ALGONQUIN COLLEGE

CST8502
MACHINE LEARNING

Week 1
Introduction to Machine Learning

Professor: Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T315

Data Analytics

- Definition: Data Analytics is the process of aggregating large data sets in order to detect underlying patterns that might not be visible by just looking at raw data.
- These patterns give insight to maximize profits, improve health, lower electricity usage, etc.
- Enables businesses to make data-driven decisions

Mountains of Data

- We now have more data being gathered/collected. Governments are starting to adopt openness policies of making public data freely available on the internet.
- Canada Open Government: http://open.canada.ca/en
- Seattle Open Data https://data.seattle.gov/
- Ontario Open Data https://www.ontario.ca/search/data-catalogue
- Ottawa Open Data: http://data.ottawa.ca/

Seattle Bicycle Traffic

- https://data.seattle.gov/Transportation/daily-bike-traffic/d4dx-u56x
- A traffic counter counts number of bicycles on the East and West sidewalks.
- There are traffic spikes from 7-9 am on the West side, and from 5-6pm, but only 5 days a week.

Profit

- As an entrepreneur, where would you sell hot dogs, or advertise?

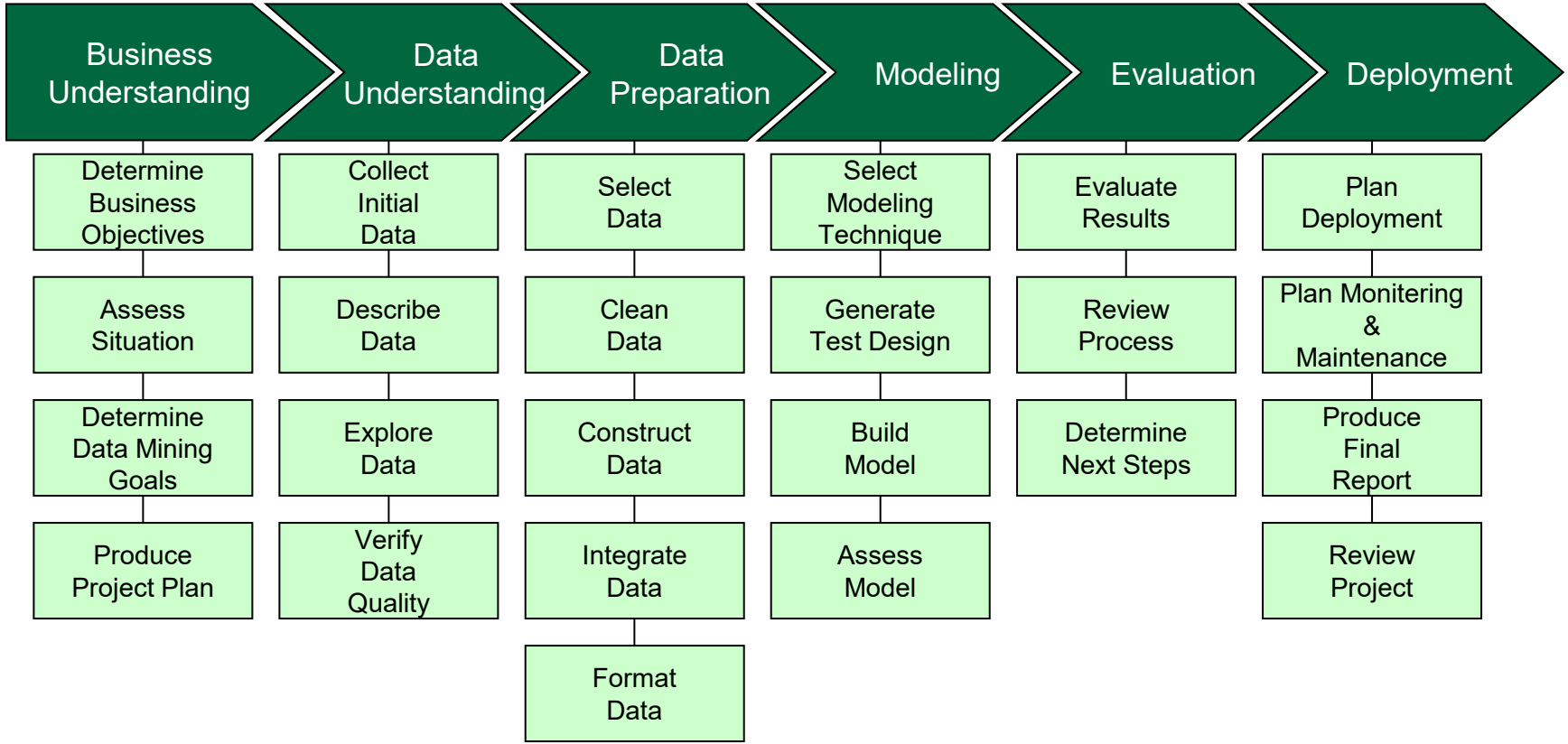
http://www.blogto.com/eat_drink/2015/07/everything_to_know_about_hot_dog_stands_in_toronto/

