email: abran012@uottawa.ca

SYSC 4700 - Lecture 10
Carleton University

Part 1: *Introduction to Big Data Analytics*

Part 2: *Introduction to Machine Learning*

Presenter: André Brandão
abran012@uottawa.ca
February 08, 2018

email: abran012@uottawa.ca

SYSC 4700 - Lecture 10 - Part 2

What is "machine learning?"

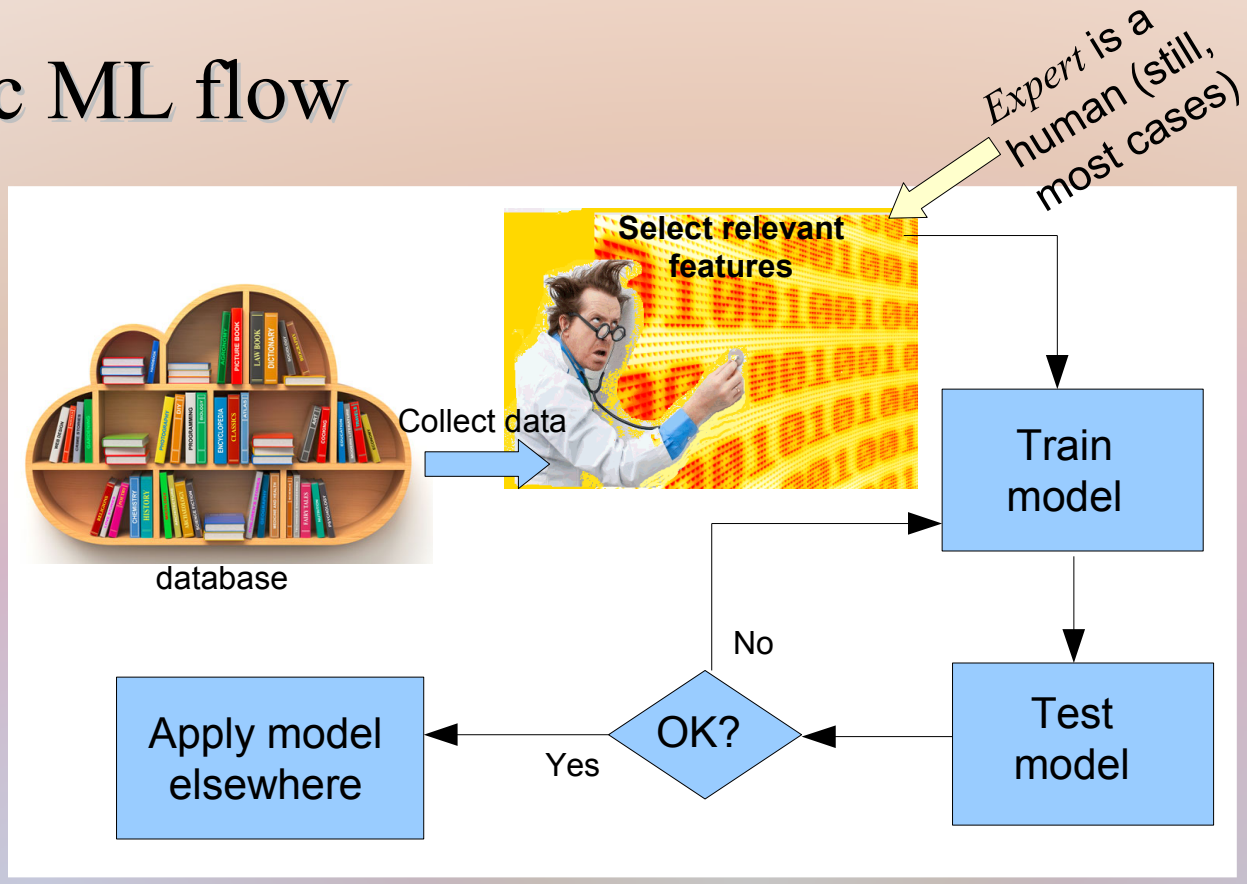
We will discuss:

- Concepts.
- ML in *Prediction* and *Classification*.
- Examples.
- ML, AI and Big Data.

