# ECE 449 Machine Learning (21Fa): Assignment 2

Xu Ke 3190110360

Nov. 10, 2021

## 1 Problem 1

### 1.1 Draw on Figure 1 the decision boundary obtained by the linear hard margin SVM method with a thick solid line. Draw the margins on either side with thinner dashed lines. Circle the support vectors.

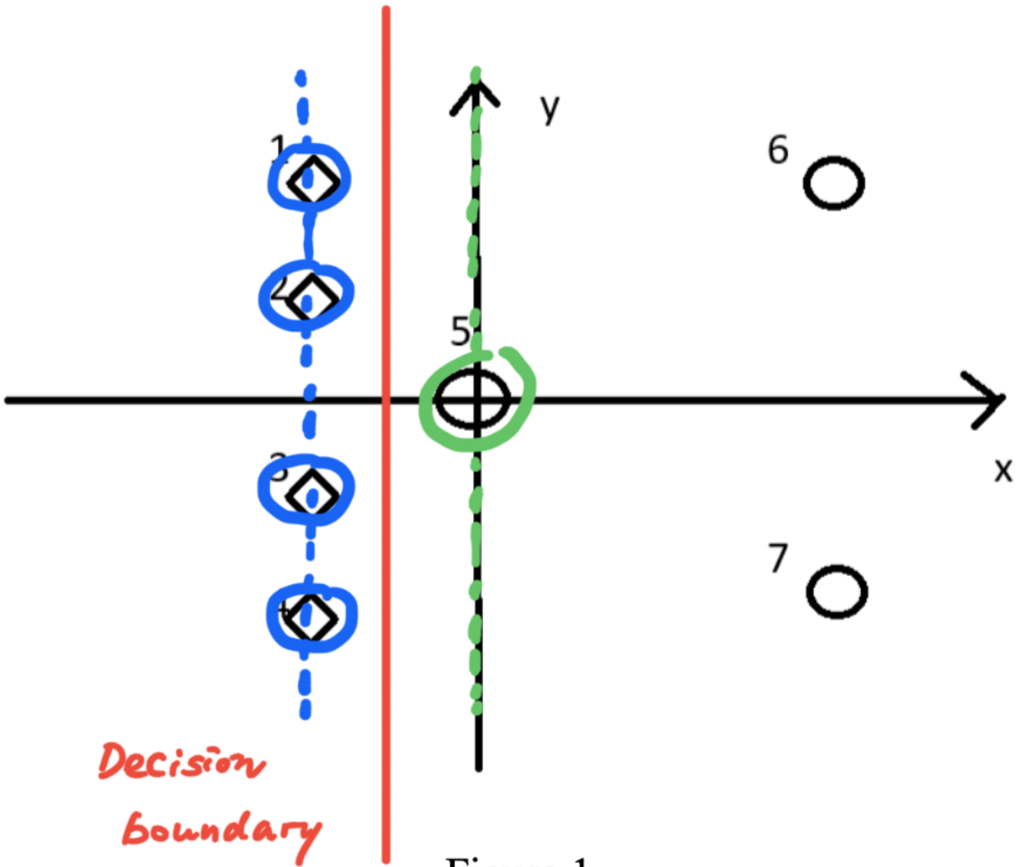
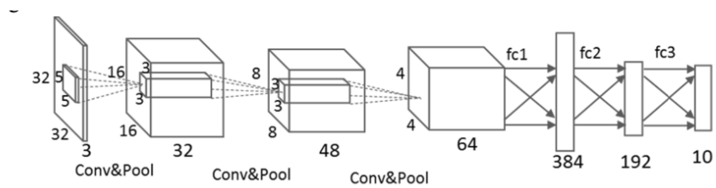



Figure 1

### 1.2 The removal of which sample will change the decision boundary?

The removal of sample 5 will change the decision boundary. Since in this case, the decision boundary is based on sample 1-4 and sample 5 and remove any one of sample 1-4 will not make any difference due to the invariance of the dashed line but the removal of 5 will shift the boundary rightward for a large scale.

# 2 Problem 2




## 2.1 Please calculate the total number of parameters (including all weights and all biases, suppose biases are used wherever possible) of this CNN. [6 pts]

Let's calculate the number of parameters from the left to the right one by one:
For the first $Conv\&Pool$, the number of parameters is (5*5*3+1)*32 = 2432;
For the second $Conv\&Pool$, (3*3*32+1)*48 = 13872;
For the third $Conv\&Pool$, (3*3*48+1)*64 = 27712;
For the first fc (fc1), (4*4*64+1)*384 = 393600;
For the second fc (fc2), (384+1)*192 = 73920;
For the thirf fc (fc3), (192+1)*10 = 1930
Thus, by adding them up we can get the total number of parameters is:

$$2432 + 13872 + 27712 + 39360 + 73920 + 1930 = 513466 \tag{1}$$

# 3 Problem 3

## 3.1 Explain the process of gradient descent algorithm. What is the difference between gradient descent algorithm and gradient ascent algorithm?

