INTRODUCTION TO MACHINE LEARNING

DATA REPRESENTATION
Data Representation?

- How does a computer represent data?
  - 0 and 1 in the aspect of “general” computer science
  - Vector/Matrix in the aspect of “Machine Learning”
## Data Representation

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalar</td>
<td>single number</td>
<td>- usually write in italics  &lt;br&gt; - lower-case variable names  &lt;br&gt; - e.g., $s \in \mathbb{R}, n \in \mathbb{N}$ [1]</td>
</tr>
<tr>
<td>Vector</td>
<td>array of numbers</td>
<td>- arranged in order  &lt;br&gt; - lower-case names written in bold typeface  &lt;br&gt; - $\mathbf{x} = \begin{bmatrix} x_1 \ x_2 \ \vdots \ x_n \end{bmatrix}$, $\mathbf{x} = {x_1, x_2, \ldots, x_n}$  &lt;br&gt; - what is $\mathbf{x}<em>s$ when $s = {1, 3, 6}$?  &lt;br&gt; - Then, $\mathbf{x}</em>{-s}$?</td>
</tr>
</tbody>
</table>

---

## Data Representation

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
</table>
| **Matrix** | 2-D array of numbers | - an element is identified by two indices  
- upper-case variable name with bold typeface, e.g., $\mathbf{X}$  
- $\mathbf{X} \in \mathbb{R}^{m \times n}$: matrix has a height of $m$ and a width of $n$, and elements are real numbers  
- e.g., $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix}$ |
| **Tensor** | array with more than two axes | - three indices to identify an element |
Types of Variable

- **Categorical variable:** discrete or qualitative variables
  - Nominal:
    - Have two or more categories, but which do not have an intrinsic order
  - Ordinal
    - Have two or more categories, which can be ordered or ranked.
- **Continuous variable**
Data Representation

- **Features**
  - An individual measurable property of a phenomenon being observed
  - The number of features or distinct traits that can be used to describe each item in a quantitative manner
  - May have implicit/explicit patterns to describe a phenomenon

Reference: https://en.wikipedia.org/wiki/Feature_(machine_learning)
Feature Examples:

- **Speech recognition**
  - noise ratios, length of sounds, relative power

- **Spam detection**
  - presence or absence of certain email headers, email structure, frequency of specific terms

- **Image recognition**
  - Edges, curves, ..
Data Representation

- **Samples**
  - Items to process (classify or cluster)
  - Can be a document, a picture, a sound, a video, or a patient
  - Features are characteristics of a sample

Reference: https://en.wikipedia.org/wiki/Feature_(machine_learning)
Data Representation

- Feature vector
  - An N-dimensional vector of numerical features that represent some objects
  - A sample consists of feature vectors

- Feature extraction (feature selection)
  - Preparation of feature vector
  - Transforms the data in the high-dimensional space to a space of fewer dimensions

Reference: http://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937
Example - Survey

- Convert Data to a feature vector/sample matrix

\[
\begin{bmatrix}
time = agree \\
audio = yes \\
\vdots
\end{bmatrix}
\]
Example – Structured data

