BIG DATA:
TECHNOLOGIES AND APPLICATIONS

10. Machine Learning and Data Mining

Il-Yeol Song, Ph.D.
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104

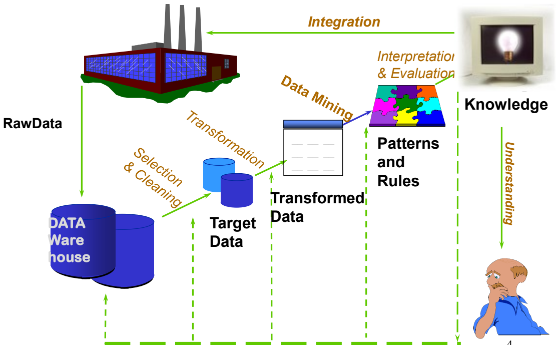
In this Lecture

- Introduction to Analytics and Machine Learning
- Machine Learning, Data Mining, Predictive Analytics
- Major Types of Machine Learning
  - Supervised Learning
    - Decision Tree
    - Regression
    - Logistic Regression
    - Ensemble
    - Neural Network and Deep Learning
  - Unsupervised Learning
    - K-Means Clustering
- Applications
- Validation
- Summary

Analytics

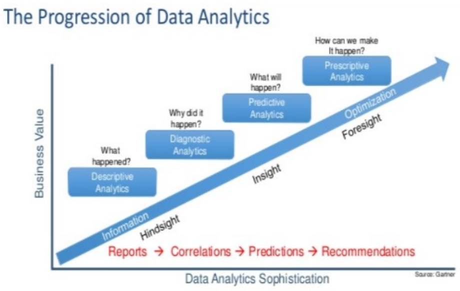
- **(Big Data) Analytics deals with "Value" of 5Vs of Big Data**
- **Data Analysis**
  - Explore, visualize, and evaluate data (descriptive statistics)
  - Focused on simple *structured data*
- **Data Analytics (Business Analytics, Data Science)**
  - Applications of DS (ML/DM) techniques to real-world problems
  - Techniques and tools applied to *big data with 4Vs* for deriving data-driven insight for better decision
  - ETL + Data Analysis + BI + Lifecycle + Business cases + Data mining and ML + Parallel processing techniques +etc

Knowledge Discovery Process



Source: Gregory Piatetsky-Shapiro

Levels of Analytics



Source: Incorporating the Data Lake into Your Analytic Architecture, Joe Cacerta, 2015

Big Data Analytics

- **Big data analytics (Data Science)**
  - Analyzing big data with the 4Vs to extract value to the business
  - It uses a variety of techniques such as parallel processing frameworks (Hadoop and Spark), machine learning, data mining, and statistics
  - Business analytics is a big data analytics with less emphasis on computational aspects, but focuses on business aspects
  - Data science in a broad sense is similar to big data analytics

Data Mining

- Data mining is knowledge discovery from data, discovering unknown patters and relationships among variables in data
  - May have no pre-formulated questions
- Analyzing massive amounts of data to uncover hidden trends, patterns, and relationships;
  - E.g., What are the characteristics of people using 5G Phone?
    - Age, education, income, occupation, etc.
  - E.g., Will a person buy iPhone or Samsung Galaxy?
    - Classification (data mining/machine learning)
  - E.g., Can we group these people based on the characteristics?
    - Clustering (data mining/machine learning)
- DM typically uses **batched information** to reveal a new insight at a particular point in time rather than an on-going basis;

Machine Learning

- A subfield of AI
- Machine learning gives computer systems the ability to automatically "**learn**" with data, without being explicitly programmed.
- ML and DM uses the same key algorithms to discover patterns in data
- ML differs from DM in process and utility:
  - ML automatically learn parameters of the models from the data
  - ML uses self-learning algorithms to improve its performance at a task with experience over time
  - ML may be batch or continuous, while DM is batch-oriented
  - This ability to learn and adapt makes it the optimal choice for improvements in ongoing processes, marketing campaigns and continuous customer service improvements

Predictive Analytics

- ***Predictive analytics*** focuses on creating actionable models to predict future behaviors and events
  - A subset of machine learning techniques that predict future outcome from data based on previous patterns
  - Employs data mining/machine learning techniques to create actionable predictive models based on available data
  - Used in areas such as finance, customer relationships, customer service, customer retention, fraud detection, targeted marketing, and optimized pricing
    - Ex: Should we approve the loan?
    - Ex: Will the person buy my product?

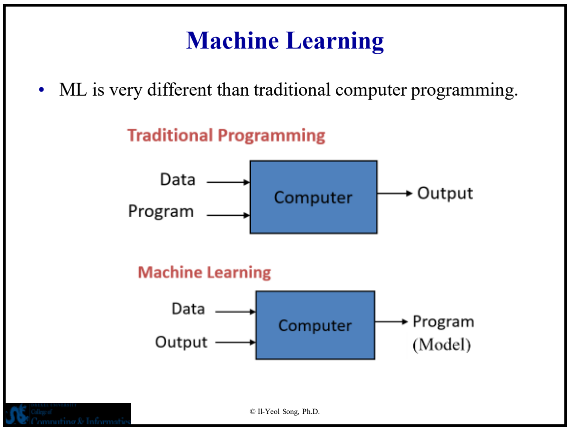
Introduction to Machine Learning Techniques

Key Points of Machine Learning

- *A collection of algorithms and techniques used to* ***build systems (models) that learn from data.***
- Learn from data and build a model
- Data transformation is the biggest and hardest portion of Machine Learning.
- Only as good as the data you use to train the machine.
- There is a chance for errors in Machine Learning.
- AI is not going to out-smart the human species and take over soon, but maybe some day (**Singularity point**).

Example Problem of Machine Learning

- It is very hard to write programs that solve problems like **recognizing a face**.
  - We don't know what program to write because we don't know how our brain recognize human faces.
  - Even if we had a good idea about how to do it, the program might be **horrendously complicated**.
- Instead of writing a program by hand, we **collect lots of examples** that specify **the correct output** for a given input.
- A machine learning algorithm then takes these examples and produces a program that does the job.
  - If we do it right, the program works for new cases as well as the ones we trained it on.



Machine Learning

- ML is very different than traditional computer programming.

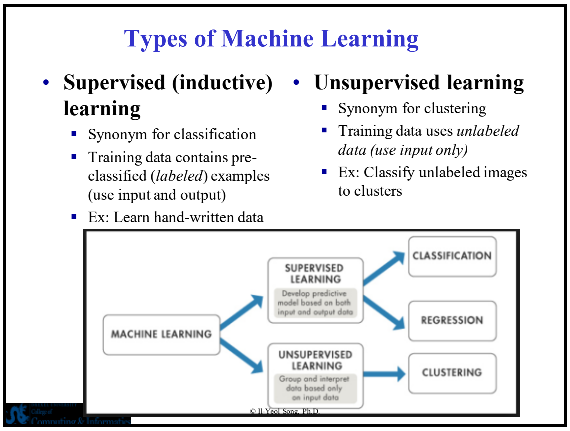
Traditional Programming

Data → Computer → Output
Program →

Machine Learning

Data → Computer → Program (Model)
Output →

© Il-Yeol Song, Ph.D.