email: abran012@uottawa.ca

ML Concepts

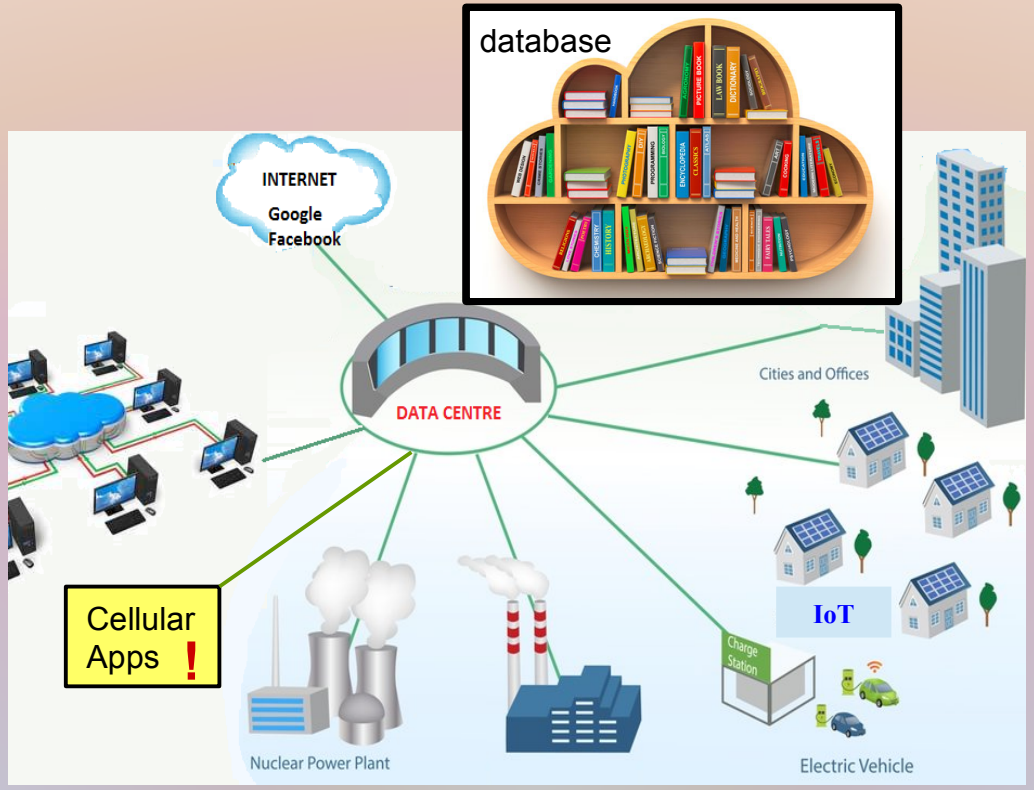
Basic ML flow



email: abran012@uottawa.ca

ML Concepts

Database



email: abran012@uottawa.ca

ML Concepts

Domain Expert

- There is no deterministic method capable of finding a thesis that proves a given hypothesis.
- For this reason, the search for cause & effect starts with the intuition of the expert, the "Eureka" moment.
- ML is a tool in this process. It allows us to test our hypothesis faster than ever.

*"**Modelling**" in ML culture is the process of applying software methods to classify and predict features labelled as significant (by a domain expert) in the quest for knowledge of cause & effect.*

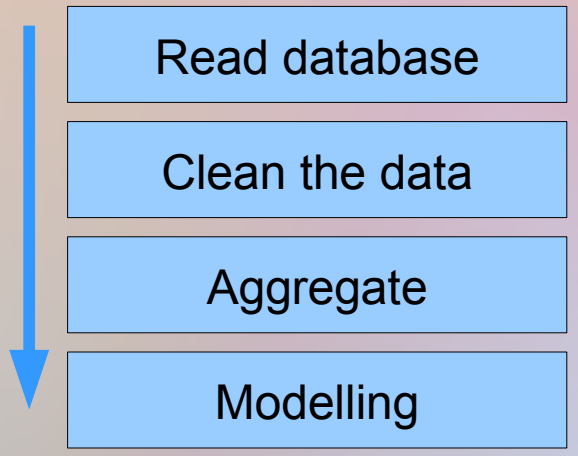
email: abran012@uottawa.ca

ML Concepts

Modelling

There is a large array of development platforms, frameworks and computer languages containing classification and prediction methods for big data.

- SparkR/Sparklyr with Rstudio and R;
- Pyspark and Python;
- Cloudera machine learning;
- Hortonworks;
- Databricks;
- Spark for Java, Scala, just to name a few.



