obtains 0 training error on this dataset? Assume the length of all training examples is bounded by 12, i.e. $\max_i \|x_i\|_2 \leq 12$. Show your derivation, and your answer should not exceed three sentences.

Problem 2: Kernels

(8 + 8 + 9 + 8 = 33 points)

Consider one-dimensional datasets of the form $D_n^6 = \{x_i, y_i\}_{i=1...n}$ where n is a natural number, x_i is an integer, and $y_i \in \{-1, +1\}$. The data are generated as follows:

$$y_i = \begin{cases} +1, & x_i \mod 6 < 3 \\ -1, & \text{else} \end{cases}$$

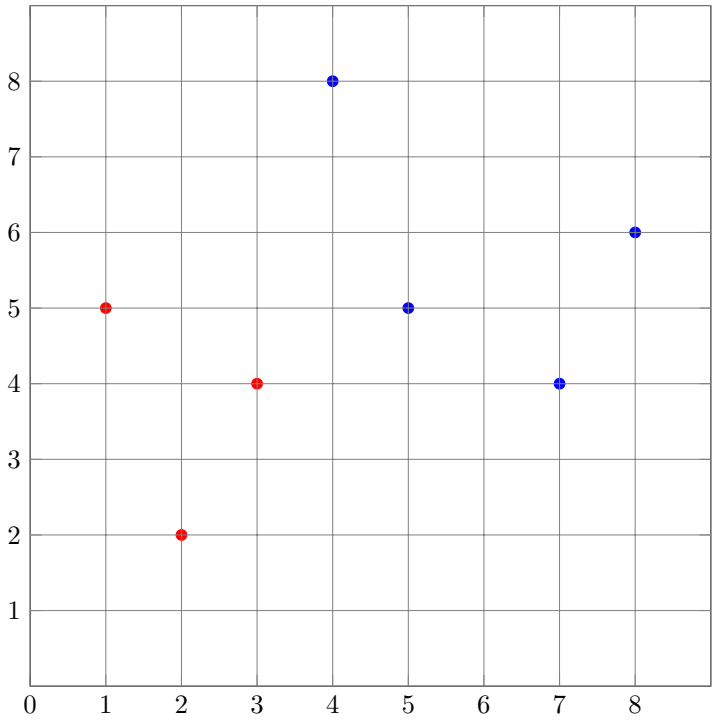

Figure 1: Toy dataset.

(a) Show that any such dataset D_n^6 is linearly separable in the feature space

$$\phi(x_i) = (\cos Ax_i, \sin Ax_i)$$

where $A = \frac{\pi}{3}$, by plotting the data in the feature space and (visually) identifying a linear separator. See Overleaf template for plotting assistance.

(b) Show that the kernel function $K(x, x')$ for the feature space above is given by:

$$K(x, x') = \cos(A(x - x'))$$

You may refer to this link for a list of common trigonometric identities.

(c) Suppose that you train a hard-margin kernelized SVM with the above feature space on a dataset D_{19}^6 where x_i ranges over all the integers in the interval $[-8, 10]$. Draw the resulting decision boundary in the original (one-dimensional) instance space. List all instances that lie on the margin.

(d) Now consider datasets of the form $D_n^8 = \{x_i, y_i\}_{i=1\ldots n}$ where

$$y_i = \begin{cases} +1, & x_i \mod 8 < 4 \\ \text{-1}, & \text{else} \end{cases}$$

Define a feature space $\phi(x_i)$ such that any such dataset D_n^8 is linearly separable in it.

Problem 3: More on SVMs

(17 + 17 = 34 points)

In this problem we will investigate the training of linear classifiers on n training examples, $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where each $x_i \in \mathbb{R}^N$ and each $y_i \in \{-1, 1\}$.

(a) Suppose that our dataset S has the following property:

- The maximum length of all feature vectors x_i is 1, i.e. $\max_{1 \le i \le n} x_i^\top x_i = 1$.

i	y_i	$w^\top x_i + b$	α_i
1	-1	-1	0
2	1	0.4	0.1
3	-1	-0.1	0.1

Table 1: Problem 3a

We train a linear SVM classifier on this data, and get (w, b). The first three instances are as follows (Table 1). α_i is the dual variable of example i.

What is the upper bound on the leave-one-out error for the three instances? Explain your solution in at most 3 sentences.

(b) Now suppose we have a different dataset S', which has the following two properties

- The length of all feature vectors x_i is 1, i.e. $\forall i : x_i^\top x_i = 1$
- Any two feature vectors in our training set are orthogonal, i.e. $\forall i \neq j : x_i^\top x_j = 0$.

Suppose we train a homogeneous hard margin SVM on S'. For any training example (x_i, y_i), what is the corresponding dual variable α_i? Show your work

Hint: Recall from lecture that the dual optimization problem requires to maximize the objective

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\vec{x}_i \cdot \vec{x}_j)$$

such that $\forall i = 1, \ldots n, \quad \alpha_i \geq 0$. Note that because we train a homogeneous linear classifier we do not need the equality constraint.