

## Problem Set 3

Gen Nishida

*Handed In: November 6, 2014*

# 1 Questions

1. Fitting an SVM classifier by hand (source: Machine Learning, a probabilistic perspective. K Murphy)

- (a) Write down a vector that is parallel to the optimal vector  $w$ .

The decision boundary is perpendicular to the vector  $\phi(x_2) - \phi(x_1)$ .

$$\phi(x_2) - \phi(x_1) = [1, 2, 2]^T - [1, 0, 0]^T = [0, 2, 2]^T$$

Thus, the vector that is parallel to the optimal vector  $w$  is  $[0, 1, 1]^T$ .

- (b) What is the value of the margin that is achieved by this  $w$ ?

The maximum margin is the half of the distance between two points in the 3d feature space. Thus,

$$\frac{\|\phi(x_2) - \phi(x_1)\|}{2} = \frac{\|[0, 2, 2]^T\|}{2} = \sqrt{2}$$

- (c) Solve for  $w$ , using the fact the margin is equal to  $1/\|w\|$ .

Let  $w = k[0, 1, 1]^T$ . Then,

$$\|w\| = k\|[0, 1, 1]^T\| = \sqrt{2}k$$

Since the margin is  $\sqrt{2}$ ,

$$\sqrt{2} = \frac{1}{\|w\|} = \frac{1}{\sqrt{2}k}$$

By solving this, we obtain  $k = 1/2$ . Thus,  $w = [0, 1/2, 1/2]^T$ .

- (d) Solve for  $w_0$  using your value for  $w$  and Equations 1 to 3.

By substituting  $w$  of Equations 2 and 3, we get

$$\begin{cases} y_1(w^T \phi(x_1) + w_0) = -([0, 1/2, 1/2]^T \cdot [1, 0, 0]^T + w_0) = -w_0 \geq 1 \\ y_2(w^T \phi(x_2) + w_0) = ([0, 1/2, 1/2]^T \cdot [1, 2, 2]^T + w_0) = 2 + w_0 \geq 1 \end{cases}$$

By solving this, we obtain

$$-1 \leq w_0 \leq -1$$

Thus,  $w_0 = -1$ .

- (e) Write down the form of the discriminant function  $f(x) = w_0 + w^T \phi(x)$  as an explicit function of  $x$ .

$$f(x) = w_0 + w^T \phi(x) = -1 + [0, 1/2, 1/2]^T \cdot [1, \sqrt{2}x, x^2]^T = \frac{1}{2}x^2 + \frac{1}{\sqrt{2}}x - 1$$

2. We define a concept space  $C$  that consists of the union of  $k$  disjoint intervals in a real line. A concept in  $C$  is represented therefore using  $2k$  parameters:  $a_1 \leq b_1 \leq a_2 \leq b_2 \leq \dots \leq a_k \leq b_k$ . An example (a real number) is classified as positive by such concept iff it lies in one of the intervals. Give the VC dimension of  $H$  (and prove its correctness).

The answer is  $2k$ .

**Proof:**

Let  $VC(k)$  be the VC dimension for  $k$  disjoint intervals in a real line. I prove by induction that  $VC(k) = 2k$  in the following. When  $k = 1$ , two examples have four patterns in total, and all the cases can be correctly classified (Figure 1). Thus,  $VC(1) = 2$ , which satisfies the above hypothesis.

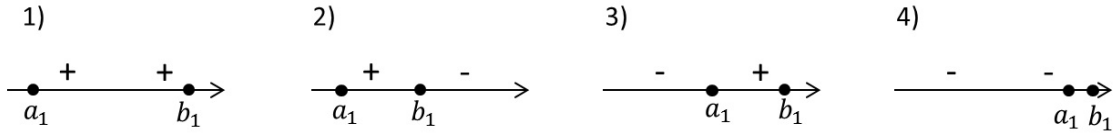


Figure 1: For the case of  $k = 1$ , two examples are correctly classified in all the four cases. Thus,  $VC(1) = 2$ .