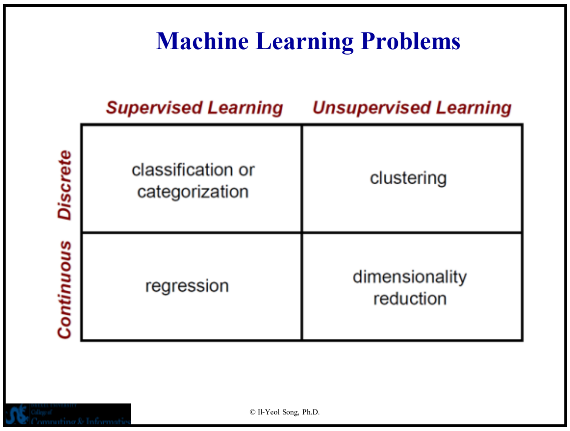


Types of Machine Learning

- **Supervised (inductive) learning**
  - Synonym for classification
  - Training data contains pre-classified (*labeled*) examples (use input and output)
  - Ex: Learn hand-written data
- **Unsupervised learning**
  - Synonym for clustering
  - Training data uses *unlabeled data (use input only)*
  - Ex: Classify unlabeled images to clusters

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

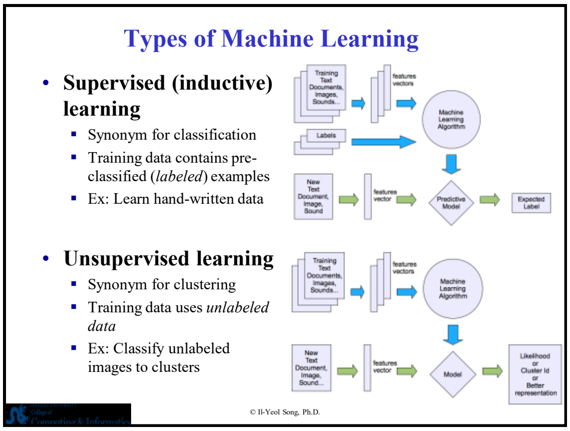
© Il-Yeol Song, Ph.D.



Machine Learning Problems

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

© Il-Yeol Song, Ph.D.

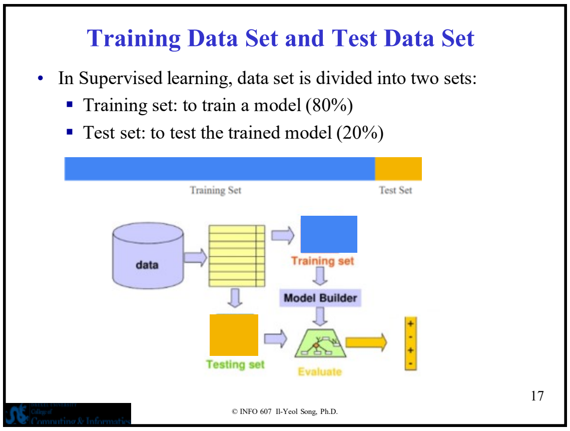


Types of Machine Learning

- **Supervised (inductive) learning**
  - Synonym for classification
  - Training data contains pre-classified (*labeled*) examples
  - Ex: Learn hand-written data
- **Unsupervised learning**
  - Synonym for clustering
  - Training data uses *unlabeled data*
  - Ex: Classify unlabeled images to clusters

Training Text Documents, Images, Sounds...
features vectors
Labels
Machine Learning Algorithm
New Text Document, Image, Sound
features vector
Predictive Model
Expected Label
Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
New Text Document, Image, Sound...
features vector
Model
Likelihood or Cluster Id or Better representation

© Il-Yeol Song, Ph.D.

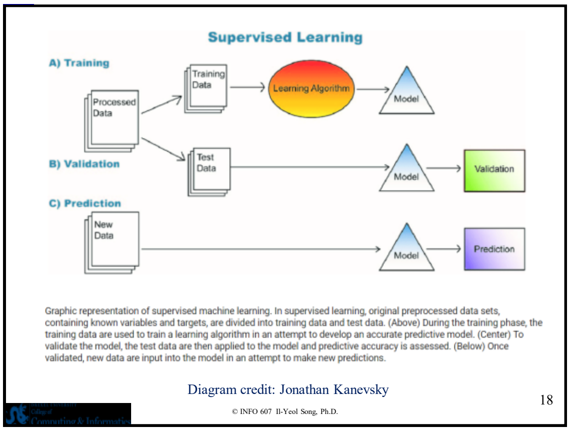


Training Data Set and Test Data Set

- In Supervised learning, data set is divided into two sets:
  - Training set: to train a model (80%)
  - Test set: to test the trained model (20%)

17

© INFO 607 Il-Yeol Song, Ph.D.



Supervised Learning

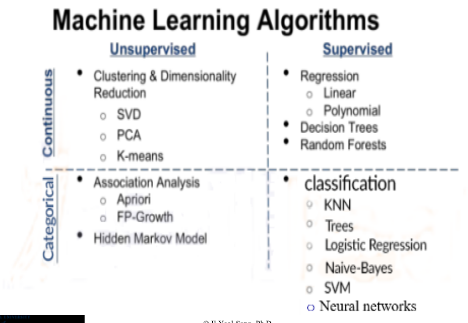
Graphic representation of supervised machine learning. In supervised learning, original preprocessed data sets, containing known variables and targets, are divided into training data and test data. (Above) During the training phase, the training data are used to train a learning algorithm in an attempt to develop an accurate predictive model. (Center) To validate the model, the test data are then applied to the model and predictive accuracy is assessed. (Below) Once validated, new data are input into the model in an attempt to make new predictions.

Diagram credit: Jonathan Kanevsky

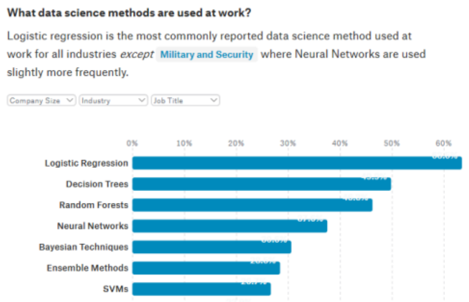
18

© INFO 607 Il-Yeol Song, Ph.D.

Machine Learning Problems



Kellog 2017 Survey: Representative Algorithms


What data science methods are used at work?

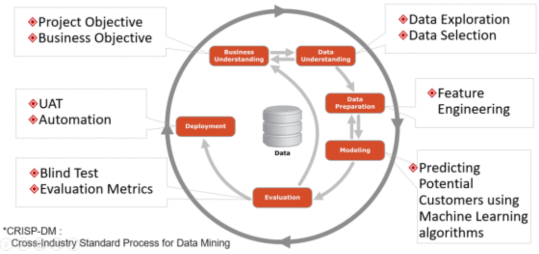
Logistic regression is the most commonly reported data science method used at work for all industries *except* Military and Security where Neural Networks are used slightly more frequently.

https://www.kaggle.com/surveys/2017

How Can We Do Data Mining?

