CSC 735 – Data Analytics

Chapter 24

Machine Learning Overview

Dr. Mo Mirbagheri

Advanced Analytics

- Supervised learning
- Unsupervised Learning
- Recommender System
- Graph analytics
- Deep learning

Advanced Analytics Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Supervised Learning

- Most common type of machine learning.
- Goal is to **predict** for each data point based on various features
- Trains over historical data.
- Requires the dependent variable of the historical data (that which you are trying to predict on future data) needs to already be known (i.e. labeled).
- Usually an iterative process, starting on a basic model, making adjustments, with the goal of generalizing it to be able to make predictions.

Classification

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.

"Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar"

Classification: Definition

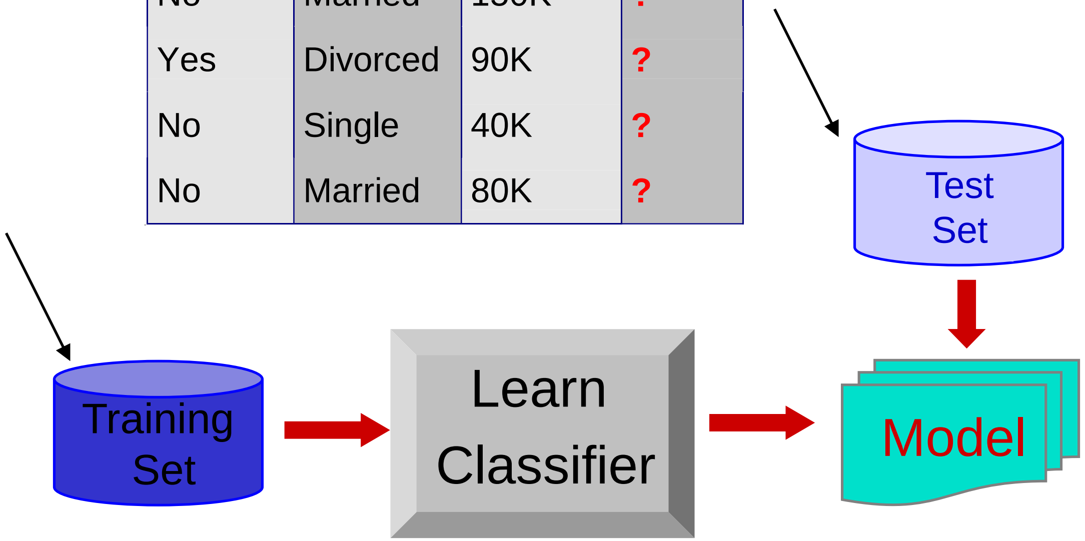
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

“Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar”

Classification Example

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	**Marital Status**	**Taxable Income**	**Cheat**
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



"Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar"

Classification

Binary classification

- Resulting model will make a prediction that a given items belongs to one of two groups

Multiclass classification

- When we classify items into more than just two categories

Classification Applications

- Predicting disease
    - Doctors use a historical dataset of behavioral and physiological attributes of a set of patients to predict whether or not a patient has heart disease
- Classifying images
    - Classify images or label the objects in images
- Prediction customer churn
    - Predict whether current customer likes to stop using a service
- Buy or won’t buy
    - Predict whether visitors of their website will purchase a given product

Regression

Classification vs Regression

Similarities

- Trains on historical data
- Makes predictions on new data

Differences

- Classification predicts a label, i.e. a discrete value
- Regression predicts a real-value

