

## Exercise 1.1 (Learning From Data — Abu-Mostafa et al.)

### Problem Rubric

Each task must be expressed within the *learning from data* framework by clearly identifying:

- a) the input space  $\mathcal{X}$ ,
- b) the output space  $\mathcal{Y}$ ,
- c) the (unknown) target function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ ,
- d) and the nature of the data set from which the function is learned.

### (a) Medical Diagnosis

- $\mathcal{X}$ : Medical history of a patient, including symptoms, test results, and prior conditions.
- $\mathcal{Y}$ : Set of possible medical problems or diagnoses.
- $f$ : A function mapping medical histories to the corresponding diagnosis.
- Data set: Patient medical records labeled with confirmed diagnoses.

### (b) Handwritten Digit Recognition

- $\mathcal{X}$ : Pixel intensity values of handwritten digit images.
- $\mathcal{Y}$ : Digits  $\{0, 1, \dots, 9\}$ .
- $f$ : A function mapping pixel values to the digit represented in the image.
- Data set: A collection of labeled handwritten digit images.

### (c) Spam Detection

- $\mathcal{X}$ : Words or extracted features from email messages.
- $\mathcal{Y}$ : {spam, not spam}.
- $f$ : A binary classification function determining whether an email is spam.
- Data set: Emails labeled as spam or not spam.

### (d) Electric Load Prediction

- $\mathcal{X}$ : A vector consisting of electricity price, temperature, and day of the week.
- $\mathcal{Y}$ : Electric load demand.
- $f$ : A function predicting electric load from the input variables.
- Data set: Historical observations of electric load paired with price, temperature, and calendar information.

**(e) Example of a Personal Interest Problem**

- $\mathcal{X}$ : User interaction data with an online form system (e.g., time spent, number of edits).
- $\mathcal{Y}$ : Probability of successful form submission.
- $f$ : An unknown function inferred empirically from data.
- Data set: Logged interaction records with observed submission outcomes.