- By Utilizing the CRISP-DM Methodology
- CRoss Industry Standard Process for Data Mining
- Initiative launched Sept.1996
- Funding from European commission
- Over 200 members of the CRISP-DM SIG worldwide
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify, ..
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, …
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, ...

Phases in the DM Process: CRISP-DM



*CRISP-DM :
Cross-Industry Standard Process for Data Mining

Model selection

- There are a lot of different kinds of models out there!
- Choosing the right model starts with knowing what exists.
- **Exploratory data analysis** should be a guide. So, selection of a model can occur early on after exploration,
- Can **perform a comparative analysis of multiple models** and then choose the best-performing model.
- However, performance should not be the only consideration.
- Some other selection factors:
  - What model is a good theoretical match for the data?
  - How does it match a business model?
  - How domain-portable is a model? How difficult is a model to implement?
  - Will the model scale across multiple machines? How transparent are a model's inner workings?

Representative ML/DM Algorithms

Machine Learning Algorithms

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM ○ Neural networks

Queries ML/DM/DS Answers:

- **Classification**: Is this A or B?
  - Deny or approve a loan; spam or not? Cancer or not?, Coupon/discount? etc
- **Abnormal Detection**: Is this abnormal, weird? (detect unexpected or unusual events or behaviors)
  - Identify CC fraud in real-time? Any abnormal sensor data?
- **Regression**: For a given input data set, how many or how much
  - Temperature tomorrow? Sales next quarter?
- **Clustering**: How is this organized?
  - profitable customer groups? movie/book recommendation; which patients are responding to the therapy?
- **Reinforcement learning**: What should do next?
  - Driverless car: accelerate/stop on yellow light, robot vacuum
- **Association**: Find the correlation between buying patterns of products of type A and those of type B.

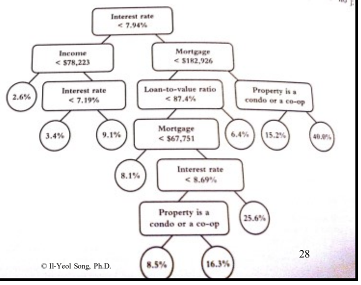
What is a Decision Tree Classifier?

Decision Tree - An Intuitive Introduction
https://www.kdnuggets.com/2019/02/decision-trees-introduction.html

Decision Tree

“Classify the data in the order of important variables that distinguish the target variable”

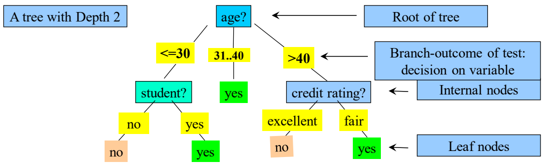
“Decision tree showing probability of prepayments: 10 segments.”



What is a Decision Tree Classifier?

- Given input X = {X1, X2, …., Xn}, the goal is to predict a response output variable (or target variable) Y.
- The prediction logic is structured by a tree that specifies sequences of decisions and consequences
  - Theory is well-defined
  - Easy to understand the decision rules it has learned
- Input variables can be continuous or discrete
- Output: A tree that describes the decision flow
- Leaf nodes return either a probability score or simply a classification

Decision Tree Induction: Explanation

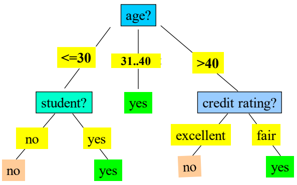


- ❑ Tuples flow along the tree and are partitioned at each branch
- ❑ Internal node denotes an attribute
- ❑ Branch represents the values of the node attribute
- ❑ Leaf nodes represent **class labels**

Decision Tree Induction: An Example

- ❑ Training data set class variable: **Buys_computer**
- ❑ Popular DT algorithms: Quinlan's ID3, C4.5, CART.
- ❑ Example tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Algorithm for Decision Tree Induction

- **Basic Tree Construction Algorithm**
  - At start we choose one attribute as the root and put all its values as branches
  - We choose recursively internal nodes (attributes) with **the greatest information gain value**.
  - We Stop when
    - all the samples (records) are of the same class, then the node becomes the leaf labeled with that class
    - or there is no more data left
    - or there is no more new attributes to be put as the nodes. In this case we apply MAJORITY VOTING to classify the node. All samples for a given node belong to the same class
  - **Tree pruning**
    - Identify and remove branches that reflect noise or outliers

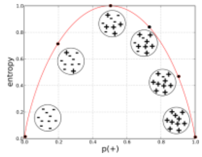
A Criterion for Attribute Selection

- Which is the best attribute?
  - The one which will result in the smallest tree.
  - The attribute with the greatest information gain
  - Heuristic: ***choose the attribute that produces the "purest" nodes!***
- Popular ***impurity*** criterion: ***Entropy, which measures the level of uncertaionty***
  - *High entropy: Impure*
  - *Low entropy: pure*
- We can then compare a tree **before** the split and **after** the split using ***Information Gain*** = ***Entropy (before) – Entropy (after).***
  - Statistical quantity measuring how well an attribute classifies the data.
- ***Strategy: choose attribute that results in the greatest information gain.***

Selecting Informative Attributes

- The most common splitting criterion is called **information gain** (IG)
  - It is based on a **purity measure** called **entropy**
    - ***A measure of uncertainty*** *(disorder)* associated with a variable
    - $entropy\ = -\mathrm{p}_1 \log_2(p_1) - p_2 \log_2(p_2) - .. =$ $-\sum_{i=1}^{m} p_i \log_2(p_i)$
      P_i = probability of property i within the set
    - Entropy = 0 (data is pure)
    - Entropy = 1(max disorder)
    - Example for R={a,a,a,b,b,b,b,b}

$$entropy\ (R) = I(R) = -\left[\underbrace{\left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right)}_{\text{a-values}} + \underbrace{\left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right)}_{\text{b-values}}\right]$$

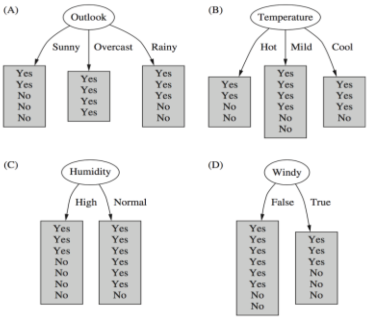


Weather Data: Play or not Play?

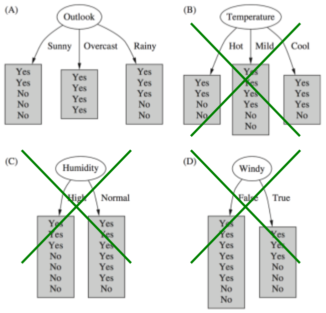
Shall we play tennis today?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

Which attribute to select?

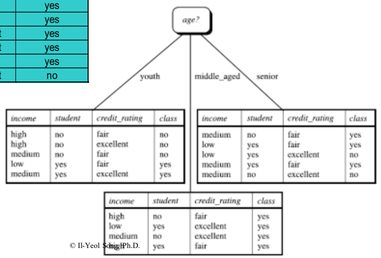


Which attribute to select?



