1 Learning Outcomes

- Distinguish between tabular, semi-structured, and unstructured data
- Describe how features relate to measurements of objects' properties
- Explain how features for objects are organized into a matrix (One row per object, One column per feature)
- Give an example of representing a data set as a feature matrix
- Explain what a label vector contains and how to interpret it
- Explain what reshaping an array means, why you would want to reshape an array, and how reshaping is accomplished

2 Structured vs Unstructured Data

Data exists in different formats and sizes. It can be structured, semi-structured, or unstructured. Each one of this data is collected in different ways, and stored in a different type of database.

Structured Data
Structured data refers to the data that is highly organized and that is stored in a pre-defined tabular format: information is organized into rows and columns, where each row represents a record (an example, a sample or an observation), each column represents a field (feature) and entries of the same column has the same type. Structured or tabular data is commonly stored in relational databases or spreadsheets, or as a comma-separated values (CSV) file. Figure 1 shows an example of tabular data: each row represents one order and each column represents an information about an order i.e. country, salesperson, units and amount. A single row has all the information of a particular order.

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amoun
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00

Figure 1: An example of tabular data (Source).

Semi-structured Data
Semi-structured data comes in a form that does not follow a strict schema like tabular data in databases, but it still has some structure. Figure 2 shows a JSON file which is an example of a

file that stores semi-structured data. A JSON file is a text file that can store nested information and can represent more complex data. As shown in figure 2, the data is stored in key-value pairs enclosed in curly brackets. The value of a key can be a set of key-value pairs.


```
{
    "orders": [
        {
            "orderno": "748745375",
            "date": "June 30, 2088 1:54:23 AM",
            "trackingno": "TN0039291",
            "custid": "11045",
            "customer": [
                {
                    "custid": "11045",
                    "fname": "Sue",
                    "lname": "Hatfield",
                    "address": "1409 Silver Street",
                    "city": "Ashland",
                    "state": "NE",
                    "zip": "68003"
                }
            ]
        }
    ]
}
```


Figure 2: An example of JSON file (Source).

Unstructured Data

Unstructured data is the data that does not come with a pre-defined format, and it is stored in its raw native format. Examples of unstructured data include text (like social media posts, product reviews), videos, images, audios, sensor data, etc.

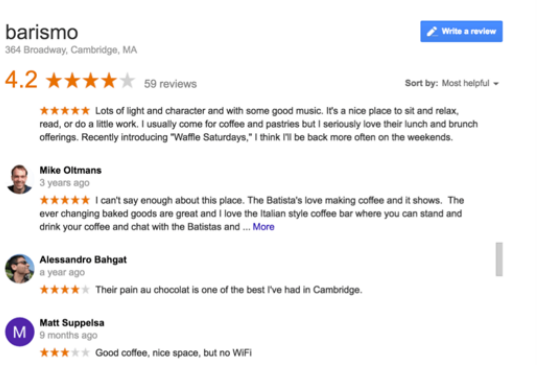

Figure 3: Textual data is an example of unstructured data (Source).

3 Data Format for Machine Learning

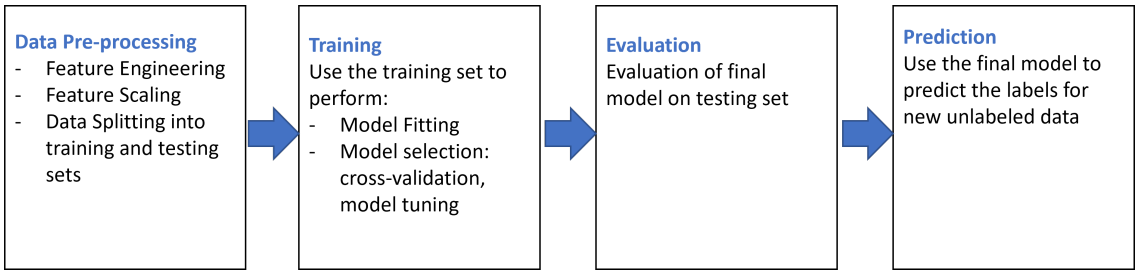

Figure 4: Steps of building a machine learning system

Structured or tabular data is the data format that is closest to the one that is expected by traditional machine learning models. Classic machine learning models expect data to be a set of samples, where each sample represents an object for which we want to predict a label and each sample should consist of a set of features or measurements. Many of the columns of tabular data can be directly used as features for machine learning models and thus structured data requires less processing than semi-structured or unstructured data. Working with unstructured data requires features to be first engineered and extracted from the raw data, and then organized into samples of features. This step is known as feature engineering and it is part of the data pre-processing step shown in figure 5.

