Introduction to Big Data and Data Science (CSCE 5300 Section 005)*

Yunhe Feng

Assistant Professor, Department of Computer Science and Engineering

19th September, 2024

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

1. Introduction to Machine Learning
2. Polynomial Regression
3. Assignment

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

What is Machine Learning

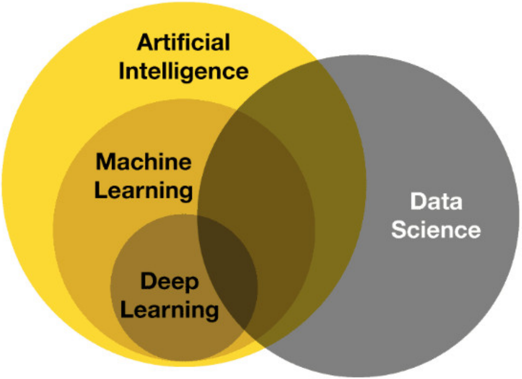


Figure 1: Deep Learning VS Data Science[1]

[1] https://www.deviq.io/insights/artificial-intelligence-vs-machine-learning-vs-data-science

Examples of Machine Learning

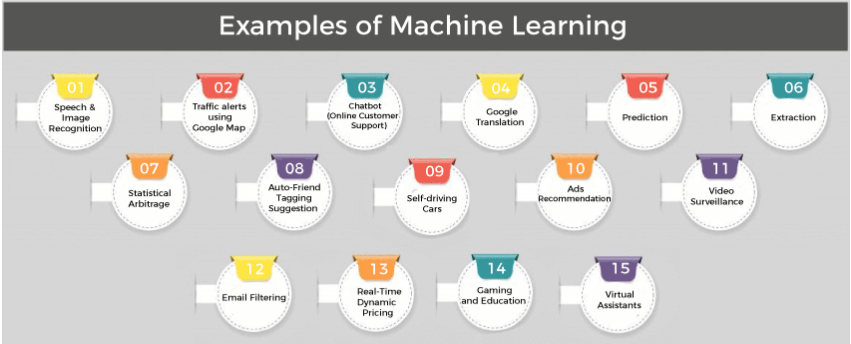
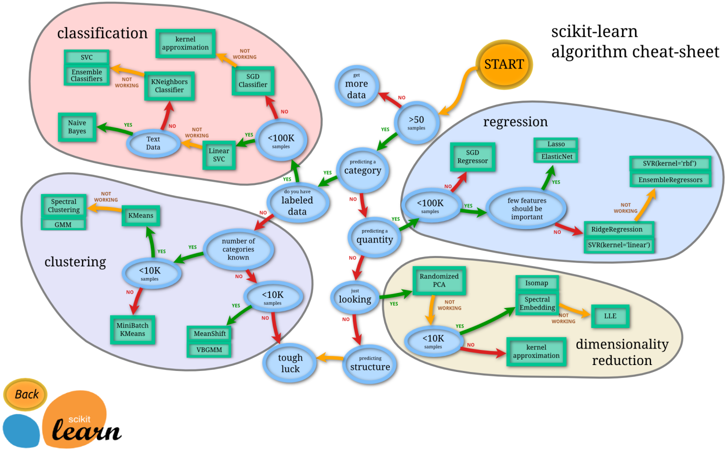


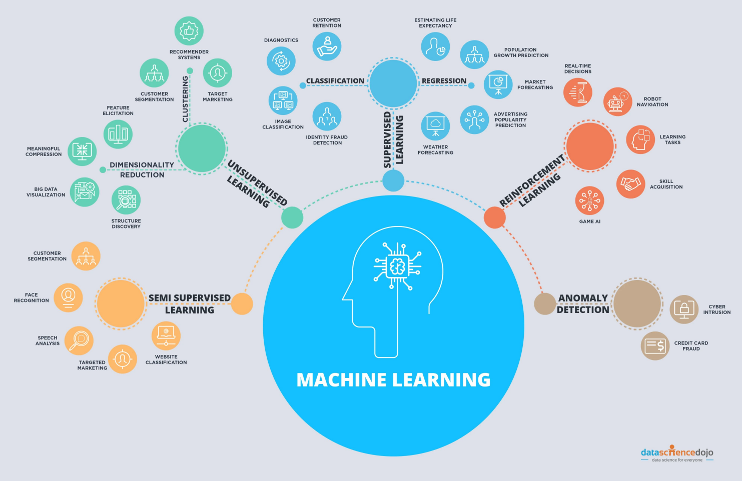
Figure 2: Applications of Machine Learning[2]

