CS 4780/5780: Machine Learning for Intelligent Systems **(Due: 10/13/20)**

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 on Canvas.
- Assignments are due by 5pm on the due date in PDF form on Gradescope.
- Late assignments can be submitted on Gradescope up to **Friday, October 16, at 5:00pm**. This is also when the solutions will be released. Note that the latest time to submit HW3 late is **earlier than usual**. This is a trade-off between giving you the flexibility to submit late, and being able to release the HW3 grades and solution before the prelim.
- You can do this assignment in groups of 1-2. 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.

Problem 1: Hard-Margin SVMs

(10 + 10 + 14 = 34 points)

In parts (a) and (b) 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 neatly by hand) that achieves the largest hard margin on S without using an SVM package. **Give us the co-ordinates of at least 2 points that this hyperplane passes through.** Clearly mark all support vectors in your diagram.

(b) Now, compute by hand the weight vector $\vec{w}_{opt}$ and bias b_{opt} of the maximum margin hyperplane of the SVM in (a). Normalize your solution so that it satisfies $\|\vec{w}_{opt}\|_2 = 1$. 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 are given a new binary classification problem and a training sample of size 7500. The examples have 100,000 binary features (that take values $\{0, 1\}$), and you are wondering how well a homogeneous hard-margin SVM will work on this problem (i.e. how low the expected generalization error will be).

Here is what you know about the problem. Not all of the features are relevant, but at least the feature vectors $\vec{x}$ are quite sparse. Furthermore, you know that some of those 100,000 binary features are very reliable in predicting the class label – you just don't know which ones. In particular, you know that there is a set P of 10 features of which every positive example has at least 5 ones, and every negative example has at most 3 ones. Similarly, you know there is a set N of 10 features of which every positive example has at most 3 ones, and every negative example has at least 5 ones. The set Q contains the remaining 99,980 features (which do not tell you anything about the class label), but only at most 20 of those are ones for any example. Will the SVM learn a rule with good generalization error for this problem, even though we don't know which features are in P, which ones are in N, and which ones are in Q?