Now, given that  $VC(k-1) = 2(k-1)$ , we want to show that two additional examples can be correctly classified by an additional interval. Assume without loss of generality that two additional numbers are greater than the existing  $2(k-1)$  ones that are already classified. Then, there are only four cases in terms of the labels of two additional examples, and they are correctly classified by  $k$ -th interval in all cases (Figure 2).

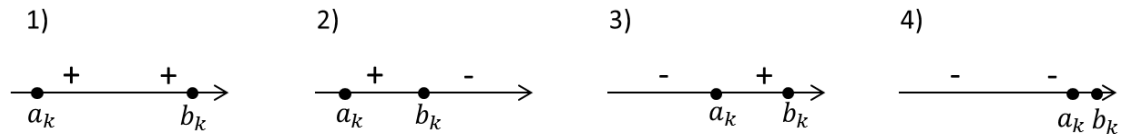


Figure 2: When adding  $k$ -th interval, two additional examples are correctly classified in all the four cases.

Thus,

$$VC(k) \geq VC(k-1) + 2 = 2(k-1) + 2 = 2k$$

Therefore,  $VC(k)$  is at least  $2k$  by induction. Also, Figure 3 shows a case in which  $k+1$  positive examples and  $k$  negative examples cannot be correctly classified. This concludes  $VC(k) = 2k$ .

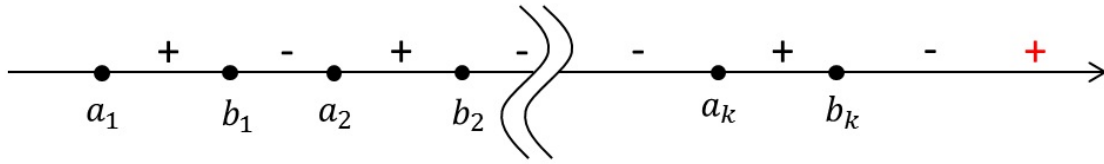


Figure 3: This figure shows a case in which  $k + 1$  positive examples and  $k$  negative examples cannot be correctly classified by  $k$  disjoint intervals. The right most positive example (red color) is classified incorrectly. Thus,  $VC(k) < 2k + 1$ .

### 3. The Gradient Descent (GD) algorithm

- (a) Write in one sentence: what are the hyper parameters of the GD algorithm.

The hyper parameters are the initial weight vector which you can randomly initialize if you want, learning rate that is the step size of updating the weight vector, a parameter  $\lambda$  that defines the impact of the regularizer, and the convergence criteria when to stop the iteration such as the maximum number of iterations.

- (b) Write in one sentence: What is the difference between  $l1$  and  $l2$  regularization.

$l1$  regularization uses  $l1$  norm of  $w$  as the regularization term, which encourages the sparsity, while  $l2$  regularization uses  $l2$  norm of  $w$ .

- (c) Write down the gradient descent algorithm applied to hinge loss with  $l2$  regularization.

The algorithm is shown in Algorithm 1. Note that the bias term is included in the weight vector by extending the feature vector as  $[x, 1]^T$  and the weight vector as  $[w, b]^T$ .

---

**Algorithm 1** Gradient descent algorithm applied to hinge loss with  $l2$  regularization

---

```

1: procedure HINGEREGULARIZEDGD()
2:   Initialize  $w$  randomly
3:   for  $i = 0$  to  $maxIterations$  do
4:      $\Delta w = (0, \dots, 0)$ 
5:     for all training data  $x_d$  for  $d = 1, \dots, D$  do
6:       if  $y_d w \cdot x_d \leq 1$  then
7:          $\Delta w = \Delta w + y_d x_d$ 
8:      $\Delta w = \Delta w - \lambda w$ 
9:      $w = w + \eta \Delta w$ 
10:  return  $w$  when it has converged

```

---

## 2 Programming Assignment