[2] https://tinyurl.com/yhbmpmz6

Machine Learning Models

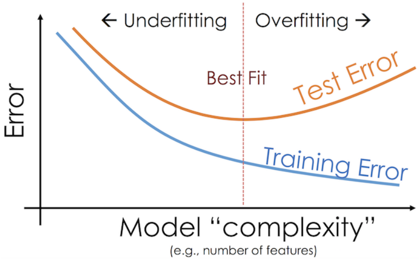


Machine Learning Models



Train Machine Learning Models

- Training Data: train the model
- Test Data: test the model performance
- Accuracy?
- Doesn't work, train again



Think About

- Can I use all data for training?
- What part of data for training is a good estimate?
- How do I measure goodness of a fit?
- What do I do if the fit is not good?

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

Terminology - Training, Test, and Validation Datasets

- **Training dataset**: Train the model.
- **Test dataset**: Evaluate the performance of the model on unseen data.
- **Validation dataset**: Fine-tune the model's hyperparameters and evaluate the performance of the model on unseen data during training.

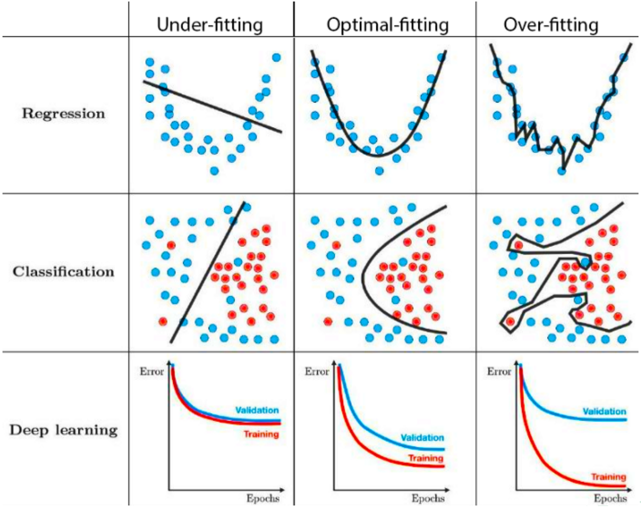
Ratios of Training, Test, and Validation Datasets

- 80/10/10
- 70/15/15
- 60/20/20
- 50/25/25
- Why and How to select?

More Terminologies

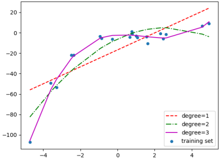
- **Hyperparameters**: Parameters that are used to control the learning process of a machine learning algorithm, e.g., learning rate and # of epochs
- **Underfit**: simple fit that may lead to lower correction
- **Overfit**: complex fit that may lead to over correction
- **Feature Selection**: the process of identifying and selecting the most informative and relevant features from a dataset.
- **Outliers**: point or group of points that do not follow the trend
- **Curse of Dimensionality**: a phenomenon that occurs when the number of features in a dataset is large

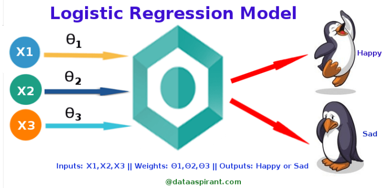
Fitting Data: Hyperparameters Trade Off in Complexity of the Fit

