

# HOMWORK 2:

## WRITTEN EXERCISE PART

### 1 Information Theory [25/4 pts]

Suppose  $X, Y$  are two random variables taking values in a discrete finite set  $V$ . Let  $H(Y)$  denote the entropy of  $Y$ , and let  $H(Y|X)$  denote the conditional entropy of  $Y$  conditioned on  $X$ . Prove that if  $X, Y$  are independent, then  $H(Y) = H(Y|X)$ .

$$H(Y | X) = \frac{H(Y \cap X)}{H(X)} = \frac{H(Y)H(X)}{H(X)} = H(Y)$$

### 2 Standardizing Numeric Features [25/4 pts]

Standardize the data set with four points in 2 dimension:  $(7, 7), (3, 7), (3, 3), (7, 3)$ .

$(0.866, 0.866), (-0.866, 0.866), (-0.866, -0.866), (0.866, -0.866)$

Mean = (5,5) Standard Deviation = (2.309401 2.309401)

### 3 $k$ -Nearest Neighbors [25/4 pts]

Consider the training data set  $x_1 = (7, 7), y_1 = 0; x_2 = (3, 7), y_2 = 1; x_3 = (3, 3), y_3 = 1; x_4 = (7, 3), y_4 = 2$ . Suppose the Manhattan distance is used. What is the label for  $x = (0, 0)$  in the following settings? Show the calculation steps.

- 1-nearest neighbors.
- 3-nearest neighbors.
- 3-nearest neighbors, distance weighted. The weight for the  $i$ -th neighbor  $z$  is  $1/d(x, z)^2$ .

1.

$$x_1 - x = (7 - 0, 7 - 0) = (7, 7) = 14$$

$$x_2 - x = (3 - 0, 7 - 0) = (3, 7) = 10$$

$$x_3 - x = (3 - 0, 3 - 0) = (3, 3) = 6$$

$$x_4 - x = (7 - 0, 3 - 0) = (7, 3) = 10$$

Therefore, the label for  $x$  will given by  $y_3, y = 1$

2.

We already know that the three closest points to  $(0,0)$  are  $x_2, x_3$  and  $x_4$ . The labels of these points is 1,1 and 2. The most frequently occurring is 1. Therefore the label of  $x$  will be 1

3.

Assuming distance weightage, we have  $x_1 = 0/14, x_2 = 1/10, x_3 = 1/6$  and  $x_4 = 2/10$ . The maximal of these would be  $x_4$ , hence the label would be  $y_4 = 2$ .

### 4 Performance Measurements [25/4 pts]

Consider the following confusion matrix for 2 classes.

	actual positive	actual negative
predict positive	76	18
predict negative	24	82

Compute the accuracy, error, true positive rate, false positive rate, precision, and recall.

$$\text{TPR} = \text{recall} = 76/100$$

$$\text{FPR} = 18/100$$

$$\text{Precision} = 76/94$$

$$\text{Error} = 24/100$$

$$\text{Accuracy} = 76+82/(76 + 18 + 24 + 82) = 158/200$$