The Decision Tree after Splitting by Age

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



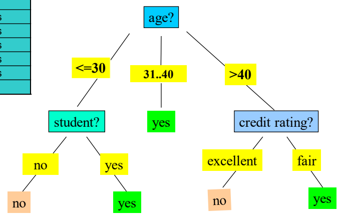
income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

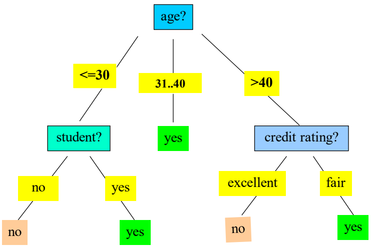
The Final Decision Tree

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



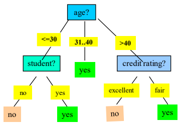
Predicting New Data

- Suppose Bill Clinton (age=55, income=high, student=no, credit_rating=excellent). Wil he buy a computer?



Rule Extraction from a Decision Tree

- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction: the leaf holds the class prediction
- Rules are mutually exclusive and exhaustive



IF *age* = young AND *student* = *no* THEN *buys_computer* = *no*

IF *age* = young AND *student* = *yes* THEN *buys_computer* = *yes*

IF *age* = mid-age THEN *buys_computer* = *yes*

IF *age* = old AND *credit_rating* = *excellent* THEN *buys_computer* = *yes*

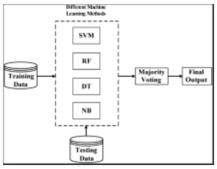
IF *age* = young AND *credit_rating* = *fair* THEN *buys_computer* = *no*

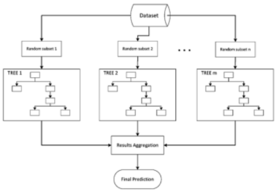
Classification in Large Databases

- ID3 and C4.5 assumes data fits into main memory
  - OK for up to several hundreds of thousands of data
- Scalability: Classifying data sets with millions of examples and hundreds of attributes with a reasonable speed
- **Why is decision tree induction popular**?
  - output is easily understood by rules
  - relatively faster learning speed (than other classification methods)
  - convertible to simple and easy to understand classification rules
  - can use SQL queries for accessing databases
  - comparable classification accuracy with other methods

Ensemble Methods

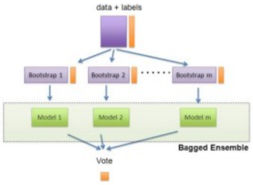
- **Methods of Using Multiple Models to Improve**
  - Bagging
  - Boosting
  - Random Forests





Ensemble Methods: Bagging

- **Bagging (Boostrap Aggregating)**
  - **Bootstrap** means ***random sampling with replacement***
  - Build a classifier for each bootstrap
  - Each model runs independently and then aggregate the outputs at the end without preference to any model

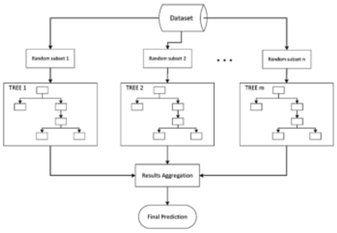


Ensemble Methods: Boosting

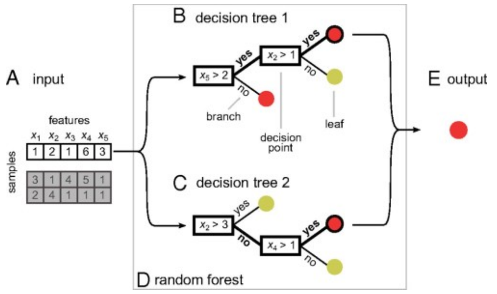
- **Boosting**
  - Creates multiple weak models whose output is added together to get an overall prediction
  - With each iteration of boosting, a new model is created and the new model is trained (updated) from the errors of the previous learners.
  - The final model is a weighted combination of all the individual model
  - A popular model: Adaboost (Adaptive boosting)

Ensemble Methods: Random Forests

- **Random Forests**
  - Each model is created from BAGGing + samples of features



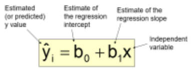
Individual decision trees vote for class outcome in a toy example random forest.



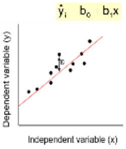
Danielle Denisko, and Michael M. Hoffman PNAS 2018;115:8:1690-1692

Regression in a Nutshell

- Regression is the most frequently used technique for ***predicting a continuous outcome.***
  - Focuses on the relationship between an outcome and its input variables as well as have a sense of how changes ir individual parameters affect the outcome :


Estimated (or predicted) y value
Estimate of the regression intercept
Estimate of the regression slope
Independent variable

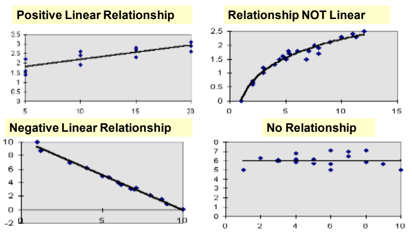
$$\hat{y}_i = b_0 + b_1 x$$



❑ Input (independent) variables: x is a continuous or discrete variable
❑ Output (dependent) variable, y, is ***continuous***:

❑ Regression optimizes *the sum of the squares of errors* ($\sum (y - \hat{y})^2$)

Types of Regression Models



Linear Regression: Use cases

- Use Linear regression ***to predict a continuous value*** as a linear or additive function of other variables.
  - The most frequently used technique for predicting a continuous outcome.
  - It is simple and works well in most instances.
  - Try this first; if it fails, try other complex methods.
- Example questions:
  - Predict the lifetime value (LTV) of customers and understand what drives LTV
  - Predicting income as a function of number of years of education, age and gender
  - Predict House price (outcome) as a function of square footage, number of rooms, number of bathrooms

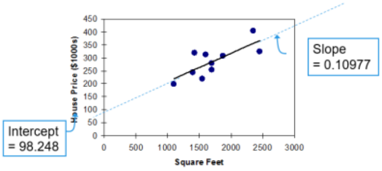
Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (y) = house price in $1000s
  - Independent variable (x) = square feet

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Graphical Presentation

- House price model: scatter plot and regression line



Slope = 0.10977

Intercept = 98.248

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

$\hat{y}_i$ b_0 b_1x

Interpretation of the Slope Coefficient, b_1

□ $\widehat{\text{house price}} = 98.24833 + \boxed{0.10977}\,(\text{square feet})$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = \boxed{.10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Prediction Example: House Prices

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

Predict the price for a house with 2000 square feet

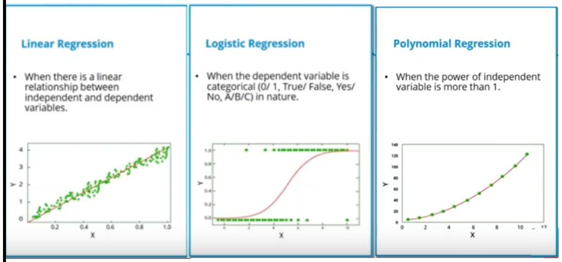
$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098\,(\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85\end{aligned}$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Linear Regression Summary

- **Pros**
  - Concise representation by coefficients
  - Robust to redundant variables and correlated variables
  - Some explanatory value by relating impact of each variable on the outcome
  - Easy to score new data
- **Cons**
  - Does not handle missing values well
  - Cannot handle variables that affect the outcome in a discontinuous way.
  - Assumes that each variable affects the outcome linearly and additively

Regression Types



Linear Regression

- When there is a linear relationship between independent and dependent variables.