Regression Example

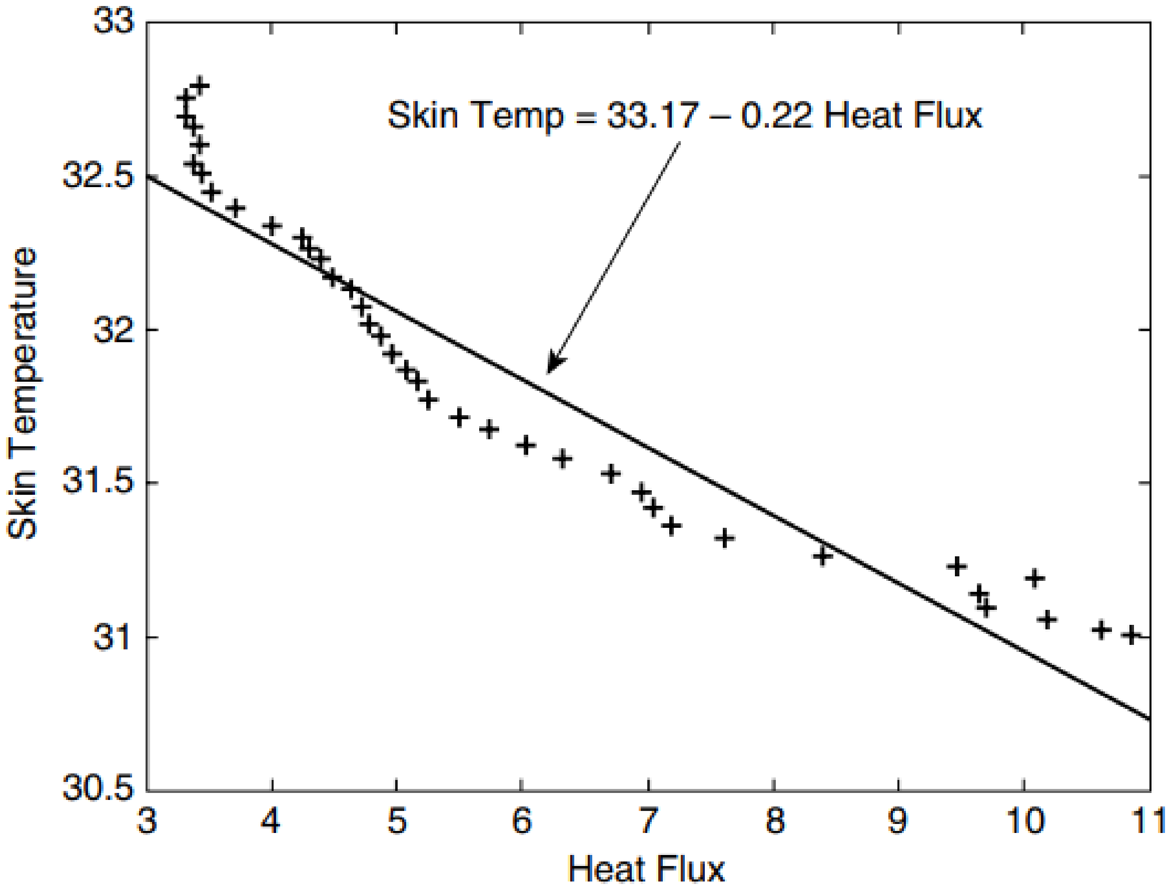


Figure D.2. A linear model that fits the data given in Figure D.1.

"Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar"

Regression Applications

- Predicting Sales
    - Retail companies
- Predicting the number of viewers of a show
    - Entertainment industry
- Predicting Real Estate Value
- Financial Forecasts
    - Stock price
- Weather Forecasts
    - Temperature

Unsupervised Learning

Unsupervised Learning

- Different from Supervised Learning:
  - No Dependent Variable (no class label)
- Goal:
  - Find Patterns
  - Discover Underlying Structure