email: abran012@uottawa.ca

ML Modelling

Linear Regression: Prediction

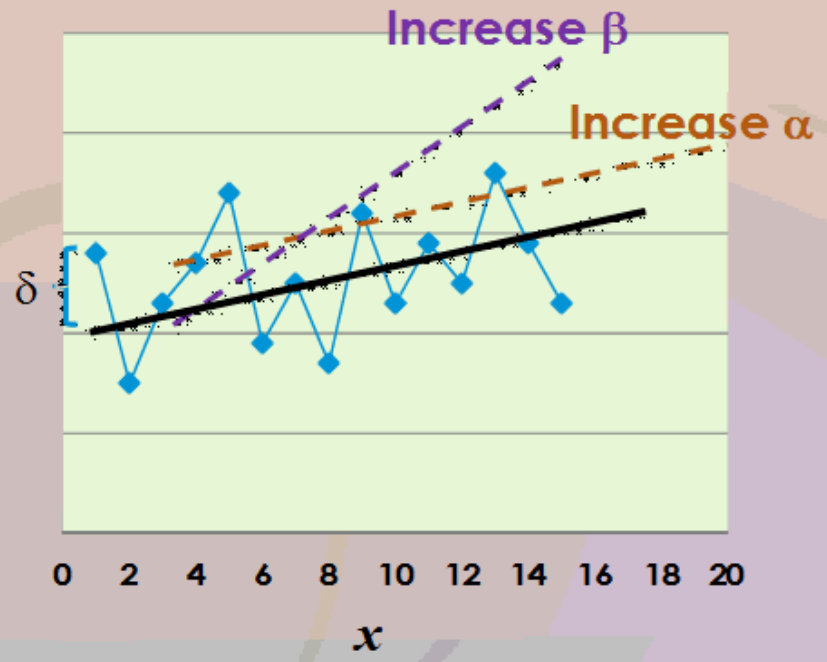
$$y_i = \alpha + \beta x_i + \delta_i$$

y: target variable
x: feature variable
δ: residual term

Prediction

$E[y_i \mid x_i]$ expected y given x :

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$



email: abran012@uottawa.ca

ML Modelling

Linear Regression: Prediction

Goal is to minimize the mean-squared error: $\min\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2\right)_{\hat{\alpha},\hat{\beta}}$

Where $\varepsilon_i = y_i - \hat{y}_i$

Solution: $\hat{\beta} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$ $\hat{\alpha} = \mathrm{avg}(\hat{y}) - \hat{\beta} \times \mathrm{avg}(x)$

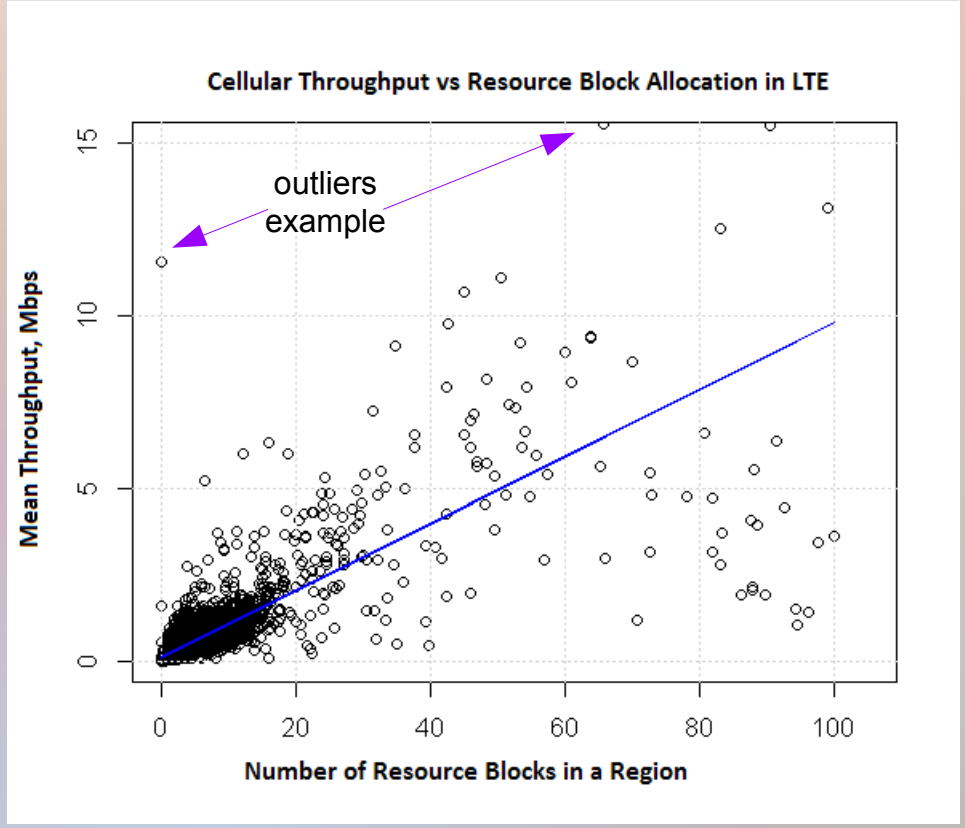
After we compute $\widehat{\propto}$ and $\hat{\beta}$ we estimate $\widehat{y_i}$ for a given x_i.

