CMSC/LING/STAT 208: Machine Learning

Abhishek Chakraborty [Much of the content in these slides have been adapted from *ISLR2* by James et al. and *HOMLR* by Boehmke & Greenwell]

What is Machine Learning?

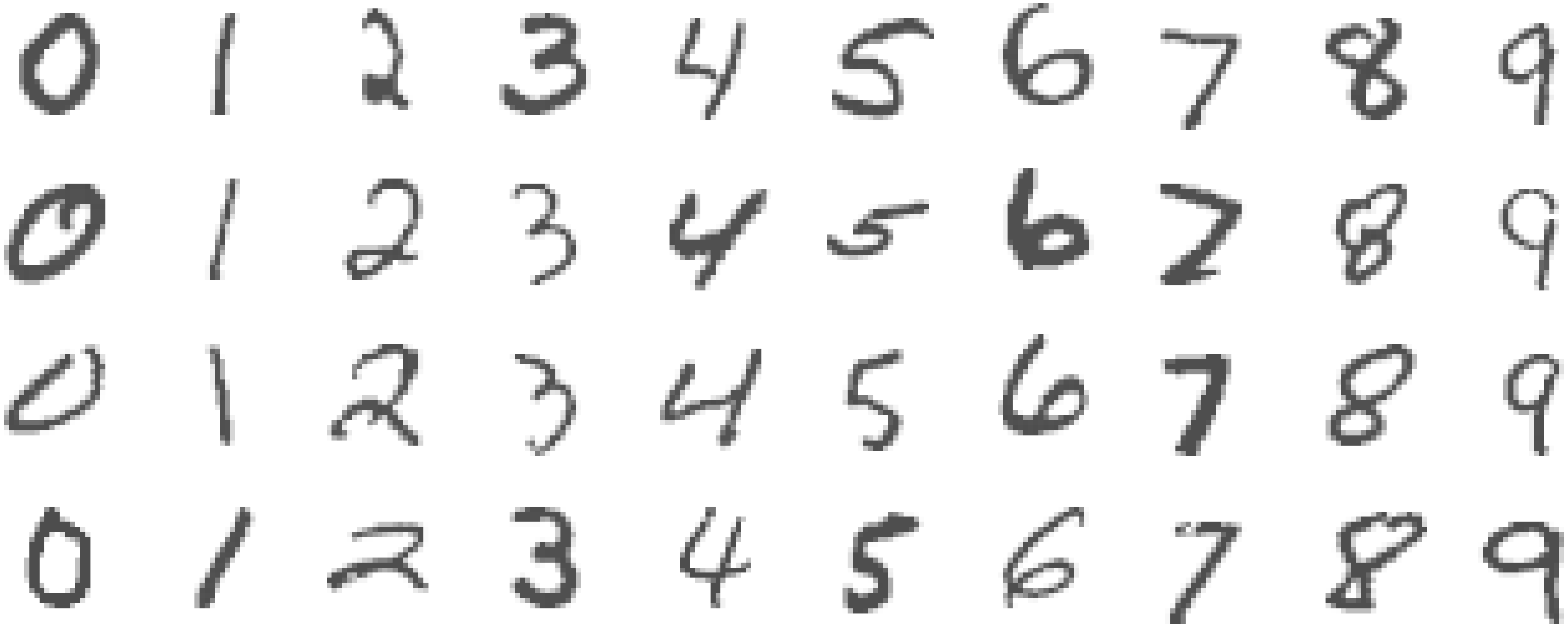
- Machine Learning is the study of tools/techniques for understanding complex datasets.
- The name machine learning was coined in 1959 by Arthur Samuel.
  - “Field of study that gives computers the ability to learn without being explicitly programmed.”

What is Machine Learning?

Tom M. Mitchell (1998) defined algorithms studied in the machine learning field as

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

What is Machine Learning?



MNIST handwritten digits (from ISLR, James et al.)

Question!!!

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. According to Tom Mitchell's definition, what is the task T, experience E, and performance measure P in this setting?

- The number (or fraction) of emails correctly classified as spam/ham.
- Classifying emails as spam or ham (not spam)
- Watching you label emails as spam or ham.

Statistical Learning vs Machine Learning vs Data Science

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap, a great deal of “cross-fertilization”.
- “Data Science” - Reflects the fact that both statistical and machine learning are about data.
- “Machine learning” or “Data Science” are “fancier” terms.