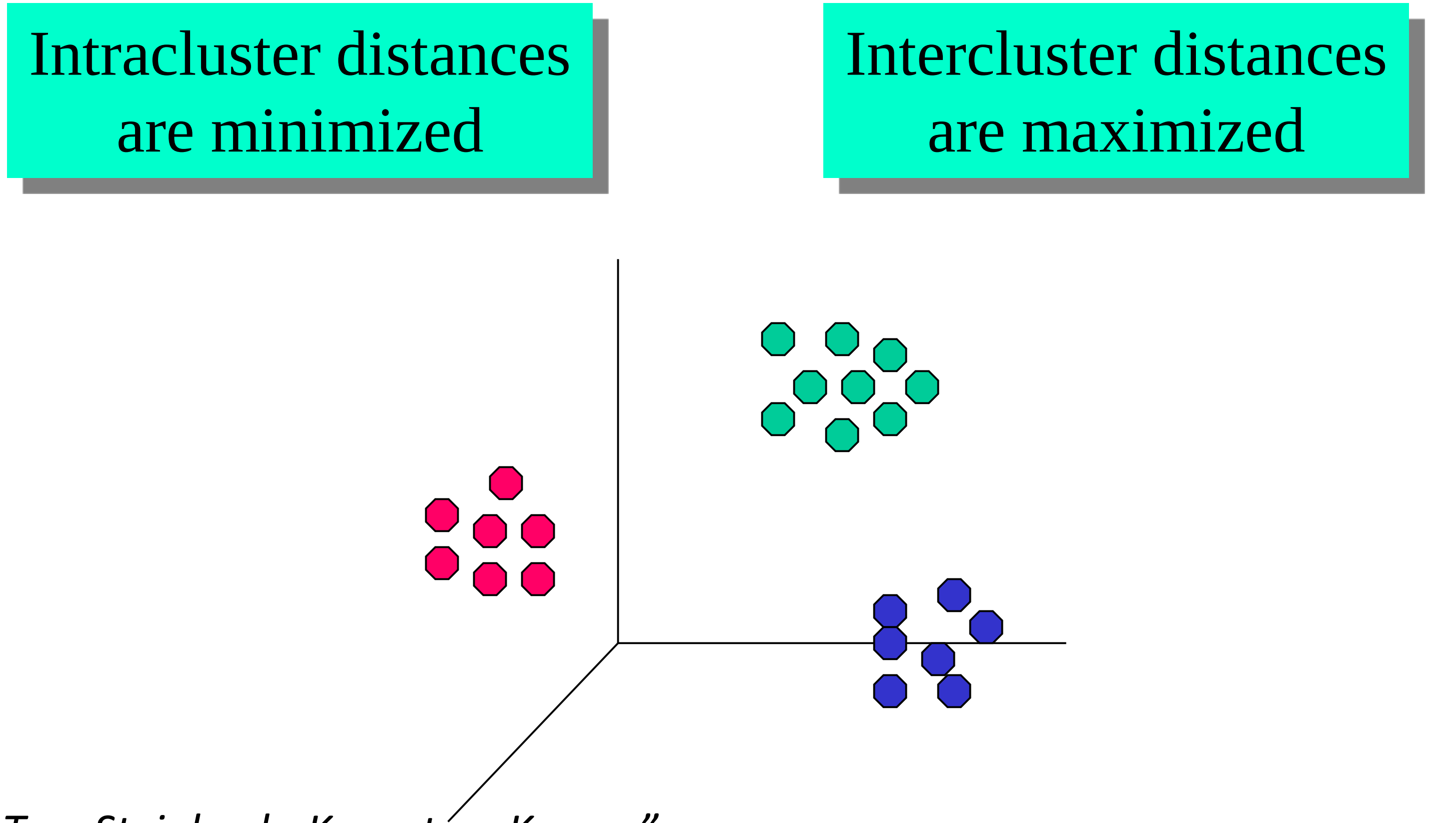
Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

“Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar”

Illustrating Clustering

- ☒Euclidean Distance Based Clustering in 3-D space.



Intracluster distances are minimized

Intercluster distances are maximized

“Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar”

Clustering Applications

- User Segmentation
  - Given a set of behaviors, casual vs frequent users/customers might display different behavior trends.
  - E.g., observing buying patterns of customers in same cluster vs. those from different clusters.

- Topic Modeling
  - Given a set of web pages on topic of data science, they could be clustered into pages about ML, SQL, streaming, etc… based on groups of shared/frequent vocabulary.

Anomaly Detection

- Goal:
  - find objects that are different from most other objects.
  - anomalous objects are known as **outliers**, since, on a scatter plot of the data, they lie far away from other data points.

