Big Data

Spark ML Algorithms

Spark Machine Learning Algorithms

❖ ML Algorithms

- Basic statistics
  - Correlation
  - Hypothesis testing (P-value)
- Classification and Regression
  - Linear models
    - SVM (Support Vector Machine)
    - Linear Regression, Logistic Regression
  - Naive Bayes
  - Decision tree

Spark Machine Learning Algorithms

❖ ML Algorithms

- Others
  - Collaborative Filtering
  - Clustering
    - *k*-means
  - Dimensionality Reduction
  - etc.

Spark Machine Learning Algorithms

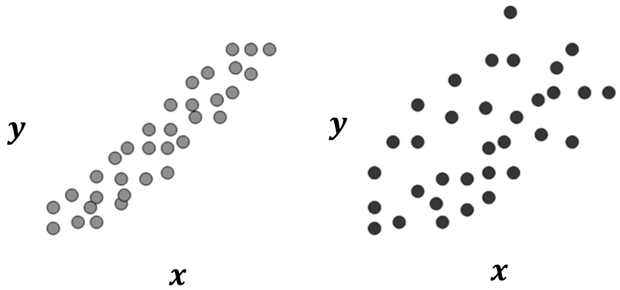
❖ Correlation

- Variables can have a relationship of being
  - Independent
  - Weakly Correlated
  - Strongly Correlated
- Correlation indicates the extent to which variables have an influence on each other (to increase or decrease)

Spark Machine Learning Algorithms

❖ Correlation Example

- In the case of blue dots, x and y are strongly correlated
- On the other hand, the red dots show a weaker correlation between x and y



Spark Machine Learning Algorithms

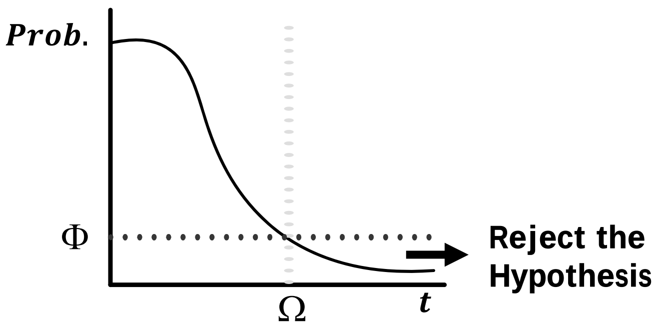
❖ Hypothesis Testing

- Hypothesis testing method
  1. Define a statistic that obeys a certain distribution if the hypothesis is correct
  2. Collect samples and then calculate the statistic probability
  3. If the sample statistics show a probability lower than the threshold of being drawn from this distribution, then the hypothesis is rejected

Spark Machine Learning Algorithms

❖ Hypothesis Testing Example

- If the probability is lower than the threshold Φ (i.e., t is larger than Ω) then the hypothesis is rejected



Spark Machine Learning Algorithms

❖ Linear Models

- Frequently used in Regression
  - SVM (Support Vector Machine)
  - Logistic Regression
  - Linear Regression, etc.
- Regression is a method to predict the value of one (or more) continuous *target* variable y given a (D-dimension vector) input x

Spark Machine Learning Algorithms

❖ Linear Regression

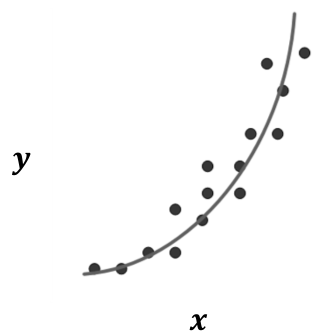
- Linear regression is the simplest regression model
- Output of Linear regression is continuous
- To obtain output y, a linear combination of input variables x_i and weights w_i

$$y(\mathbf{x},\mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

Spark Machine Learning Algorithms

❖ Linear Regression Example

- Example of linear regression
  - Red dots are input data values
  - Blue curve is the Linear regression output model



Spark Machine Learning Algorithms

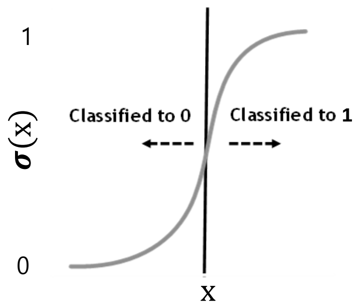
❖ Logistic Regression

- Used in classification problems that need to make a decision
- Decide among two options
  - Example
    - Decide 0 or 1 using a Sigmoid curve
- Decide among multiple options

Spark Machine Learning Algorithms

❖ Logistic Regression Example

- Sigmoid σ(x) S-shape curve with a decision boundary is frequently used to make a decision
  - Small changes on σ(x) near the decision boundary will determine the classification result of 0 or 1