Healthcare

- The Real-World Benefits of Machine Learning in Healthcare
- Machine Learning Healthcare Applications – 2018 and Beyond
- Machine Learning in Healthcare

CRISP-DM

- Cross-industry Standard Process for Data Mining
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objectives
Assess Situation
Determine Data Mining Goals
Produce Project Plan
Collect Initial Data
Describe Data
Explore Data
Verify Data Quality
Select Data
Clean Data
Construct Data
Integrate Data
Format Data
Select Modeling Technique
Generate Test Design
Build Model
Assess Model
Evaluate Results
Review Process
Determine Next Steps
Plan Deployment
Plan Monitering & Maintenance
Produce Final Report
Review Project

Business Understanding

What does the business need?

- Determine business objectives - Identify what you want to accomplish from a business perspective
- Assess situation – determine availability of resources, assess risks, think about contingency plans for risks etc.
- Determine goals
- Produce project plan

Data Understanding

- What data do we have/need? Is it clean?
  - Collect initial data
  - Describe data – check quantity of data, data format, consistent coding schemes
  - Explore data – visualize data, identify relationships among data, query data etc.
  - Verify data quality – how clean is the data? Any noise?
    - Missing data, errors, measurement errors, inconsistent representation etc.

Data Preparation

How can we organize the data to perform modeling?

- Select data – determine which data will be used and document reasons for inclusion/exclusion
- Clean data – solve all data quality issues
- Construct data – derive new attributes
- Integrate data – create new datasets by combining data from multiple sources
- Format data – re-format data as necessary (discretization etc.)

Data Preparation

- The lengthiest process!
- When getting data from different sources, some work is needed when putting it together:
  - Cleaning and filtering: Remove duplicate data, missing data, resolve incomplete data. Something like: ***Woodroffe Ave, Woodroffe, Woodroffe Avenue*** should all be the same.
  - Remove outliers: (data that is far outside the average). Every semester, some students register for a course but don't drop it. This means they get 0 for everything and lowers the class average. Another example is that sales for a store is $0 for some regional holidays.
  - Variable transformations. Changing how variables are represented (metric / imperial)

Modeling

What modeling technique should we apply?

- Select modeling technique – determine which algorithms to try
- Generate test design – how to split data for training, testing, validation etc.
- Build model – build decided models
- Assess model – check generated models, apply domain knowledge to interpret the results

Evaluation

Which model best meets the business objectives?

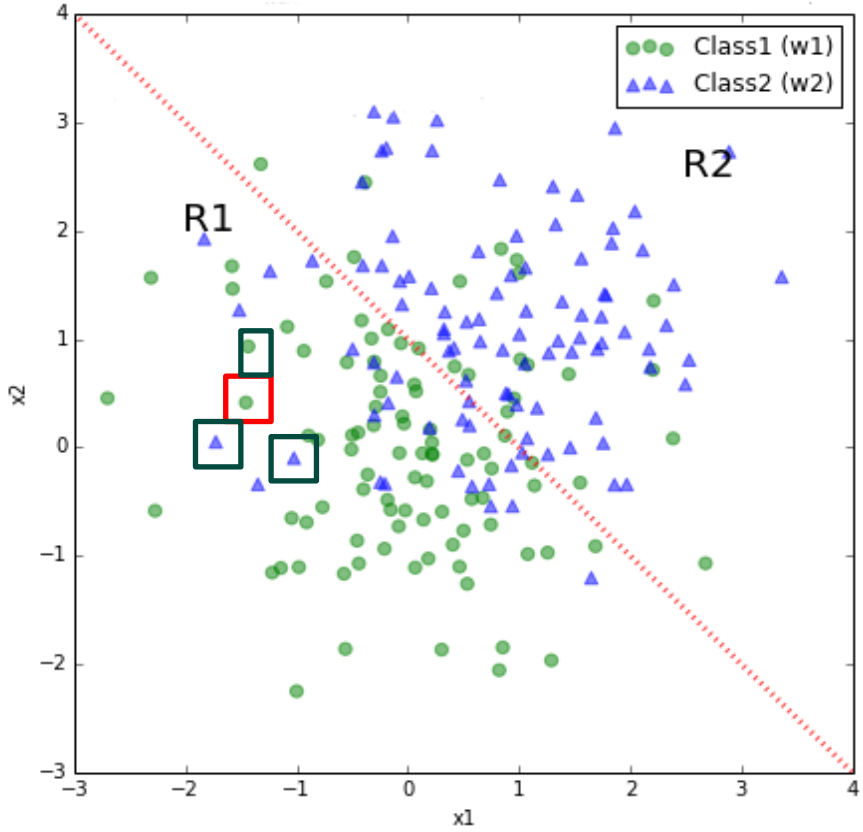
- Evaluate results – do the models meets the business criteria? Which ones should we approve?
- Review process – Review the work.
- Determine next steps – determine whether to proceed or iterate further etc.