### 3.1.1 The process of gradient descent algorithm

Step 1: Calculate the loss function $J(\theta)$.
Step 2: Calculate the gradient of loss function $\frac{\partial}{\partial\theta}J(\theta)$
Step 3: Update $\theta$ using $\theta = \theta - \alpha \cdot \frac{\partial}{\partial\theta}J(\theta)$
Step 4: Repeat step 1-3 until convergence (i.e.reach minimum).

### 3.1.2 The difference between gradient descent algorithm and gradient ascent algorithm

Gradient descent algorithm is an iteration towards local minimum while gradient ascent algorithm is an iteration towards local maximum.

# 4 Problem 4

## 4.1 Large Learning Rate:

### 4.1.1 Advantage:

In the early stage of algorithm optimization, the learning process will be accelerated (i.e. converge fast), making it easier (i.e. fewer iterations) for the model to approach the local or global optimal solution.

### 4.1.2 Disadvantage:

In the later period, there will be large fluctuations (i.e. have risks in divergence), and even the value of the loss function may hover around the minimum value with great fluctuations but very hard to reach the optimal value.

## 4.2 Small Learning Rate:

### 4.2.1 Advantage:

To some extent, it can help avoid some over-corrections and proceed in a relatively smooth path, finally settling in a minimum and improving generalization accuracy.

### 4.2.2 Disadvantage:

The loss function will converge very slowly, increasing the time to find the optimal value, and it maybe just converges to the local extreme point, and thus there is no real optimal solution found.

# 5 Problem 5

## 5.1 What is the entropy of a fair four-sided die?

$$H(x) = -\Sigma_{i=1}^{4} P_i \cdot log_2 P_i = -4 \cdot \frac{1}{4} \cdot (-2) = 2 \tag{2}$$

# 6 Problem 6

$$\begin{aligned} H(X|Y) = \Sigma_y P_Y(y) \cdot [-\Sigma_x P_{X|Y}(x|Y) \cdot log(P_{X|Y}(x|Y))] \\ = \frac{1}{4} \cdot (-\frac{1}{2} \cdot log_2 \frac{1}{2} - \frac{1}{4} \cdot log_2 \frac{1}{4} - \frac{1}{8} \cdot log_2 \frac{1}{8} - \frac{1}{8} \cdot log_2 \frac{1}{8}) \\ + \frac{1}{4} \cdot (-4 \cdot \frac{1}{4} \cdot log_2 \frac{1}{4}) + \frac{3}{8} \cdot (-2 \cdot \frac{1}{3} \cdot log_2 \frac{1}{3} - 2 \cdot \frac{1}{6} \cdot log_2 \frac{1}{6}) \\ + \frac{1}{8} \cdot (\frac{1}{8} \cdot log_2 1) = \frac{15}{16} + \frac{1}{8} \cdot log_2 54 \end{aligned} \tag{3}$$

$$\begin{aligned} H(Y|X) = \Sigma_x P_X(x) \cdot [-\Sigma_y P_{Y|X}(y|X) \cdot log(P_{Y|X}(y|X))] \\ = \frac{7}{16} \cdot (-\frac{6}{7} \cdot log_2 \frac{2}{7} - \frac{1}{7} \cdot log_2 \frac{1}{7}) + \frac{1}{4} \cdot (-2 \cdot \frac{1}{4} \cdot log_2 \frac{1}{4} - \frac{1}{2} \cdot log_2 \frac{1}{2}) \\ + 2 \cdot \frac{5}{32} \cdot (-\frac{1}{5} \cdot log_2 \frac{1}{5} - 2 \cdot \frac{2}{5} \cdot log_2 \frac{2}{5}) = \frac{7}{16} \cdot log_2 7 + \frac{5}{16} \cdot log_2 5 - \frac{1}{4} \end{aligned} \tag{4}$$

$$\begin{aligned} I(X;Y) = H(Y) - H(Y|X) = -\Sigma_y P_Y(y) log_2(P_Y(y)) - H(Y|X) = -2 \cdot \frac{1}{4} \cdot log_2 \frac{1}{4} \\ - \frac{3}{8} \cdot log_2 \frac{3}{8} - \frac{1}{8} \cdot log_2 \frac{1}{8} - \frac{7}{16} log_2 7 - \frac{5}{16} log_2 5 + \frac{1}{4} \\ = 1 - \frac{3}{8} \cdot log_2 \frac{3}{8} - \frac{1}{8} \cdot log_2 \frac{1}{8} - \frac{7}{16} log_2 7 - \frac{5}{16} log_2 5 + \frac{1}{4} \end{aligned} \tag{5}$$