Spark Machine Learning Algorithms

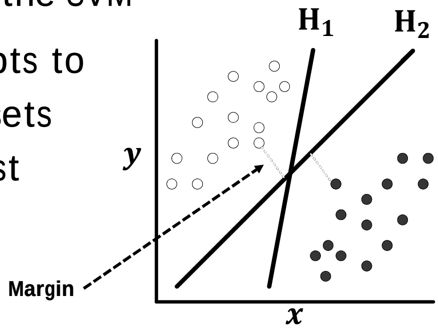
❖ SVM (Support Vector Machine)

- SVM technique is frequently used in solving problems in classification, regression, and novelty detection
- SVM algorithm creates a decision boundary that maximizes the margins between the groups that are being classified

Spark Machine Learning Algorithms

❖ SVM Example

- $\mathbf{H_1}$ represent another algorithm
  - Example: ML (Maximum Likelihood) algorithm
- $\mathbf{H_2}$ represents the SVM
- SVM $\mathbf{H_2}$ attempts to separate datasets with the largest margin



Spark Machine Learning Algorithms

❖ Naive Bayes Classifier

- Conditional independence is assumed to simplify the classification decision
- Bayes Theory is based on conditional probability
  - P(x|y,z) is the conditional probability that x occurs based on the condition that y and z occurred earlier
  - If x is independent of z then P(x|y,z) ➔ P(x|y)

Spark Machine Learning Algorithms

❖ Naive Bayes Classifier

- Everything is dependent to everything else
  - But the relations are too complex to fully analyze
- In order to simplify the computation process, the Naive Bayes model "Naively" assumes that events are independent
  - Pros: Provides fast and easy-to-compute results
  - Cons: Accuracy and reliability is sacrificed
  - Used when the resulting accuracy is sufficient to be applied to its purpose

Spark Machine Learning Algorithms

❖ Naive Bayes Classifier Example

- Conditioned on class c
  1. If it is "assumed" that the probability distributions of the input variables x_1,..., x_D are independent (which they are actually not independent) then…

  2. Class conditional probability density equation can be written as a simple multiplication (product) of one dimensional probability density functions

Spark Machine Learning Algorithms

❖ Naive Bayes Classifier Example

- Lets find the probability of occurrence of dataset $\boldsymbol{x}$ through class $\mathbf{c}$

$$p(\mathbf{x}|y = \mathrm{c})$$

- "Naively" assume conditional independence between the features

$$p(\mathbf{x}|y = \mathrm{c}) = \prod_{\boldsymbol{j}=\mathbf{1}}^{D} \boldsymbol{p}(\boldsymbol{x}_{\boldsymbol{j}}|\boldsymbol{y} = \mathrm{c})$$

  - D is the number of features

Spark Machine Learning Algorithms

❖ Naive Bayes Classifier Example

$$p(\boxed{\mathbf{x}}|y = \mathrm{c}) = \boxed{\prod_{j=1}^{D}} p(\boxed{x_j}|y = \mathrm{c})$$

Multi-dimension

Product of one dimensional densities

- Multi-dimensional probability is easily obtained from multiplications (product) of many one dimensional densities

Spark Machine Learning Algorithms

❖ Naive Bayes Classifier Example

$$p(\boxed{\mathbf{x}}|y = \mathrm{c}) = \boxed{\prod_{j=1}^{D}} p(\boxed{x_j}|y = \mathrm{c})$$

Multi-dimension

Product of one dimensional densities

- This model is called “naive” since in reality these features are not independent
- But it often works very well

Spark Machine Learning Algorithms

❖ Decision Tree

- Local region is identified (classified) in a sequence of recursive splitting decisions in an efficient way
  - Fewer number of processing steps
  - Each step has low computation
  - Results in a hierarchical tree shape form
- Hierarchical algorithm is used in supervised learning (training) of the Decision Tree

Spark Machine Learning Algorithms

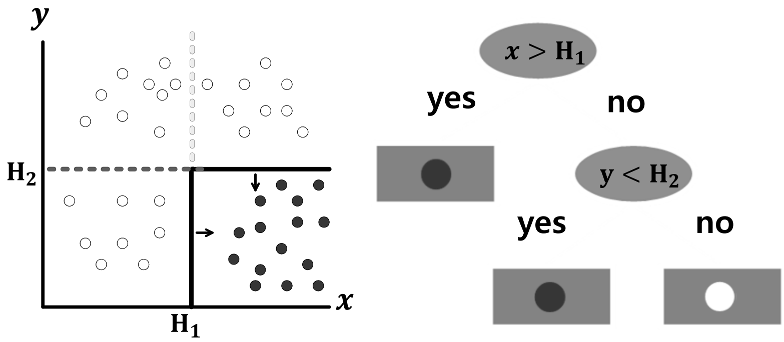
❖ Decision Tree

- Supervised Learning
  - Training a ML (Machine Learning) system with labeled data (desired outputs)
  - Since the desired outputs are known for the inputs (during the training), error values are available for each training step
  - Backpropagation of errors are used in training the ML system to make it more accurate