$\{x\}$ and $\{y\}$ are past data (feature and target columns) from the data base.

email: abran012@uottawa.ca

ML Modelling

Generalized Linear Models
- GLM library: comes with SparkR package



Read database

Clean the data

Aggregate

Apply the model

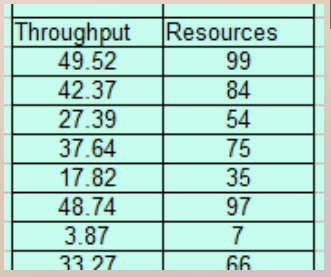
Throughput vs.
Resource allocation

One resource block in LTE:

- 12 subcarriers,
- 0.5 *ms* duration
- Maximum of 100 resource blocks/carrier/20MHz BW

email: abran012@uottawa.ca

ML Modelling



Throughput	Resources
49.52	99
42.37	84
27.39	54
37.64	75
17.82	35
48.74	97
3.87	7
33.27	66

Read database

70% training

30% testing

$$\hat{\beta} = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

$$\hat{\alpha} = \text{avg}(y) - \hat{\beta} \times \text{avg}(x)$$

Training = obtain a model

Testing = execute a prediction.

$E[y_i \mid x_i]$ expected y given x :

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Feature (x)

Target (y)

Verify the error variance

$$\varepsilon_i = y_i - \hat{y}_i$$

email: abran012@uottawa.ca

ML Modelling

Example with SparkR syntax, once you have read the “inputData” csv file:

```
df_list <- randomSplit(inputData, c(0.7, 0.3),0)
training_df <- df_list[[1]]
test_df <- df_list[[2]]
```

```
linearModel <- glm(throughput ~ rbUsage,
                   training_df, family = “gaussian”)
```

```
result <- predict(linearModel,test_df)
```

email: abran012@uottawa.ca

ML Modelling

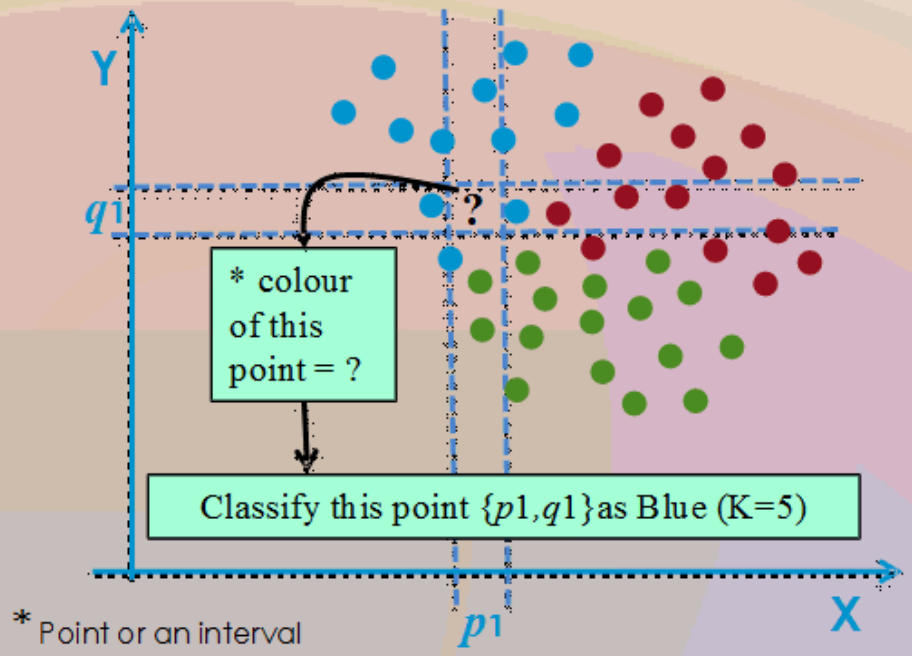
K-NN: Classification

Training Phase

- Choose a {X,Y} training set.
- The Training set becomes a reference or (blue-print) for subsequent classifications.

Test Phase

A point {p,q} from the test set is compared with the K nearest points from the reference {X,Y}.



Majority vote decides the colour of {p,q}.

email: abran012@uottawa.ca

ML Modelling

K-NN: Classification

Distance calculation

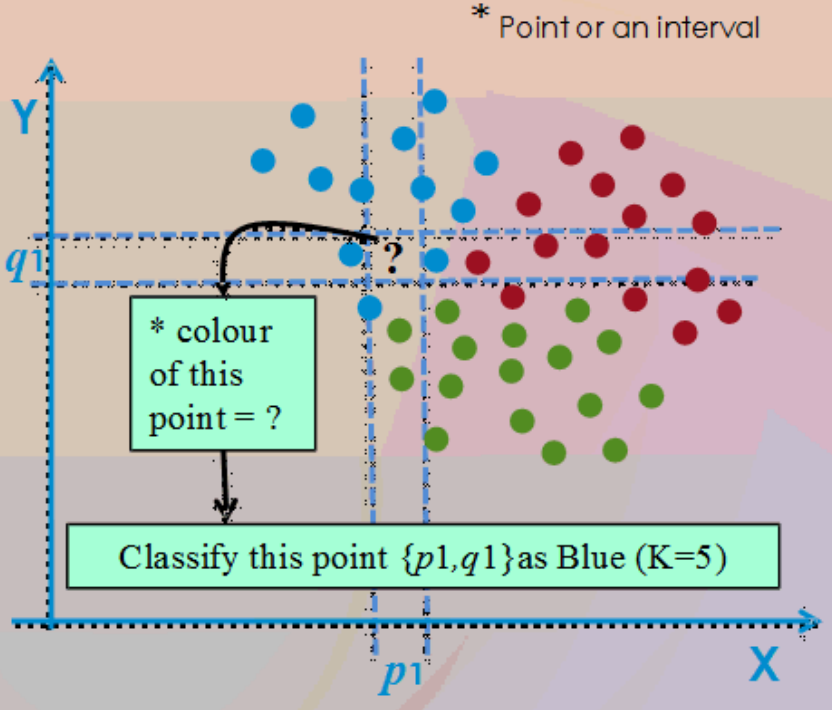
Euclidean:

$$D(m,r) = \sqrt{\sum_{i=1}^{N} (m_i - r_i)^2}$$

where $r_i = \sqrt{x_i + y_i}$; $m_i = \sqrt{p_i + q_i}$

N is dimension (N=1 for single feature)

Other distances:
Manhattan; Minkowski; Canberra.



email: abran012@uottawa.ca

ML Modelling

K-Means: Clustering

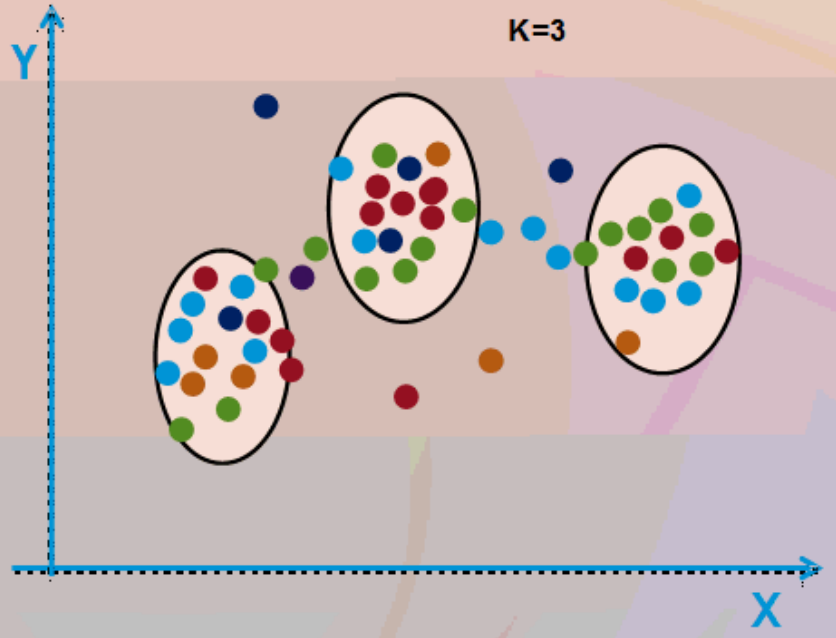
Basic idea of clustering

- It ignores classes (colours of dots here).
- It focuses on **Densities**.

Example shows 3 clusters (K=3) or distinct areas.