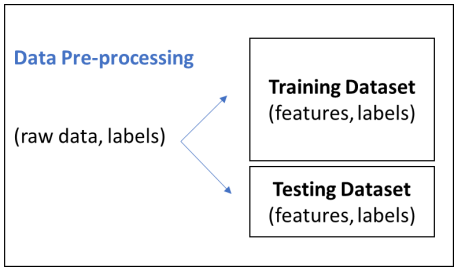

Figure 5: Data pre-processing requires extracting features from raw data

Feature engineering may also be applied to structured data. For example, traditional learning models expect all features to be numerical, however some of the fields in tabular data can be non-numerical (categorical: color, vehicle type; boolean: married, automatic engine), which need to be transformed into numerical features. Feature engineering might also be needed in order to create new features and enrich the data with them. In this course, when discussing any machine learning model, we will not discuss feature engineering deeply; most of the time, we will assume that we have some numerical features that are of interest for us and for we which we want to build a machine learning model.

Let's now discuss the exact data format for machine learning models using vector and matrix notation. In particular, we introduce feature vector, feature matrix and label vector.

3.1 Feature Vector

Given a set of m features, we define the feature vector $\mathbf{x} \in \mathbb{R}^m$ as the vector that collects these features:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

We use the subscript i to denote the i^{th} feature (x_i) of $\mathbf{x}$. These features represent the measurements or properties of an object for which we want to predict a label. The vector of features can be used to represent a single sample or example in a training data set or a testing set or a future data set (naive) for which we want to predict the unknown labels. Moreover, a machine learning model can be defined as the mathematical equation or formula $f(.)$ that maps a vector of features $\mathbf{x}$ to its label y:

$$\mathbf{x} \xrightarrow{\text{ML model f(.)}} y, \text{ where } y = f(\mathbf{x})$$

The goal of the training phase is to find this mathematical mapping $f(.)$, which can be used to predict the label y for a given vector of features $\mathbf{x}$. The space of all possible feature vectors $\mathbb{R}^m$ is called ***feature space***.

3.2 Feature Matrix

A training, testing or future/naive data set consists of a collection of feature vectors. For each set, we can collect these feature vectors into a matrix that we call feature matrix. Collecting feature vectors into feature matrix helps create a more compact notation and apply vectorized operations. Assume a data set consists of n feature vectors. We use the superscript j to denote the j^{th} feature vector in the data set: $\mathbf{x}^{(j)}$, i.e, the feature vector of the j^{th} sample. The feature matrix X is obtained by first transposing each feature vector into a row vector and then vertically stacking them, as in:

$$X = \begin{bmatrix} \mathbf{x}^{T(1)} \\ \mathbf{x}^{T(2)} \\ \vdots \\ \mathbf{x}^{T(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_m^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \dots & x_m^{(n)} \end{bmatrix}$$

Therefore, the feature matrix $X \in \mathbb{R}^{n \times m}$ consist of n rows and m columns: each row represents a single sample or example of the data set, and each column represents a measurement or feature.

3.3 Label Vector

In supervised learning, there is a label that corresponds to each sample in the training and testing sets. If a given set consists of n data samples, then we must have n labels. Let $y^{(j)}$ be the label

that corresponds to the j^{th} sample, then the label vector is given by:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

where $\mathbf{y} \in \mathbb{R}^n$.

3.4 Putting All together

During the training phase, a training algorithm expects feature matrix and its corresponding label vector that corresponds to the training data set, and uses them in order to fit a machine learning model (i.e., to find specific properties or characteristics of the model) as shown in figure 6.

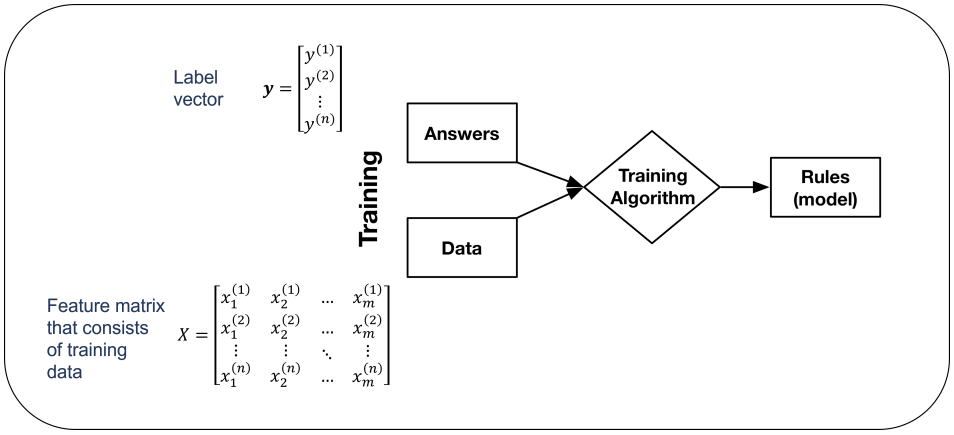

Figure 6: Model fitting expects a feature matrix and its corresponding label vector.