Spark Machine Learning Algorithms

❖ Decision Tree Example

- For classification, decision boundaries on the dataset (white and red dots) can be determined using a decision tree



Spark Machine Learning Algorithms

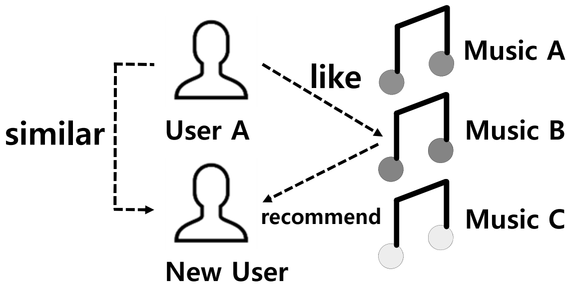
❖ Collaborative Filtering

- ML algorithm that collects preferences or taste information from many users (collaborative) and uses this information to make automatic predictions (filtering) about the interests of other users
  - Collaborative
    - Combining collected information
  - Filtering
    - Filter out less probable options to find the most probable prediction

Spark Machine Learning Algorithms

❖ Collaborative Filtering Example

- Music vendor can recommend music to a New User based on information on User A who's characteristics seem similar



Spark Machine Learning Algorithms

❖ Clustering

- Process of finding similar characteristics in a dataset to form groups of data
- Training data consists of a set of input vectors without any corresponding target values
  - Dataset contains no information (labels) on data and cluster relation
  - Unsupervised Learning is needed

Spark Machine Learning Algorithms

❖ *k*-means Algorithm

- One of the most famous clustering (classification) algorithms
- Unlabeled data is classified to *k* classes
- Mean (average) of each class is updated when new data (vector) is received
- Mean value is used to update the division of the classes (clusters)

Spark Machine Learning Algorithms

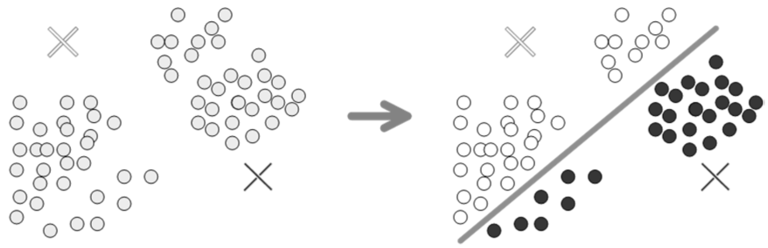
❖ *k*-means Algorithm Example

- Classify unlabeled data into the two classes
  ➔ Red or White
  - All data (vector) is originally colorless

Spark Machine Learning Algorithms

❖ *k*-means Algorithm Example

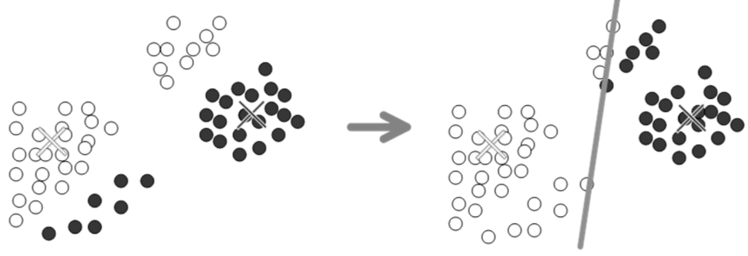
- Step 1: Decide the two centers of the classes randomly
- Step 2: Allocate the data into the nearest center



Spark Machine Learning Algorithms

❖ *k*-means Algorithm Example

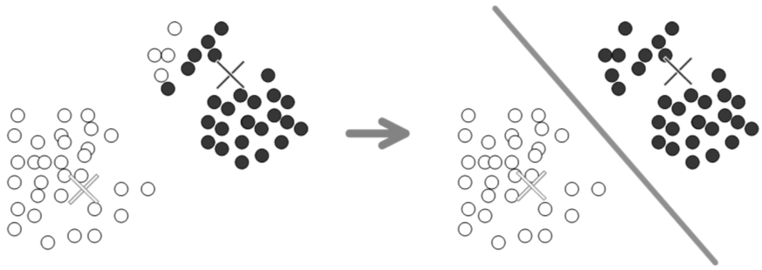
- Step 3: Calculate a mean of each class based on the average distance
- Step 4: Allocate the data into the nearest center