- Clustering is an unsupervised process

No training



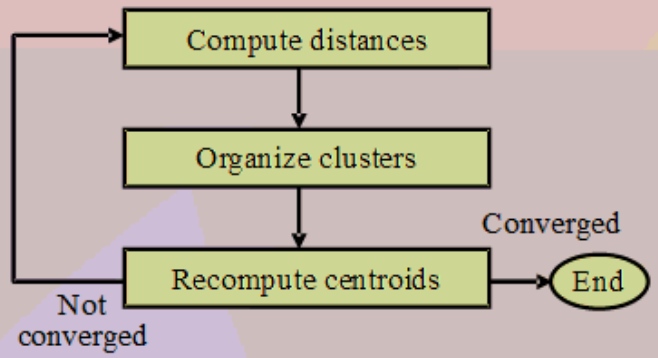
email: abran012@uottawa.ca

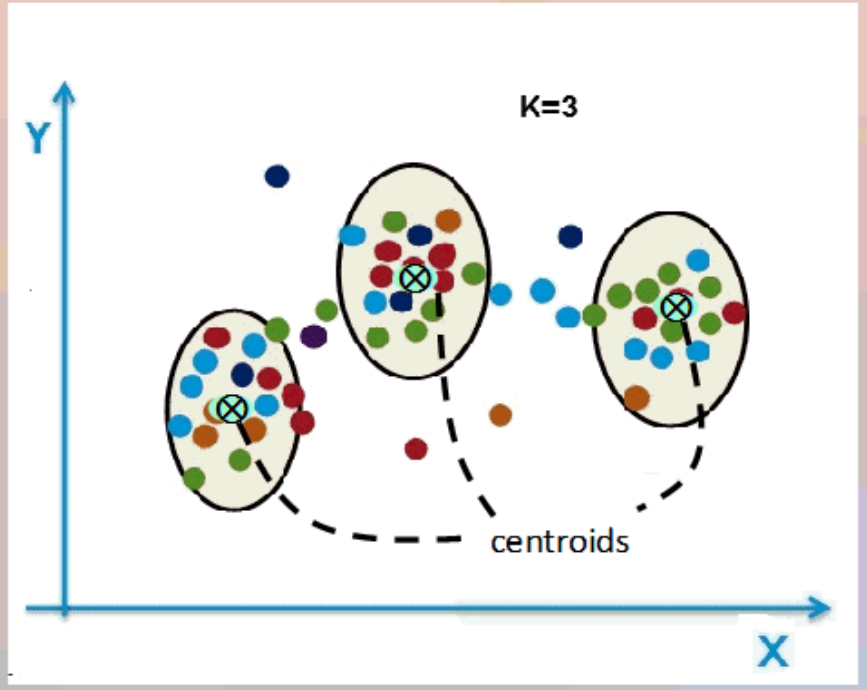
ML Modelling

K-Means: Clustering

Each cluster has one centroid.

The nearest centroid defines the membership of a data point.





Recompute *centroids = find the mean distance values of the elements for each cluster.*

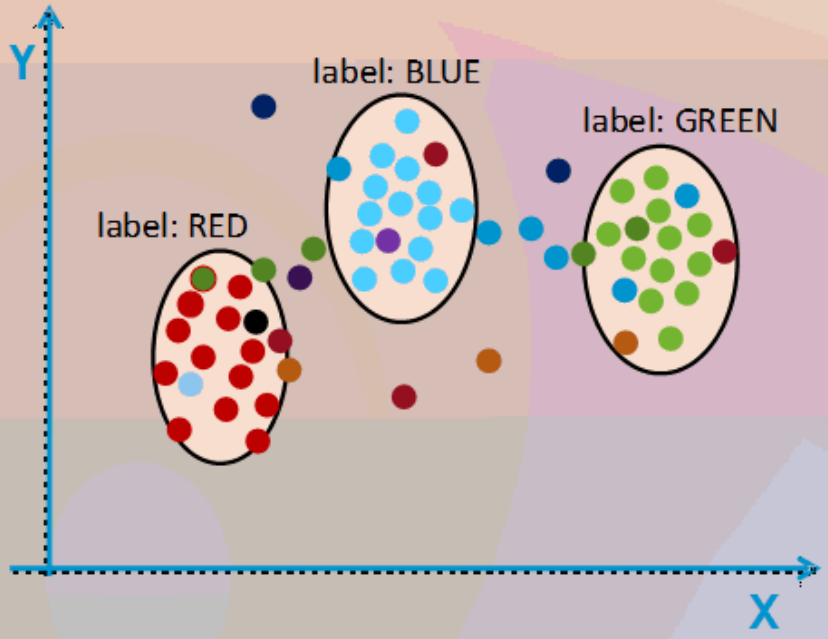
email: abran012@uottawa.ca

ML Modelling

K-Means: Classification

These clusters contain unique elements

- Data is classified according to the membership of associated clusters.
- This is **Supervised** algorithm: training provides the membership label of each centroid.



Classification is not possible with pure clustering.

email: abran012@uottawa.ca

ML Modelling

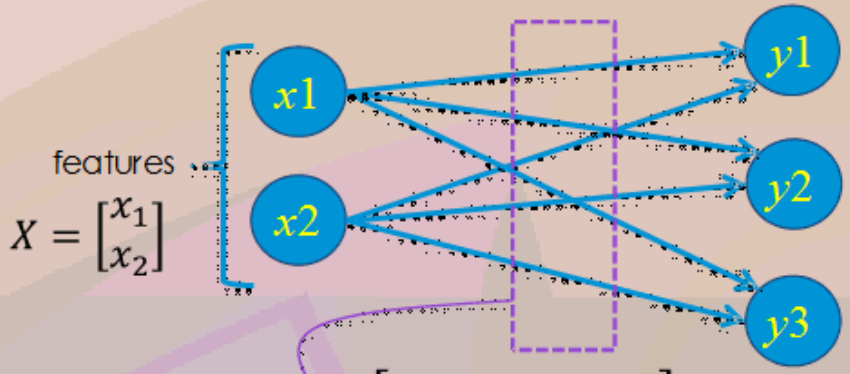
Other Examples of Classification Methods

- Naive Bayes
- Linear Discriminant Analysis
- Linear Logistic Regression
- Support Vector Machines
- Random Forests
- Neural Networks

email: abran012@uottawa.ca

ML Modelling

A Glimpse into Neural Networks



features

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$W^{(1)} = \begin{bmatrix} w_{1,1}^{(1)}, w_{1,2}^{(1)}, w_{1,3}^{(1)} \\ w_{2,1}^{(1)}, w_{2,2}^{(1)}, w_{2,3}^{(1)} \end{bmatrix}$$

connection weights for layer (1)