Anomaly Detection Applications

- Fraud Detection
  - Purchasing behavior of someone who steals a credit card is **different** from that of the original owner
- Intrusion Detection
  - Can be detected by monitoring systems and networks for **unusual** behavior
- Medicine
  - Anomalous spot in an image may indicate a potential health problem for a patient

Recommender System

Recommender System

- **Goal**: Find similarities between the users or items
- Study people's explicit preference (through rating) or implicit ones(though observed behavior) for various products or items
- Algorithms make recommendations to users based on what similar users liked, or what other products resemble the ones the user already purchased

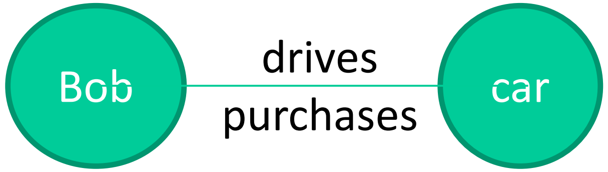
Recommender System Applications

- Movie Recommendation
    - Netflix uses Spark to make large-scale movie recommendation to its users
    - It does this by studying what movie users watch and do not watch in the Netflix
- Product Recommendation
    - Amazon uses product recommendation as one of its main tools to increase sales
    - Based on the items in our shopping cart, Amazon may recommend other items that were added to similar shopping cart in the past

Graph Analytics

Graph Analytics

- Analyzing graph networks data including vertices (objects) and edges (relationships) and solving machine learning related tasks
- Ex.

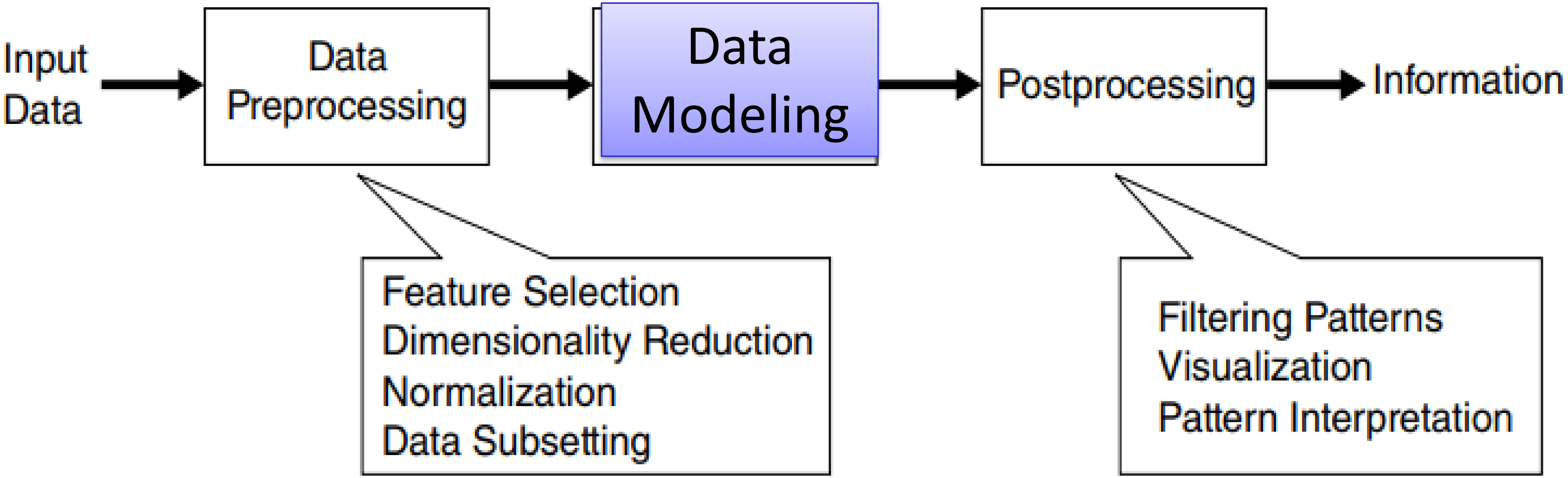


Graph Analytics Applications

- Anomaly detection
  - if typically in our data each vertex has ten edges associated with it and a given vertex only has one edge, that might be worth investigating as something strange
- Classification
  - If certain vertices have a known label, then other similar vertices with similar structure could be classified under the same label.
- Recommendation
  - Google’s original PageRank algorithm.