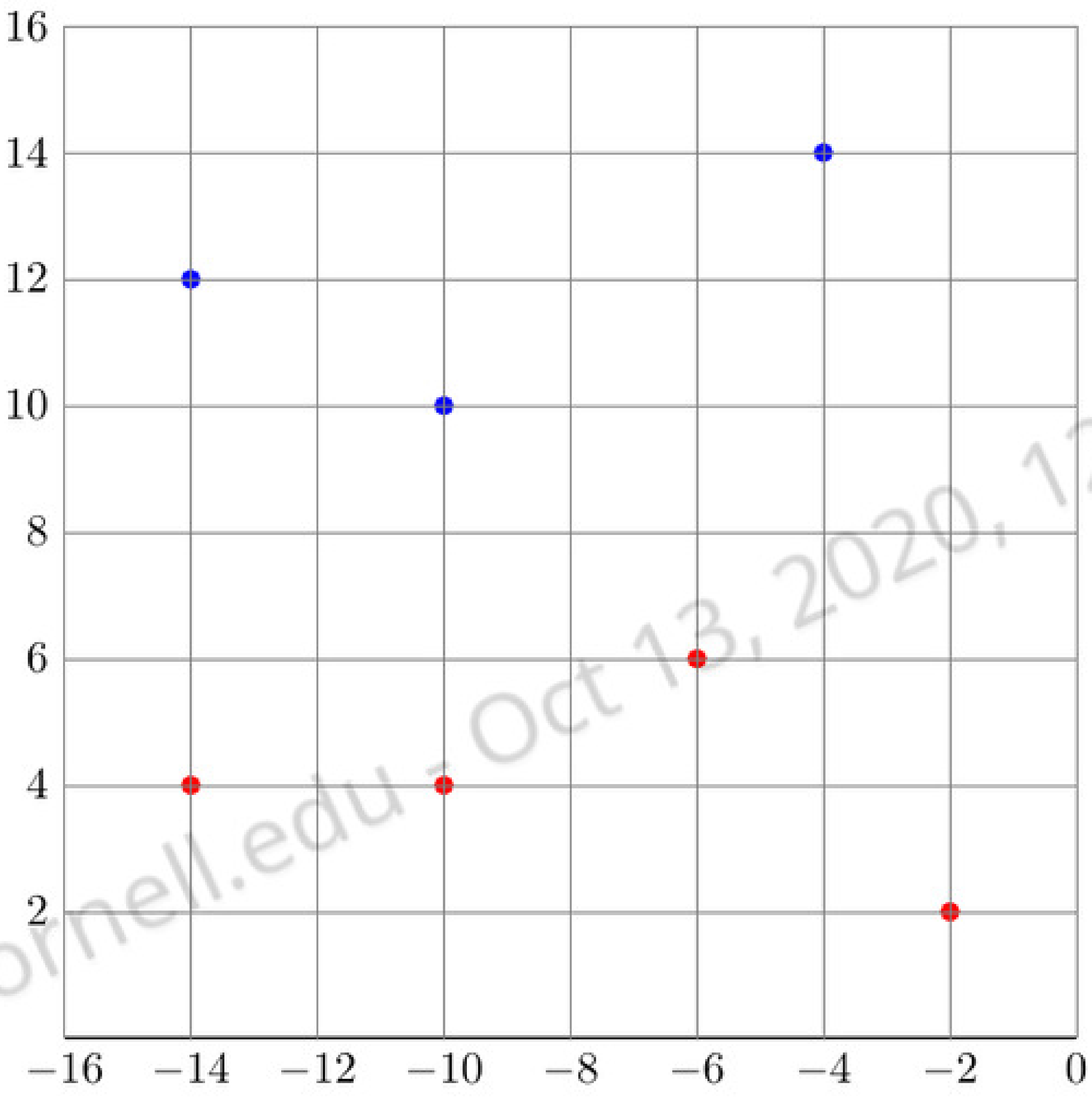

Figure 1: Toy dataset.

For the following questions, please show your work. One or two lines of derivation is sufficient. Please provide exact solutions instead of decimal approximations.

i) Calculate $\max_S R$, an upper bound on the length of the features vectors, that holds for any sample S. Your answer must be smaller than 10, and the tighter the bound, the better.

ii) Calculate $\min_S \gamma$, a lower bound on the geometric margin, that holds for any sample S. Your answer must be greater than 1/10, and the tighter the bound, the better. Hint: Construct a separating hyperplane and analyze its margin.

iii) Recall from the slides that the expected generalization error of a homogeneous hard-margin SVM trained on $|S'| = m - 1$ examples is upper bounded by the expected margin of a homogeneous hard-margin SVM trained on $|S| = m$ examples as follows:

$$E_{S'}[err_P(SVM(S'))] \leq \frac{1}{m} E_S \left[\frac{R^2}{\gamma^2}\right]$$

Instead of calculating the expected value in the RHS of the inequality, we will calculate an upper bound on the RHS. Write an expression that upper bounds $\frac{1}{m} E_S \left[\frac{R^2}{\gamma^2}\right]$ in terms of $\max_S R, \min_S \gamma$ and calculate the final value by substituting in the quantities you calculated in parts i) and ii). How large is your bound on the expected generalization error?

Problem 2: Kernels (7 + 8 + 9 + 8 = 32 points)

Consider one-dimensional datasets of the form $D_n^4 = \{x_i, y_i\}_{i=1...n}$ where n is a natural number, x_i is an integer, and $y_i \in \{-1, +1\}$. The data are generated as follows:

$$y_i = \begin{cases} +1, & x_i \mod 4 < 2 \\ \text{-1}, & \text{else} \end{cases}$$

(a) Show that any such dataset D_n^4 is linearly separable in the feature space

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

where $A = \frac{\pi}{2}$, by plotting the data in the feature space and (visually) identifying a linear separator. See Overleaf template for plotting assistance.

(b) What is the kernel $\kappa(x, x')$ of this feature space? Simplify so your final answer is a single term. You may refer to this link [1] 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}^4 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 which examples that lie on the margin.

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

[1] https://www2.clarku.edu/faculty/djoyce/trig/identities.html

Problem 3: More on SVMs (17 + 17 = 34 points)

In this problem we will investigate the training of linear classifiers on n training examples, $S = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}$ where each $\vec{x}_i \in \mathbb{R}^N$ and each $y_i \in \{-1, 1\}$.

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

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

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

i	y_i	$\vec{w}^\top \vec{x}_i + b$	α_i
1	1	0.4	0.1
2	-1	0.2	0.1
3	1	0.1	0.1
4	1	1	0

Table 1: Problem 3a

Recall the definition of leave-one-out error from lecture:

$$\mathrm{err}_{\mathrm{loo}}(A) = \frac{1}{m} \sum_{i=1}^{m} \Delta\left(h_i\left(\vec{x}_i\right), y_i\right)$$

Which of the instances in the above table is guaranteed to not be a leave-one-out error? Explain your solution, and include calculations for $2\alpha R^2$ and ξ_i.

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

- The length of all feature vectors $\vec{x}_i$ is $\sqrt{3}$, i.e. $\forall i : \vec{x}_i^\top \vec{x}_i = 3$
- Any two feature vectors in our training set are orthogonal, i.e. $\forall i \neq j : \vec{x}_i^\top \vec{x}_j = 0$.

Suppose we train a homogeneous hard margin SVM on S'. For any training example $(\vec{x}_i, y_i)$, what is the value of 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 $\forall i = 1, \ldots n, \quad \alpha_i \geq 0$. Because we train a homogeneous linear classifier we do not need the equality constraint.