Terminologies/Notations

Ames Housing dataset - Contains data on 881 houses in Ames, IA. We are interested in predicting sale price.

The first ten observations are shown below.

```
## # A tibble: 10 × 6
##    Sale_Price Gr_Liv_Area Garage_Type Garage_Area Pool_Area Neighborhood
##         <int>       <int> <fct>             <dbl>     <int> <fct>
##  1     244000        2110 Attchd              522         0 North_Ames
##  2     213500        1338 Attchd              582         0 Stone_Brook
##  3     185000        1187 Attchd              420         0 Gilbert
##  4     394432        1856 Attchd              834         0 Stone_Brook
##  5     190000        1844 Attchd              546         0 Northwest_Ames
##  6     149000          NA Attchd              480         0 North_Ames
##  7     149900          NA Attchd              500         0 North_Ames
##  8     127500        1069 Attchd              440         0 Northpark_Villa
##  9     395192        1940 Attchd              606         0 Northridge_Heights
## 10     290941        1544 Attchd              868         0 Northridge_Heights
```

Terminologies/Notations

Default dataset - Contains credit card default data on 10,000 individuals. We are interested in predicting whether somebody will default or not.

The first ten observations are shown below.

```
##    default student   balance   income
## 1       No      No  939.0985 45519.02
## 2       No     Yes  397.5425 22710.87
## 3      Yes      No 1511.6110 53506.94
## 4       No      No  301.3194 51539.95
## 5       No      No  878.4461 29561.78
## 6      Yes      No 1673.4863 49310.33
## 7       No      No  310.1302 37697.22
## 8       No      No 1272.0539 44895.59
## 9       No      No  887.2014 41641.45
## 10      No      No  230.8689 32798.78
```

Terminologies/Notations

- **Response/Target/Outcome** - variable we are interested in predicting, denoted as Y
- **Features/Inputs/Predictors** - variables used to predict the response, denoted as X
- **Feature Matrix** - all features taken together, denoted as $\mathbf{X}$
- Number of data points/observations denoted as n
- Number of features/inputs/predictors denotes as p
- Missing entries in R are denoted as `NA`

Question!!!

For each of the **Ames Housing** and **Default** datasets,

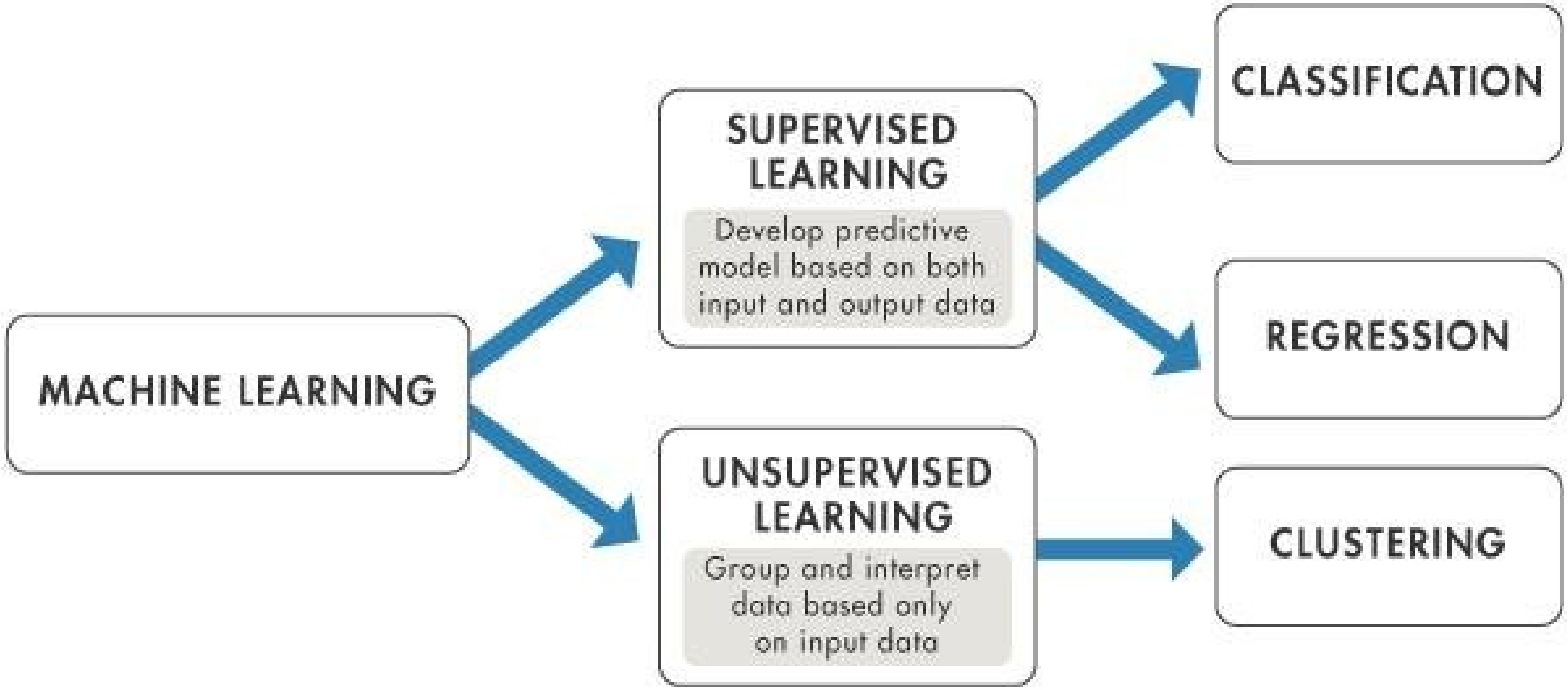
- What are the corresponding values of n and p?
- What will be the dimension of the corresponding response vector Y?
- What is the value of the 3rd feature for the 2nd observation?