- Convert Data to a feature vector/sample matrix

<table>
<thead>
<tr>
<th>Name</th>
<th>Finance</th>
<th>Marketing</th>
<th>Statistics</th>
<th>Strategy</th>
<th>Operations</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aparna</td>
<td>B</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td>1.4</td>
</tr>
<tr>
<td>Bikas</td>
<td>D</td>
<td>D</td>
<td>F</td>
<td>F</td>
<td></td>
<td>2.4</td>
</tr>
<tr>
<td>Chandra</td>
<td>D</td>
<td>A</td>
<td>F</td>
<td>F</td>
<td>D</td>
<td>2.4</td>
</tr>
<tr>
<td>Deepak</td>
<td>A</td>
<td>B</td>
<td>D</td>
<td>D</td>
<td></td>
<td>3.2</td>
</tr>
<tr>
<td>Fazal</td>
<td>D</td>
<td>B</td>
<td>B</td>
<td>D</td>
<td></td>
<td>2.4</td>
</tr>
<tr>
<td>Gowri</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>B</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>Hari</td>
<td>B</td>
<td>A</td>
<td></td>
<td>D</td>
<td></td>
<td>2.8</td>
</tr>
<tr>
<td>Ismet</td>
<td>B</td>
<td>B</td>
<td></td>
<td>A</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>Jagdeep</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>Kunal</td>
<td>F</td>
<td>A</td>
<td>F</td>
<td>F</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>Leena</td>
<td>B</td>
<td>A</td>
<td></td>
<td>F</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>Manab</td>
<td>A</td>
<td>B</td>
<td></td>
<td>B</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>Nisha</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>A</td>
<td>F</td>
<td>3.6</td>
</tr>
<tr>
<td>Osman</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td></td>
<td>4.6</td>
</tr>
<tr>
<td>Preeti</td>
<td>F</td>
<td>D</td>
<td></td>
<td>D</td>
<td></td>
<td>3.2</td>
</tr>
<tr>
<td>Rahul</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>F</td>
<td></td>
<td>4.2</td>
</tr>
<tr>
<td>Sameer</td>
<td>C</td>
<td>F</td>
<td></td>
<td>F</td>
<td></td>
<td>2.4</td>
</tr>
<tr>
<td>Tara</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.4</td>
</tr>
<tr>
<td>Utkarsh</td>
<td>F</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td>3.0</td>
</tr>
<tr>
<td>Vipul</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>F</td>
<td></td>
<td>2.4</td>
</tr>
</tbody>
</table>

\[
\begin{bmatrix}
\text{Finance} \\
\text{Marketing}
\end{bmatrix}
\]
Example – Image data

Ref: https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html
Example – Unstructured data

Unstructured data (e.g., text data)

Feature Extraction

Structured data (e.g., Bag-of-Words Model)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>love</th>
<th>dogs</th>
<th>hate</th>
<th>and</th>
<th>knitting</th>
<th>is</th>
<th>my</th>
<th>hobby</th>
<th>passion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 2</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Doc 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Data in Machine Learning

- $x_i$: input vector, independent variable

\[
x_i = \begin{bmatrix}
x_{i,1} \\
x_{i,2} \\
\vdots \\
x_{i,n}
\end{bmatrix}, \quad x_{i,j} \in \mathbb{R}
\]

- $y$: response variable, dependent variable
  - $y \in \{-1, 1\}$ or \{0, 1\}: binary classification
  - $y \in \mathbb{Z}$: multi-label classification
  - $y \in \mathbb{R}$: regression
  - Predict a label when having observed some new $x$
Data Visualization

- Vector space model
  - Data is a set of features, $d_i = \{f_1, f_2, \ldots, f_p\}$
  - All data can be represented by vector

Ref: https://www.slideshare.net/pkgosh/the-vector-space-model
Data Visualization

- Hand-written data (MNIST)
  - High-dimensional data
  - Can visualize data using Principle Component Analysis
Comparison

What would you like to show?

Relationship
- Scatter plot
  - Two variables
  - Scatter plot bubble size
  - Three variables

Distribution
- Bar chart horizontal
- Bar chart vertical
- Circular area chart
- Line chart
- Bar chart vertical
- Line chart

Composition
- Variable width chart
- Table or tables with embedded charts
- Bar chart horizontal
- Bar chart vertical
- Circular area chart
- Line chart
- Bar chart vertical
- Line chart


Scatter Plots

Show relationships between two variables

Line Plots

- Best choice when showing that a variable varies with another

Data source: nces.ed.gov/programs/digest/2013menu_tables.asp
Author: Randy Olson (@randal_olson)
Note: Some majors are missing because the historical data is not available for them

Histograms

- Showing a distribution of data points
Bar Plots

- Effective to visualize categorial data (<10)

---

Box Plots

- Visualizing distributions of variables with quartiles (25% and 75% of the data) and median

Violin Plots

Ref: https://mode.com/blog/python-data-visualization-libraries
Let’s look at Python codes for data representation and visualization