Spark Machine Learning Algorithms

❖ *k*-means Algorithm Example

- Step 5: Calculate a mean of each class again based on the average distance
- Final Step: Allocate the data (vector) into the nearest center's class



Spark Machine Learning Algorithms

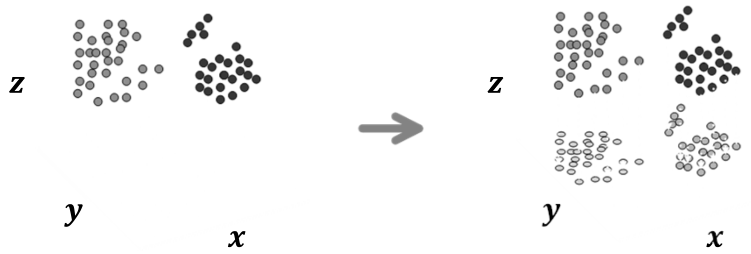
❖ Dimensionality Reduction

- Reduces dimensionality by projecting the dataset to a lower dimensional subspace
  - Captures the essence of the dataset
  - Reduces the complexity of the classifier and regressor
    - Complexity depends on the number of inputs
    - Both the time and space complexity is considered

Spark Machine Learning Algorithms

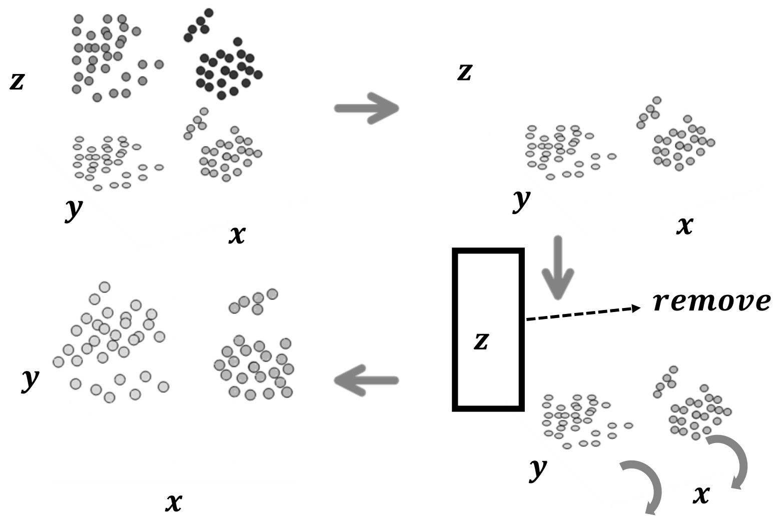
❖ Dimensionality Reduction Example

- Drone management map generation
- 3D (Longitude, Latitude, Altitude) Info
  - Longitude → x, Latitude → y, Altitude → z
- 2D (Longitude, Latitude) Drone Map



Spark Machine Learning Algorithms

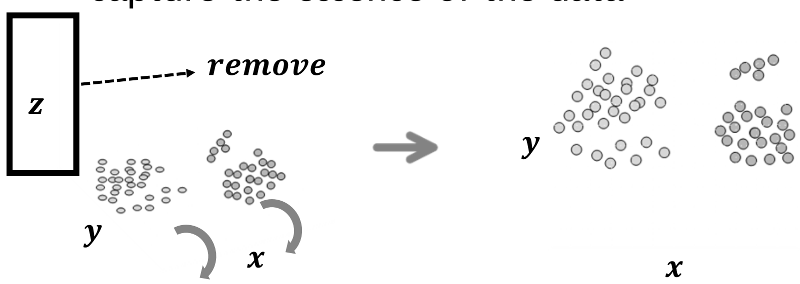
❖ Dimensionality Reduction Example



Spark Machine Learning Algorithms

❖ Dimensionality Reduction - Example

- Assuming that altitude information z is not needed
- By eliminating z from data, we can capture the essence of the data



Spark Machine Learning Algorithms

❖ ML Algorithms Summary

- Basic statistics
  - Correlation, Hypothesis testing (P-value)
- Classification and Regression
  - Linear models (SVM, Linear & Logistic Regression), Naive Bayes, Decision tree
- Others
  - Collaborative Filtering, Clustering (*k*-means), Dimensionality Reduction, etc.

Big Data

References

References

- Spark 2.2.0 Machine Learning Library Guide [Online]. Available: https://spark.apache.org/docs/latest/ml-guide.html