Logistic Regression

- When the dependent variable is categorical (0/ 1, True/ False, Yes/ No, A/B/C) in nature.

Polynomial Regression

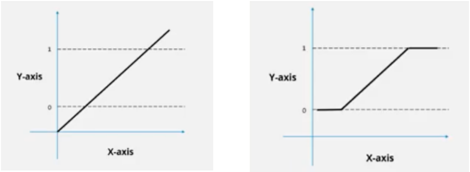
- When the power of independent variable is more than 1.

Logistic Regression

- What is it?
  - ***A classifier that predicts a categorical outcome variable*** from one or more categorical or continuous predictor variables.
  - Used to ***estimate the probability that an event will occur*** as a function of other variables
    - Ex: The probability that a borrower will default as a function of his credit score, income, the size of the loan, and existing debts.
- Why not Linear Regression?
  - Having a categorical outcome variable violates the assumption of linearity in normal regression

Logistic Regression in a Nutshell

- ***A classifier that predicts a categorical outcome variable*** with probability
  - Categorical Outcome (0 or 1, yes or no, A or B)



- Since the value of Y is between 0 and 1, the linear line has to be clipped at 0 or 1.
- The new curve cannot fit by a single linear line, and thus we need a different formula

Logistic Regression Formulation

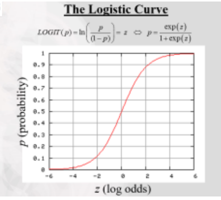
- Logistic Regression model consists of vector $\boldsymbol{\beta}$ in d dimensional space.
- For a point X in the d-space,

$$z = \alpha + \boldsymbol{\beta} \cdot \mathbf{x} = \alpha + \beta_1 x_1 + \cdots + \beta_d x_d$$

- Z has range of (-∞,∞).
- ***Logistic regression computes the probability P using the logistic function:***

$$P = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

- The logistic function maps z(-∞,∞) to P (0,1).



The logistic regression maps point x in d-space to a value in the range of o to 1 using the logistic function.

Logistic Regression: Use Cases

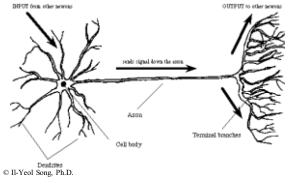
- Logistic regression is the preferred method for many binary classification problems
  - Especially, if you are interested in the probability of an event, not just predicting "yes" or "no"
- Use Cases for Binary Classification:
  - The probability the borrow will default
  - The probability the customer will churn
  - Respond to medical treatment/no response
  - Will purchase from a website/no purchase
  - Likelihood Spain will win the next World Cup
  - Approve/deny; cured/not-cured; dead/alive; progressed/Not-progressed/
- Logistic regression can also be used for multi-class Classification:

Logistic Regression: Summary

- **Pros**
  - Easy to score data
  - Returns good probability estimates of an event
  - Makes no assumption about distributions of classes in feature space
  - Easily extended to multiple classes
  - Quick to train
  - Very fast at classifying unknown records
  - Good accuracy for many simple data sets
  - Resistant to overfitting
- **Cons**
  - Does not handle missing values well
  - Does not work well with discrete variables that have many distinct values such as zip code.

Neural Network for Classification

- **A classifier, loosely based on the ideas of human's neurons interconnected by synapses in brain**
  - Each node receives data, performs an operation, and passes the new data to another node via a link.
  - Inputs are approximately summed
  - When the input exceeds a threshold the neuron sends an electrical spike

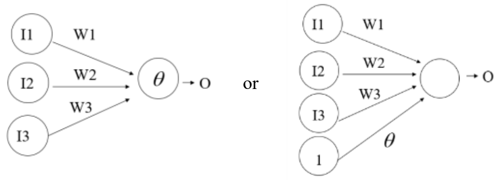


Perceptron

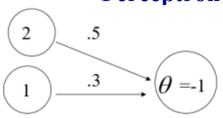
- Initial proposal of connectionist networks
- Essentially, a linear formula composed of nodes and weights

Activation Function with bias θ

$$O = \begin{cases} 1 : \left(\sum_i w_i I_i\right) + \theta > 0 \\ 0 : otherwise \end{cases}$$



Perceptron Example



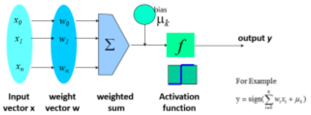
2(0.5) + 1(0.3) + -1 = 0.3 , O=1

Learning Procedure:

- Randomly assign weights (between 0-1)
- Get inputs from training data
- Get output O, adjust weights to gives results toward our desired output T. The parameter with a higher predictive power gets a higher weight
- Repeat; stop when no errors, or enough epochs completed

Neural Network for Classification

- A neural network: A set of connected input/output network, where
  - An input vector **x** is mapped into variable y.
  - The inputs are multiplied by their corresponding weights to form a weighted sum, which is added to the bias associated with unit. Then a nonlinear activation function is applied to it.

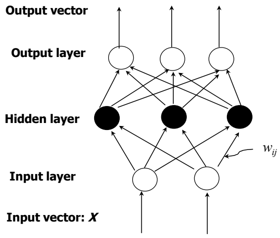


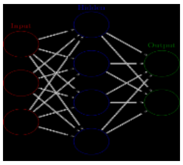
- Also referred to as **connectionist learning** due to the connections between units
- **Backpropagation**: The most popular neural network learning algorithm

Neural Network as a Classifier

- **Weakness**
  - Long training time
  - Require a number of parameters typically best determined empirically, e.g., the network topology or ``structure."
  - Poor interpretability: Difficult to interpret the symbolic meaning behind the learned weights and of ``hidden units" in the network
- **Strength**
  - High tolerance to noisy data
  - Ability to classify untrained patterns
  - Well-suited for both classification and numeric prediction (continuous-valued inputs and outputs)
  - Successful on a wide array of real-world data
  - Algorithms are inherently parallel

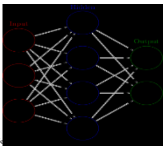
A Multi-Layer Feed-Forward Neural Network



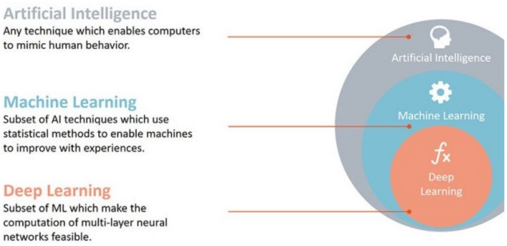


Classifying Unknown Tuples in a Trained NN

- The unknown tuple X is input to the trained network
  - If there is one output node per class, then the output node with the highest value determines the predicted class label for X.
  - If there is only one output node, then output values greater than or equal to 0.5 may be considered as belonging to the positive class, while values less than 0.5 may be considered negative.
  - The closer the value is to 1, the more likely the output variable has a higher prediction power



From Neural Networks to Deep Learning



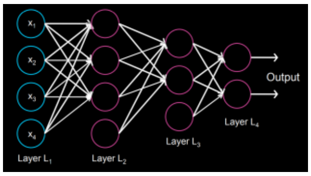
Artificial Intelligence
Any technique which enables computers to mimic human behavior.