Once a final model is chosen, it is then used on future or naive data to find its expected label as shown in figure 7.

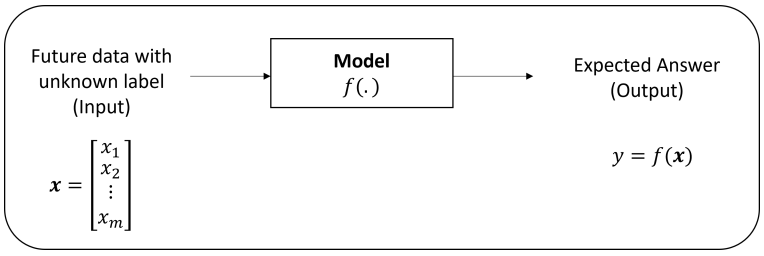

Figure 7: A model is mathematical formula that maps a feature vector to its expected label.

3.5 Examples

The Iris dataset - *The dataset consists of measurements in centimeters of the sepal length and width, and the petal length and width, respectively, for 150 flowers from three species (setosa,*

versicolor, and verginica) of iris plants.

What is the feature matrix? What is the vector label?

Each flower can be represented by vector of 4 features: sepal length, sepal width, petal length and petal width and there are 150 flowers. The feature matrix has then 4 columns and 150 rows, where each row contains the measurements of one flower. The vector of labels contains 150 entries where each entry represents the label of a flower, which could 0 (setosa), 1 (versicolor) or 2 (virginica).

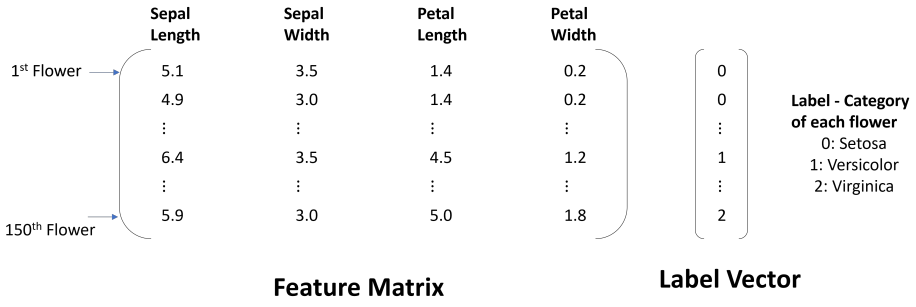

Figure 8: Feature matrix and label vector for the iris dataset

The Diabetes dataset *- The dataset consists of 10 baseline variables: age, sex, body mass index (BMI), average blood pressure (BP), and six blood serum measurements (S1, S2, S3, S4, S5, S6), for each of 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.*

What is the feature matrix? What is the vector label?

Each patient can be represented by vector of 10 features: age, sex, BMI, BP, S1, S2, S3, S4, S5, S6, and there are 442 patients. The feature matrix has then 10 columns and 442 rows, where each row corresponds to one patient. The vector of labels contains 442 entries where each entry represents the disease progression for each patient.

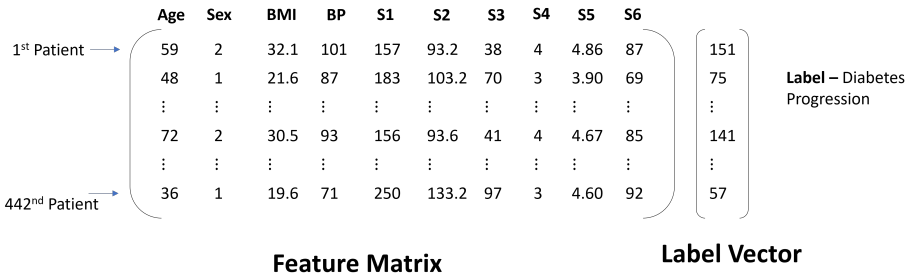

Figure 9: Feature matrix and label vector for the diabetes dataset

4 Summary of notations

Here's a summary of the mathematical notations introduced in these notes:

Symbol	Meaning
X	Feature Matrix
$\mathbf{x}$	Feature vector
$\mathbf{x}^{(j)}$	The j^{th} feature vector that corresponds to the j^{th} data sample
$x_i^{(j)}$	The i^{th} entry of the j^{th} feature vector
$\mathbf{y}$	Label vector
$y^{(j)}$	The label that corresponds to the j^{th} data sample