Output is:

$$Y = g(WX + C)$$

where

$$C = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

This is a Constant (bias)

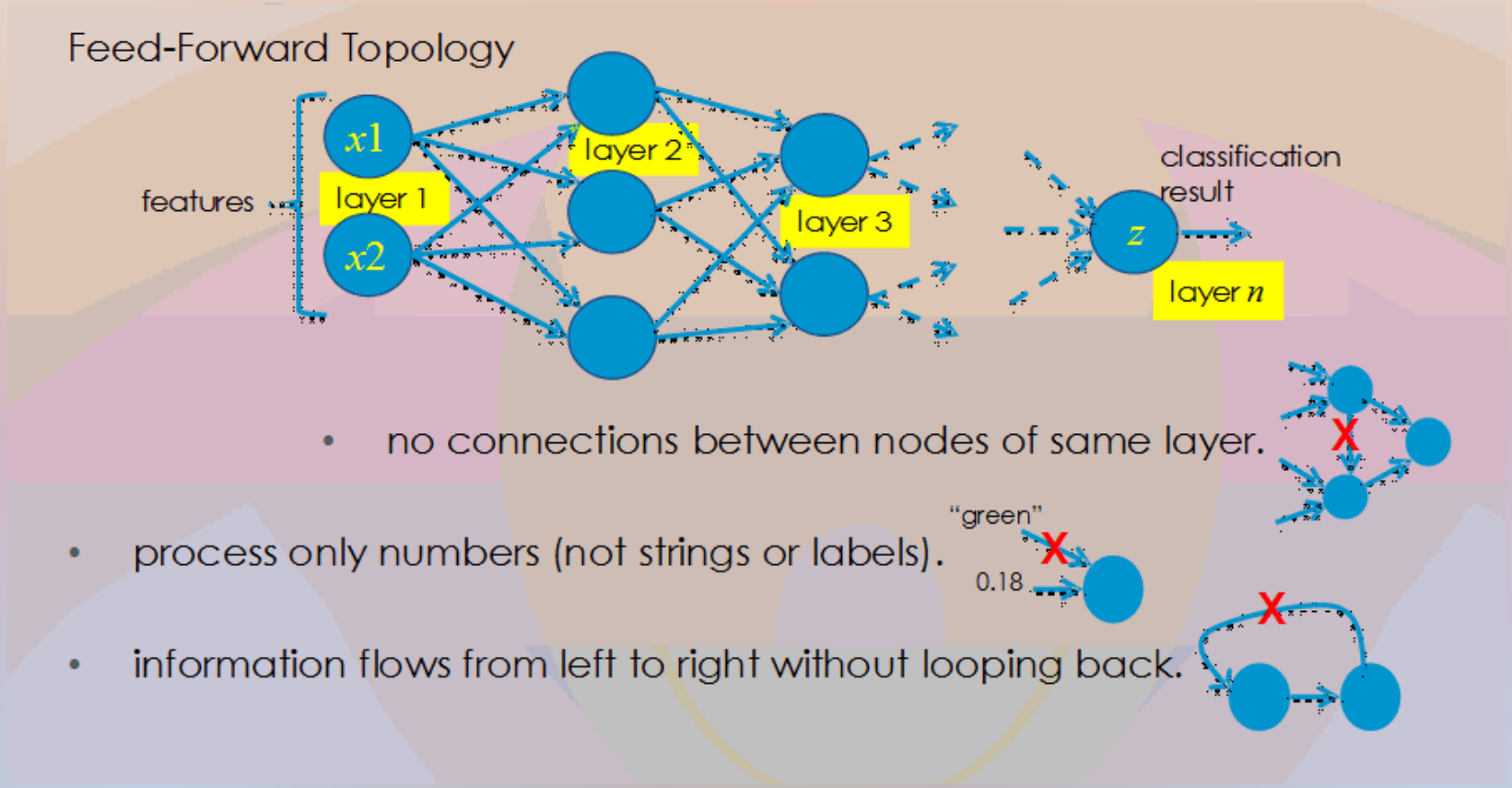
and $g(\cdot)$ is a function named *"activation function"*

The *activation function* can be anything, but it is common for $g(\bullet)$ to be a sigmoid transformation.