Machine Learning
Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

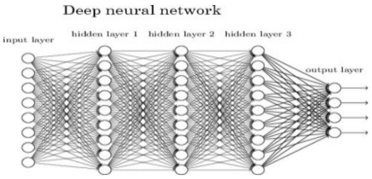
Deep Learning
Subset of ML which make the computation of multi-layer neural networks feasible.

From Neural Networks to Deep Learning

- Train networks with many layers
  - Multiple layers work to build an improved feature space
  - First layer learns 1st order features (e.g., edges, …)
  - 2nd layer learns higher order features (combinations of first layer features, combinations of edges, etc.)
  - Then final layer features are fed into supervised layer(s)

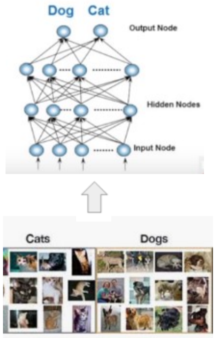


From Neural Networks to Deep Learning

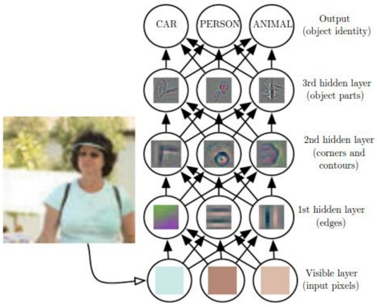


- Real-world NNs may contain thousands of nodes and billions of links.
- A Google X team built 16,000 central processing units (CPUs) to power a NN with over a billion connections. Later, they showed sixty-four **GPU**s could handle the same amount of work as 16,000 CPUs
- It was trained to process 10 million images from randomly selected YouTube videos to recognize cat images.

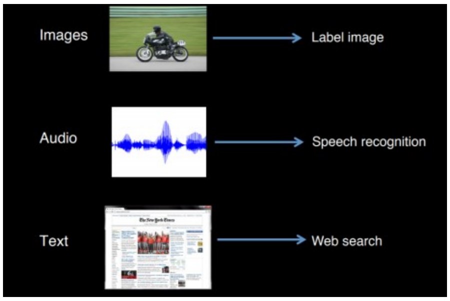
Classifying Unknown Tuples in a Trained NN



Deep Learning

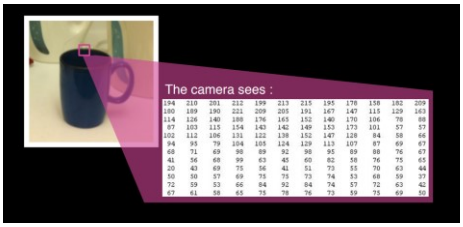

CAR
PERSON
ANIMAL
Output (object identity)
3rd hidden layer (object parts)
2nd hidden layer (corners and contours)
1st hidden layer (edges)
Visible layer (input pixels)

Deep Learning Applications



Slide Credit: Andrew Ng, Stanford University.

Why is Computer Vision Hard?



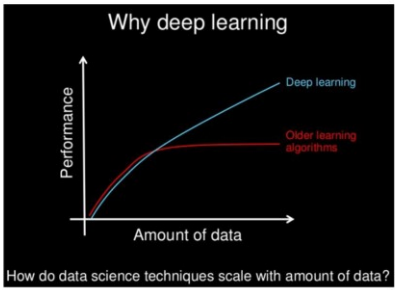
The camera sees :

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	75	65
20	43	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

Why is Speech Recognition hard?

Microphone recording:
Please find the coffee mug

Why Deep Learning?



Slide Credit: Andrew Ng, Stanford University.

Deep Learning Applications

- Good for inexact learning
  - Image/audio/video recognition (e.g, face recognition, translating voice into text)
  - NLP
- Applications
  - Determine which online ads to display in real time
  - Identify and tags friends in photos
  - Translate text into different languages on a Web page
  - Drive autonomous vehicles
  - Use for fraud detection (CC companies)
  - Predict whether you will cancel a subscription
  - Provide personalized customer recommendations
  - Use it to predict bankruptcy and loan risk
  - Hospitals use it for detection, diagnosis, and treatment of diseases
  - Other options include text analysis, image captioning, image colorization, x-ray analysis, weather forecasts, finance predictions, and more.

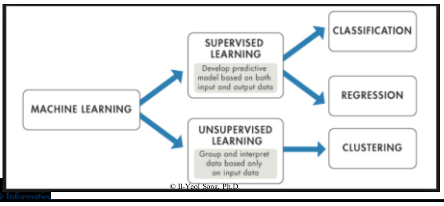
Deep Learning Applications

- NLP: Deep learning is being used to improve natural language processing tools to consistently comprehend the meaning of a sentence, not just the individual words. So if someone wants to translate 'take a hike' or 'get lost' it will not take the expression literally. It will translate the expression into a corresponding expression in the other language.

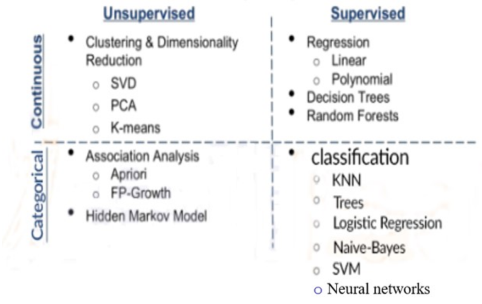
- Deep learning is overkill if your project uses small data volumes and solves simple problems. If you process large amounts of data and need to produce complex predictions, deep learning technology may be beneficial

Types of Machine Learning

- **Supervised (inductive) learning**
  - Synonym for classification
  - Training data contains pre-classified (*labeled*) examples (use input and output)
  - Ex: Learn hand-written data
- **Unsupervised learning**
  - Synonym for clustering
  - Training data uses *unlabeled data (use input only)*
  - Ex: Classify unlabeled images to clusters


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 Algorithms



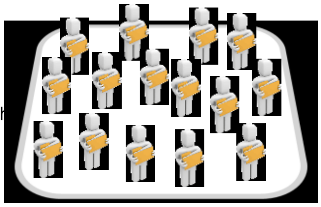
	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM ○ Neural networks

What Is Cluster Analysis?

