

# Machine Learning

## Homework 1

Due on April 8, 2020

1. Let

$$A_+ = \{(0, 0), (1, 1), (-1, 1), (1, -1), (-1, -1)\}$$

and

$$A_- = \{(1, 0), (-1, 0), (0, 1), (0, -1)\}$$

represent the *positive* and *negative* training instances respectively.

- (a) Plot the decision boundary for the 1-nearest neighbor algorithm.
- (b) What is the training set accuracy for 3-nearest neighbor algorithm?
- (c) What is the *confusion* matrix for 3-nearest neighbor algorithm on the training set?

2. Let the  $A \in R^{n \times n}$  be a symmetric positive definite matrix. Show that all eigenvalues of matrix  $A$  are positive.
3. Let  $Z = [X_1; X_2; X_3]$  be a random vector and  $\Sigma$  be a matrix with size  $3 \times 3$  where  $\Sigma_{ij} = Cov(X_i, X_j)$  and  $\Sigma_{ii} = Var(X_i)$ . Let the random variable  $W = \mathbf{a}^\top Z = a_1 X_1 + a_2 X_2 + a_3 X_3$  where  $\mathbf{a} = [a_1; a_2; a_3]$ . *i.e.*, the random variable  $W$  is the *projection* of random vector  $Z$  onto the vector  $\mathbf{a}$ . Find the variance of  $W$ .

4. Let  $S$  be a set of 10,000 random numbers generated by the *uniform distribution*,  $U[0; 1]$ . You have to estimate the *mean* of this *uniform distribution*.
  - (a) Randomly select 10 number from the set  $S$  and then use the average of these 10 number as the estimation. Repeat this experiment 20 times. What is the *sample mean* and *sample standard deviation* of these 20 random experiments?
  - (b) Randomly select 1,000 number from the set  $S$  and then use the average of these 1,000 number as the estimation. Repeat this experiment 50 times. What is the *sample mean* and *sample standard deviation* of these 50 random experiments?

5. Consider a linear system of equations as follows:

$$\begin{array}{rclcl}
 2x_1 & + & 2x_2 & & = & 4 \\
 x_1 & + & & x_3 & = & 2 \\
 & & 2x_2 & + & 2x_3 & = & 3 \\
 x_1 & + & & & x_3 & = & 4
 \end{array}$$

Find the least squares approximation solution for it.

6. Generate a training dataset with size 1000 by yourself.
  - (a) That is,  $S = \{(\mathbf{x}^i, y_i) \mid \mathbf{x}^i = (x_1^i, x_2^i) \in R^2, \text{ and } y_i \in R, i = 1, \dots, 1000\}$ , where  $x_1$  and  $x_2$  are generated by the *uniform distribution*,  $U[-1; 1]$ . The observation value  $y = 2x_1^2 + x_2^2 - 2x_1x_2 + 2x_1 - x_2 + \epsilon$  where  $\epsilon$  is the random noise generated by  $N(0, 1)$ .
  - (b) Find a *quadratic* function  $f(\mathbf{x})$ , that is fitted in the training dataset  $S$ .
  - (c) Compute the *MAE*, mean of absolute error, and plot the function you get.