Advanced Analytics Process

General Process



"Introduction to Data Mining by Tan, Steinbach, Karpatne, Kumar"

General Process

1. Collect relevant data
2. Clean, inspect, and understand the data
3. Feature engineering (ex. convert to numerical form)
4. Select some training data
5. Train potential models
6. Evaluate and compare models on testing data
7. Gain insights on the data, or utilize the model for use on future data (ex. anomaly detection, predictions, solve business problems)

Data Collection

- Gather data sets
- Can be many files, differing file types, and different sizes

Data Cleaning

- Inspecting a.k.a Exploratory Data Analysis (EDA):
  - Use interactive queries and visualization methods to understand distributions, correlations, and other details

Data Cleaning

- Inspecting a.k.a Exploratory Data Analysis (EDA):
  - Use interactive queries and visualization methods to understand distributions, correlations, and other details
- Cleaning:
  - Take notice if some values need to be removed, or have been recorded incorrectly upstream, or simply missing.
- Spark has many functions for this process

Feature Engineering

- A process that involves creating new features (variables) from existing data to improve the performance of a predictive model or to enhance the understanding of the data.
- Tasks can include
  - Normalizing data
  - Adding variables to represent interactions of other variables
  - Manipulating categorical variables
  - Converting them to a proper format for the ML model
- In MLlib, all variables will usually have to be input as vectors of doubles (regardless of what they actually represent)

Feature Engineering Activities

- **Feature Creation**: Generating new features from existing ones through mathematical operations, aggregations, or transformations.
- **Feature Selection**: Choosing the most relevant features and removing irrelevant or redundant ones.
- **Handling Categorical Data**: Converting categorical variables into numerical representations (e.g., one-hot encoding).
- **Handling Missing Data**: Dealing with missing values in a dataset through imputation or other strategies.
- **Feature Extraction**: Transforming complex data (e.g., images, text) into meaningful numerical features.
- **Interaction Features**: Creating new features that represent interactions between existing features.

Training Models

- At this point, we have:
  - A clean **dataset** of historical information(e.g. spam or not spam email)
  - A **task** we would like to complete(e.g. classifying spam emails)

Training Models

- At this point, we have:
  - A clean **dataset** of historical information(e.g. spam or not spam email)
  - A **task** we would like to complete(e.g. classifying spam emails)
- Next step: **train** a **model** to predict the correct output, given some input

Training Models

- At this point, we have:
  - A clean **dataset** of historical information(e.g. spam or not spam email)
  - A **task** we would like to complete(e.g. classifying spam emails)
- Next step: **train** a **model** to predict the correct output, given some input.
  - During the training process, we apply an ML algorithm to the data
  - To classify spam emails, our algorithm may find that certain words are better predictor of spam than others and therefore assign higher weights to those
- The output of training process is what we call a model

Data Partitioning and Evaluation

- How do we know our model is any good at what it is supposed to do?

Data Partitioning and Evaluation

- Splitting our dataset into multiple portions
  - **Training dataset** for training models
  - **Validation** set in order to try out different hyperparameters (parameters that affect the training process) and compare different variations of the same model (the validation set helps in **preventing overfitting**)
  - **Test** dataset for the **final evaluation** of our different model variations to see which one performed the best
  - **Overfitting** means that the model does not generalize well to **new** data.

Data Partitioning and Evaluation

- Various performance metrics can be used to evaluate the models
- They are relevant to the specific problem including:
  - Accuracy
  - Precision
  - Recall
  - F1-score
  - Mean squared error (MSE)
  - other methods