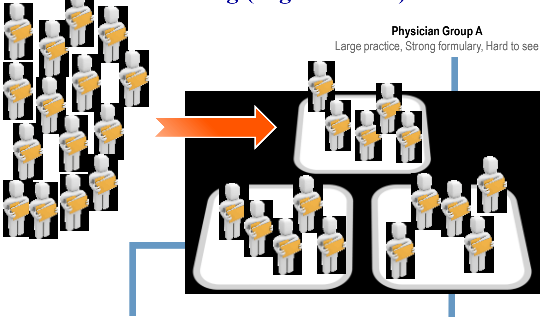
- **What is a cluster?**
  - A cluster is a collection of data objects which are
    - Similar (or related) to one another within the same group (i.e., cluster)
    - Dissimilar (or unrelated) to the objects in other groups (i.e., clusters)
- **Cluster analysis** (or ***clustering*, *data segmentation*, *automatic classification*** ...)
  - Given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible
- Cluster analysis is **unsupervised learning** (i.e., no predefined classes)
  - This contrasts with *classification* (i.e., *supervised learning*)
- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution, or
  - As a preprocessing (or intermediate) step for other algorithms (e.g, outlier detection)

Clustering (Segmentation)

"Create groups that have similar characteristics. Also a measure of how different each group is from the others."



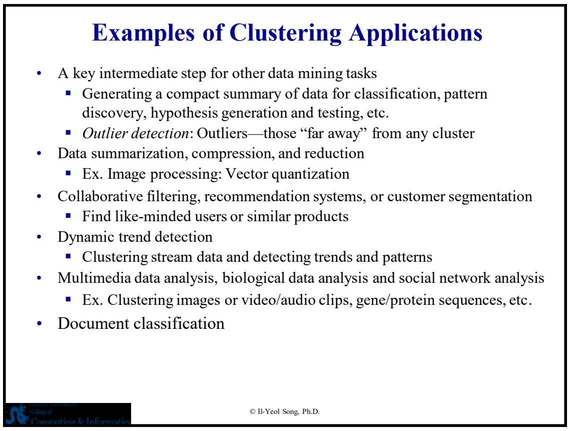
Clustering (Segmentation)



Physician Group A
Large practice, Strong formulary, Hard to see

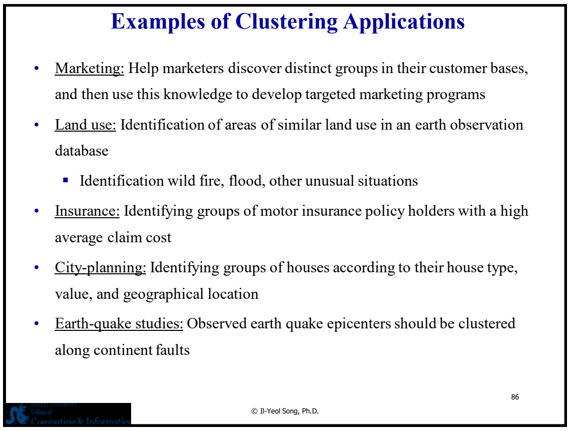
Physician Group B
Small Practice, Weak formulary, Easy to see

Physician Group C
GI Specialists, existing prescribers of Brand B



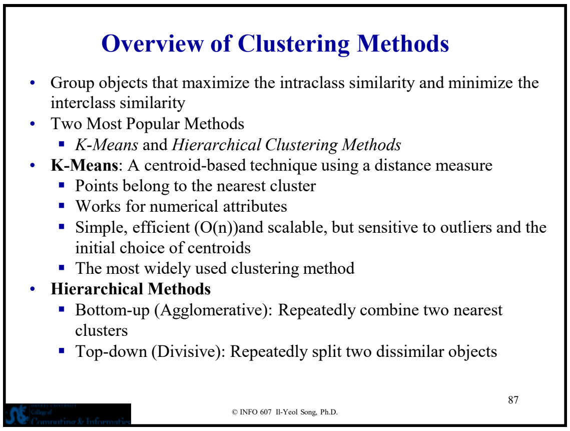
Examples of Clustering Applications

- A key intermediate step for other data mining tasks
  - Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
  - *Outlier detection*: Outliers—those "far away" from any cluster
- Data summarization, compression, and reduction
  - Ex. Image processing: Vector quantization
- Collaborative filtering, recommendation systems, or customer segmentation
  - Find like-minded users or similar products
- Dynamic trend detection
  - Clustering stream data and detecting trends and patterns
- Multimedia data analysis, biological data analysis and social network analysis
  - Ex. Clustering images or video/audio clips, gene/protein sequences, etc.
- Document classification



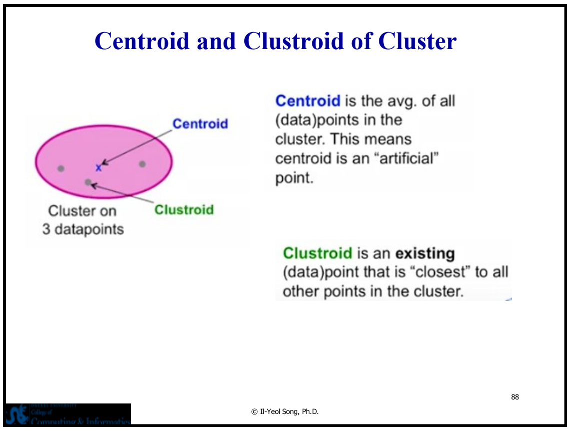
Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
  - Identification wild fire, flood, other unusual situations
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults



Overview of Clustering Methods

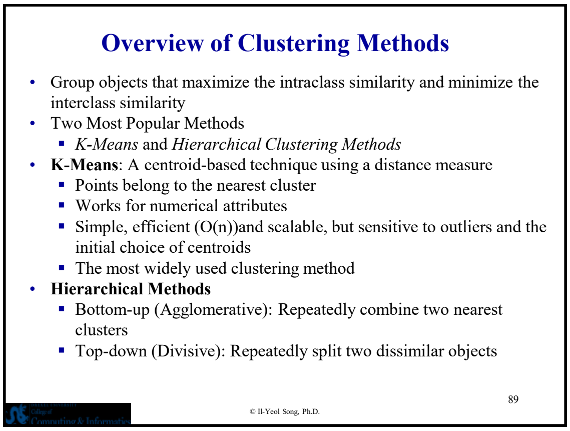
- Group objects that maximize the intraclass similarity and minimize the interclass similarity
- Two Most Popular Methods
  - *K-Means* and *Hierarchical Clustering Methods*
- **K-Means**: A centroid-based technique using a distance measure
  - Points belong to the nearest cluster
  - Works for numerical attributes
  - Simple, efficient (O(n))and scalable, but sensitive to outliers and the initial choice of centroids
  - The most widely used clustering method
- **Hierarchical Methods**
  - Bottom-up (Agglomerative): Repeatedly combine two nearest clusters
  - Top-down (Divisive): Repeatedly split two dissimilar objects



Centroid and Clustroid of Cluster

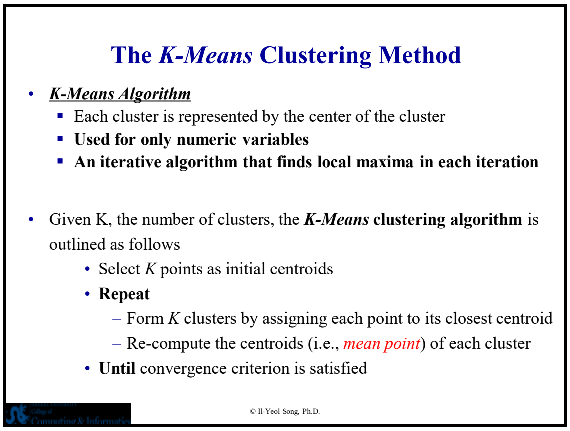
Centroid is the avg. of all (data)points in the cluster. This means centroid is an "artificial" point.