email: abran012@uottawa.ca

ML Modelling

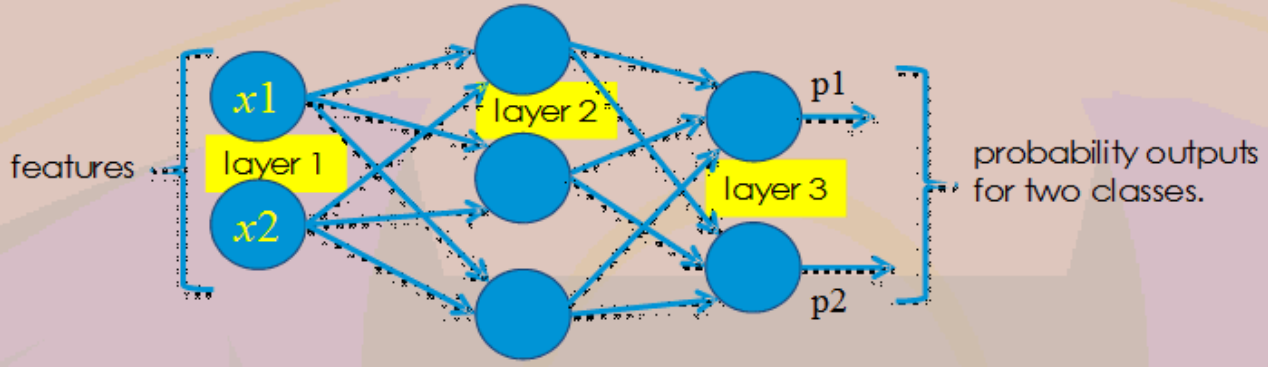
A Glimpse into Neural Networks



email: abran012@uottawa.ca

ML Modelling

A Glimpse into Neural Networks



- p1 and p2 represent probabilities for two classes of objects.
- two classes: p1+p2=100%. If they were 3 classes: p1+p2+p3=100%
- (x, y) are likely features of an object of class #1 if p1> p2 and vice-versa.
- some applications compute pi as a <u>score prediction</u> (as opposed to a probability number). Example of binary output, "rainy day p1=1", "if sunny p1=0"; "if windy p2=1", etc. , for a given set of features.

email: abran012@uottawa.ca

ML Modelling

Nature of data: qualitative and quantitative

A DATA SET {1,2,3,3,4,4,4,4,4,5,6,7,1,1,2,9,12,15,20}

Assume this is a representative sample so that bigger sample sizes produce similar statistics.

- "what is the probability that '1' is collected from the box?".
  - The answer is quantitative: P(x=1) is 3/19≈15.79%.
- "which number has the greatest frequency?"
  - The answer is also quantitative: the number 4 has P(x=4) =5/19≈26.32%.

email: abran012@uottawa.ca

ML Modelling

Nature of data: qualitative and quantitative

DATA SET X = {F150, Silverado, ..., ..., Corolla, Corolla, ..., ..., , F150,F150, Silverado, ..., ...}

Now, change the numbers in X by a string such that 1 = “Ford F150”, 2=”Chevy Silverado”, 4=”Toyota Corolla”, etc.

- “what is the probability that the ‘Ford F150’ is collected?”. $P_{x=F150}(x) \approx 15.79\%$.
- “which car model appears with greatest frequency?”. $P_{x=Corolla}(x) \approx 26.32\%$.

This set is composed of *strings*, but the model still returns a *quantitative* result.

email: abran012@uottawa.ca

ML Modelling

Nature of data: qualitative and quantitative

X = {F150, Silverado, ..., ..., Corolla, Corolla, ..., ..., ,F150,F150, Silverado, ..., ...}

Finally, if you ask

- "what is the selling performance of the F150?"
  - and define selling performance as
    - "good" for P(x)>20%,
    - "medium" for 10% <= P(x) <= 20%
    - "poor" for P(x) < 10%

The answer is: "medium performance", a *qualitative* answer (because it refers to a range within possible outcomes).

This shows that a quantitative or qualitative result is not bound by the nature of the input data (number vs string).

email: abran012@uottawa.ca

ML, AI and Big Data

ML is part of AI processing flow (modelling algorithms).

Big data is part of its infrastructure (how/where data is stored, using what resources to collect, sort, join, find, delete, insert, copy, etc.).

AI is a generic term for the ensemble that:

- Recognizes a query (interfaces with sensor or humans);
- Uses ML to predict or classify related data;
- Uses ML outcomes to feed logical decision methods;
- Feeds back the decision outcome to actuators, human interfaces or log files.

Example: data contains outside air humidity levels. Smart house uses ML to classify that data. If result is class={“is raining”} then smart house activates {“close window”} actuators; & pushes notification to owner via WiFi.

email: abran012@uottawa.ca

Bibliography

- “Machine Learning For Dummies” Paperback – May 2016, by John Paul Mueller and Luca Massaron.
- “Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners” - May 2017, by N.D. Lewis.
- “Machine Learning with R - Second Edition: Expert techniques for predictive modeling to solve all your data analysis problems” - Jul 2015, by Brett Lantz.
- “Machine Learning with Spark” - Apr 2017, by Rajdeep Dua, Manpreet Singh Ghotra and Nick Pentreath.

email: abran012@uottawa.ca

Thank You
Merci
Obrigado
متشكرم
谢谢