Question!!!

Suppose you have information about 867 cancer patients on their age, tumor size, clump thickness of the tumor, uniformity of cell size, and whether the tumor is malignant or benign. Based on these data, you are interested in building a model to predict the type of tumor (malignant or benign) for future cancer patients.

- What are the values of n and p in this dataset?
- What are the inputs/features?

Supervised vs Unsupervised


MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Machine Learning Tasks (from Bunker and Fayez, 2017)

Supervised Learning

- We have access to **labeled** data
- The objective is to learn the overall pattern of the relationship between the inputs ($\mathbf{X}$) and response (Y) in order to
  - Investigate the relationship between inputs and response.
  - Predict for potential unseen **test** cases.
  - Assess the quality of predictions.

Supervised Learning problems can be categorized into

- **Regression** problems (response is quantitative, continuous)
- **Classification** problems (response is qualitative, categorical)

Unsupervised Learning

- No response/outcome variable, just $\mathbf{X}$.
- Understand structure within data.
    - find similar groups of observations based on features (**clustering**)
    - find a smaller subset of features with the most variation (**dimensionality reduction**)
- No gold-standard.
- Easier to collect unlabeled data.
- Useful pre-processing step for supervised learning.

Unsupervised Learning

US Arrests dataset - Data on arrests for 50 US states.

The first ten observations are shown below.

```
##             Murder Assault UrbanPop Rape
## Alabama       13.2     236       58 21.2
## Alaska        10.0     263       48 44.5
## Arizona        8.1     294       80 31.0
## Arkansas       8.8     190       50 19.5
## California     9.0     276       91 40.6
## Colorado       7.9     204       78 38.7
## Connecticut    3.3     110       77 11.1
## Delaware       5.9     238       72 15.8
## Florida       15.4     335       80 31.9
## Georgia       17.4     211       60 25.8
```

Question!!!

Some of the problems below are best addressed using a supervised learning algorithm, while others with an unsupervised learning algorithm. In each case, identify whether the problem belongs to the supervised or unsupervised learning paradigm. (Assume some appropriate dataset is available for your algorithm to “learn” from.)

- Examine the statistics of two football teams, and predict which team will win tomorrow’s match (given historical data of teams’ wins/losses to learn from).
- Given genetic (DNA) data from a person, predict the probability of the person developing diabetes over the next 10 years.
- Take a collection of 1000 essays written on the US economy, and find a way to automatically group these essays into a small number of groups of essays that are somehow “similar” or “related”.
- Examine data on the income and years of education of adults in a neighborhood and build a model to predict the income from years of education.