Clustroid is an **existing** (data)point that is "closest" to all other points in the cluster.



Overview of Clustering Methods

- Group objects that maximize the intraclass similarity and minimize the interclass similarity
- Two Most Popular Methods
  - *K-Means* and *Hierarchical Clustering Methods*
- **K-Means**: A centroid-based technique using a distance measure
  - Points belong to the nearest cluster
  - Works for numerical attributes
  - Simple, efficient (O(n))and scalable, but sensitive to outliers and the initial choice of centroids
  - The most widely used clustering method
- **Hierarchical Methods**
  - Bottom-up (Agglomerative): Repeatedly combine two nearest clusters
  - Top-down (Divisive): Repeatedly split two dissimilar objects



The *K-Means* Clustering Method

- ***K-Means Algorithm***
  - Each cluster is represented by the center of the cluster
  - **Used for only numeric variables**
  - **An iterative algorithm that finds local maxima in each iteration**
- Given K, the number of clusters, the ***K-Means*** **clustering algorithm** is outlined as follows
  - Select *K* points as initial centroids
  - **Repeat**
    - Form *K* clusters by assigning each point to its closest centroid
    - Re-compute the centroids (i.e., *mean point*) of each cluster
  - **Until** convergence criterion is satisfied

The *K-Means* Clustering Method: Example

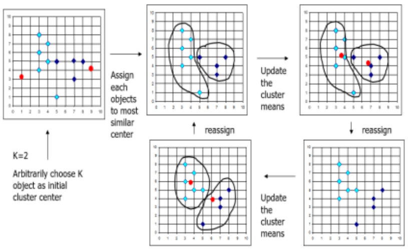
Select *K* points as initial centroids

Repeat

- Form *K* clusters by assigning each point to its closest centroid
- Re-compute the centroids (i.e., *mean point*) of each cluster

Until convergence criterion is satisfied

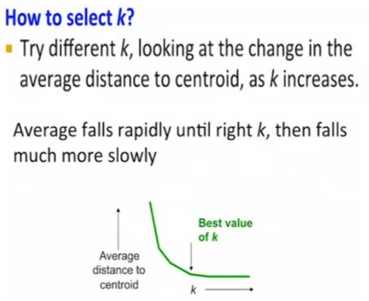
Execution of the K-Means Clustering Algorithm


K=2
Arbitrarily choose K object as initial cluster center
Assign each objects to most similar center
Update the cluster means
reassign
reassign
Update the cluster means

Comments on the *K-Means* Method

- Strength: ***Relatively simple, efficient, and scalable***
- Comment: Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weakness
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k,* the *number* of clusters, in advance
    - There are ways to automatically determine the "*best*" *K*
    - In practice, one often runs a range of values and selected the "*best*" *K* value
  - Sensitive to noisy data and *outliers*
  - Quality of clusters are *sensitive to initial seed points*

How to Select K?



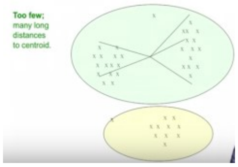
How to select *k*?

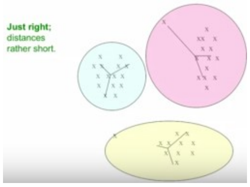
- Try different *k*, looking at the change in the average distance to centroid, as *k* increases.

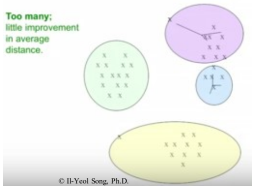
Average falls rapidly until right *k*, then falls much more slowly

https://www.youtube.com/watch?v=RD0nNK51Fp8

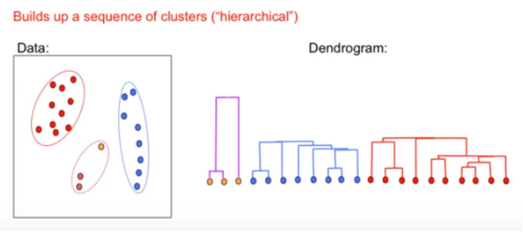
How to Select K?







Example of Hierarchical Clustering

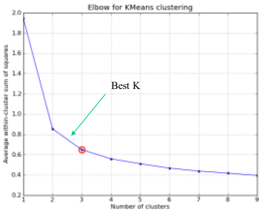


Builds up a sequence of clusters ("hierarchical")

https://www.youtube.com/watch?v=OcoE7JlbXvY&t=226s

Methods for Finding K, the Number of Clusters

- Use domain knowledge for initial k
- **Empirical method**
  - # of clusters: $k \approx \sqrt{n/2}$ for a dataset of n points (e.g., *n* = 200, *k* = 10)
- **Elbow method**: Use the turning point in the curve of the sum of within cluster variance with respect to the # of clusters
  - Try different k, looking at the change in the average distance to centroid, as k increases

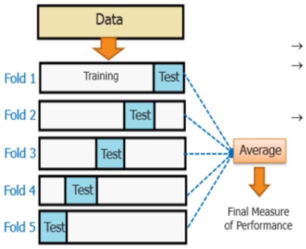


Summary on Clustering

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Two most important clustering techniques are K-Means and Hierarchical clustering techniques.
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches
- There are still lots of research issues on cluster analysis

Validation of ML Models

- How is the model validated?



→ Technique to validate models/classifiers

→ Method to estimate how accurately the model generalizes to unseen data i.e., how well it performs/predicts

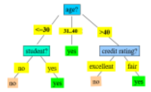
→ K-fold CV
- » Most popular
- » k is typically set to 10
- » Every sample/record is used both in training and test sets

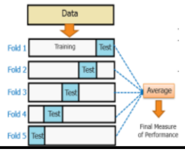
A Summary of ML Process

- Explore data –EDA
- Create a hypothesis or research goals for business questions
- Select a model
- Perform 10-fold validation
- Divide the data into training data set and test data set
- Train the model
- Compute average accuracy
- Interpret the result
- Iterate the process until satisfactory

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Discovery
Data Prep
Model Planning
Model Building
Communicate Results
Operationalize





$F_1 = \frac{2PR}{P+R}$

Question?

Cartoon

"I'M A LITTLE SURPRISED. WITH SUCH EXTENSIVE EXPERIENCE IN PREDICTIVE ANALYSIS, YOU SHOULD'VE KNOWN WE WOULDN'T HIRE YOU."
Klossner
www.jklossner.com