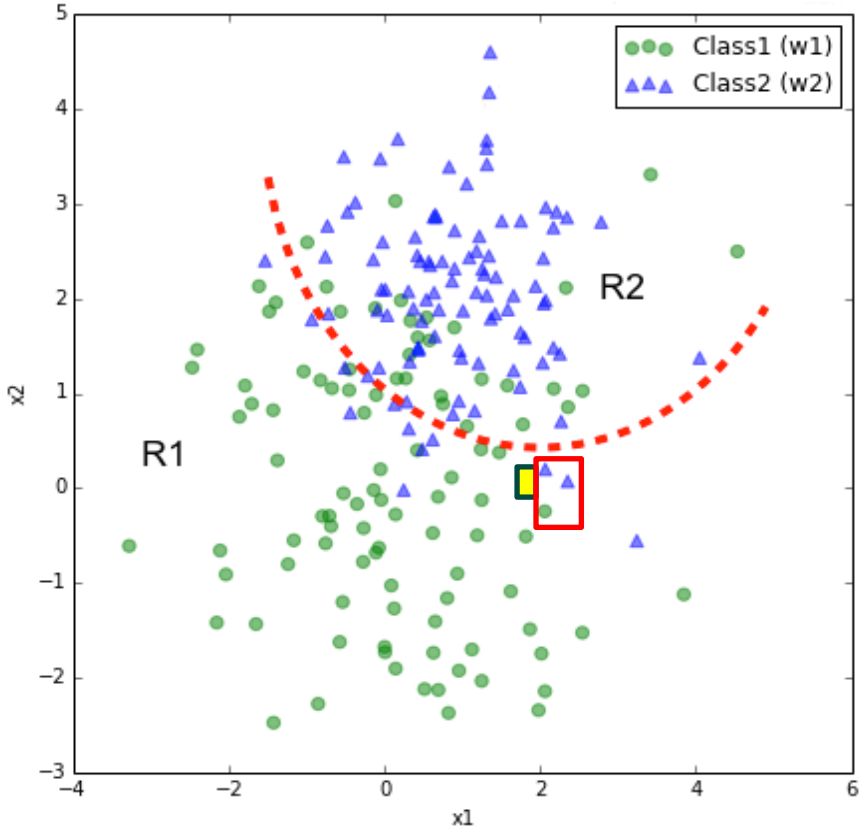
Machine Learning

- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
  - Outlier detection

Supervised learning: Classification

- Data has class labels
- Based on the labels, classifiers are generated
- New data will be classified based on the generated classifier
- Predicts a **discrete** class label
- Example 1: Cancer dataset – Malignant and benign labels are present for each instance.
- Example 2: Iris dataset – data from 3 types of flowers – every instance has a class label





https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Supervised Learning: Regression

Regression predicts **continuous** values (numbers) as the output.

Example, housing prices for various houses: number of bedrooms, garage size, property size, and the computer must interpolate predictions.

Unsupervised Learning

- Data has no class labels
- Clustering: tries to group instances
  - Similar instances grouped together to form clusters. (Ex. Insurance: grouping groups of motor insurance policy holders with a high average claim cost)
- Outlier detection: tries to find anomalies
  - Identify those instances which are distinct from the nature of the majority of instances. (Ex. Financial fraud detection)