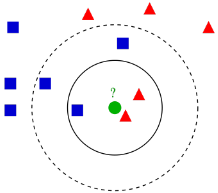


Machine Learning Algorithms

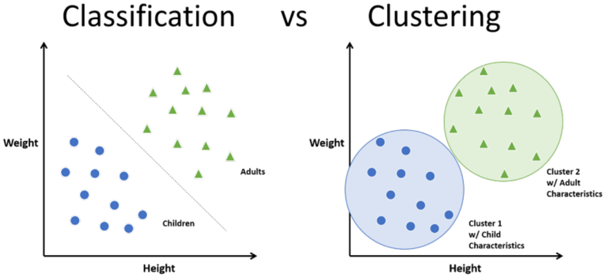
- Regression Algorithms for Machine Learning
- Most of the ML algorithms are applied to predict a class (classification) or a number (regression)
- Polynomial Regression (relationship between dependent and independent variables)
- Logistic Regression (categorical classifier)
- K-Nearest Neighbor (distance-based classification)




Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad
@dataaspirant.com



Classification VS Clustering[3]

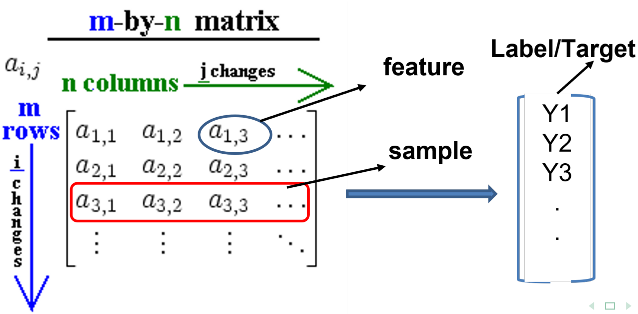

Classification vs Clustering
Weight
Height
Adults
Children
Weight
Height
Cluster 2 w/ Adult Characteristics
Cluster 1 w/ Child Characteristics

- **Classification** groups targets of a prediction (aka label data): Examples include detecting a cancer, adult, child, etc.
- **Clustering** groups similar instances together (sample data): Examples include genomic sequence, behaviors, etc.

[3] `https://tinyurl.com/yspm87sp`

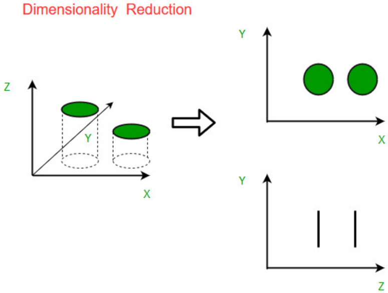
Matrices: Standard Representation of Machine Learning

- **Rows** – Instance (aka sample) of your data (SKU in supply chain, people in population, image in images, etc.)
- **Columns** - features (attribute) of your data (SKUs: how many SKUs, what type of SKUs, People: sex, disease, height)
- **Target** - what you are trying to get your system to predict (business intel, cost of care, etc.)

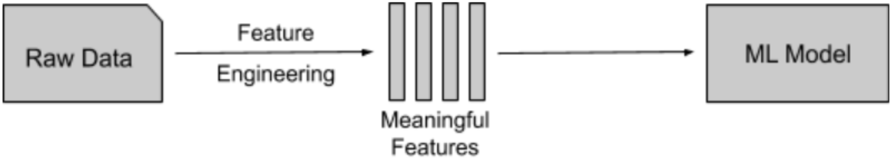

m-by-n matrix
$a_{i,j}$
n columns
j changes
m rows
i changes
$a_{1,1}$ $a_{1,2}$ $a_{1,3}$ …
$a_{2,1}$ $a_{2,2}$ $a_{2,3}$ …
$a_{3,1}$ $a_{3,2}$ $a_{3,3}$ …
feature
sample
Label/Target
Y1
Y2
Y3
.
.

Data Intelligence

- Pick good features (by hand)
- Find more data (on chosen features)
- What else can I do to improve Data Intelligence?
  - Extract meaningful relationships between data (sample -> feature)
  - Dimensionality reduction (can fewer features represent same outcome?)
  - Clustering (group related data)

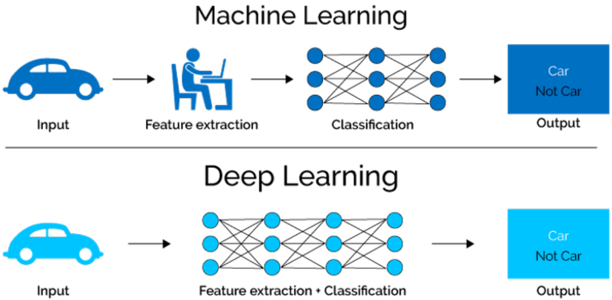


Feature Engineering



- Oftentimes, you can identify/create more features (called feature engineering) to improve the outcomes.
- Person with diabetes and gum disease is riskier than one with gum disease only.
- New features may be a result of intuition, data knowledge, or research.

Deep Learning (automatic features!)



- This is done in deep learning (“deep” = multilevel neural network)
- Deep Learning requires/employs MASSIVE AMOUNTS OF DATA. Several hidden variables/networks are employed. Also, It’s hard to understand how hidden layers connect input and output.

Recipe for Machine Learning

1. Step 1: Divide data into train and test datasets (see "cross-validation" recommended approaches)
2. Step 2: Explore dimensionality reduction and feature engineering for improved outcomes
3. Step 3: Apply models on the data – pick which one suits better such as classification, regression, etc.
4. Step 4: Validate model accuracies on test data
5. Step 5: If NOT acceptable accuracies go to STEP 3

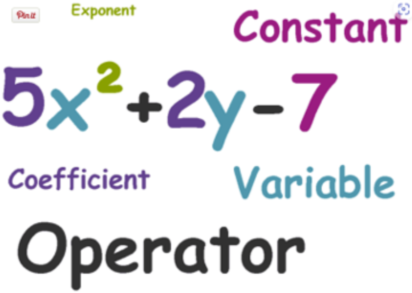
1. Introduction to Machine Learning

2. Polynomial Regression

3. Assignment

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

What is a Polynomial



- A polynomial is an algebraic expression composed of variables, constants, and exponents that are combined using mathematical operations (+, -, x, ÷)
- Represents relationship between variables
- Assists in predicting outcomes

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

What is not a Polynomial

An algebraic expression that contains

- fractional exponents
- negative exponents (example: $3X - 4X^{-2}$)
- Division by a variable (example: $3/x + 4X^2$)
- Radicals (an integer under * root, * square, cube, etc.)

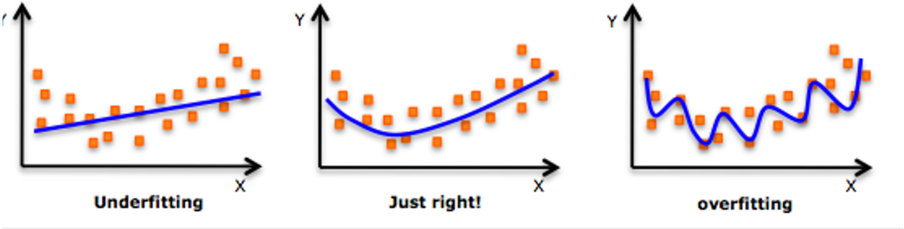
N^{th} Degree Polynomial

$Y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$ The power “n” of the polynomial Y is the degree of the polynomial

- n = 1 (linear): $Y = a_0 + a_1x$
- n = 2 (quadratic): $Y = a_0 + a_1x^+a_2x^2$
- n = 3 (cubic): $Y = a_0 + a_1x + a_2x^2 + a_3x^3$
- ...

Goal: pick the appropriate n to fit the data.

Accuracy of a Fit



Degree n = 1 Degree n = 3 Degree n = 7

- **MSE** (Mean Squared Error). The average squared difference between predicted and actual values
- **RMSE** (Root Mean Squared Error) Square root of MSE. RMSE is more commonly used because its in the same units as your prediction

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)

Mean Square Error Computation

- **MSE**: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **RMSE**: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Tutorials

https://tinyurl.com/mr3u9zdx

1. Introduction to Machine Learning

2. Polynomial Regression

3. Assignment

Assignment-4 (4.0 pts.)

- Introduction to MachineLearning (2 pts.)
- Polynomial Regression (2 pts.)

- Concept Paper / Extra Work / Research Project
- Idea selection and abstract submission (0.5 pts.)

*The teaching materials are reorganized and reformed based on Prof. Ravi Vadapalli's slides (Ravi.Vadapalli@unt.edu, UNT & University of Miami)