Modelling and Evaluation

CS5056 Data Analytics

Francisco J. Cantú, Héctor Ceballos

Tecnológico de Monterrey

March 3, 2021

Februrary-June, 2021

Look for the Model that Fits Data the Best

CS5056 Data Analytics
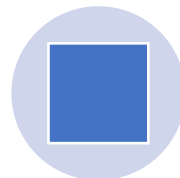## March 3, 2021
Class 4/16

Agenda

MODELLING:

FITTING AN EQUATION
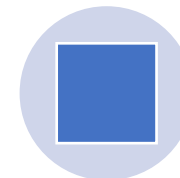
TO DATA

REGRESSION ANALYSIS

FEATURE SELECTION

PANEL DATA

SUPPORT VECTOR MACHINES

OVERFITTING & UNDERFITTING

CS5056 Data Analytics
March 10, 2021
Class 5/16
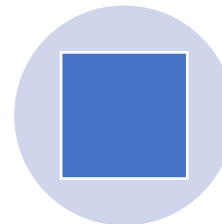
Tecnológico de Monterrey

Agenda

MODELLING:

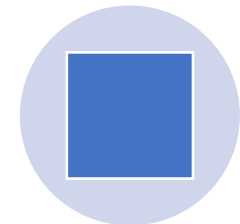**FITTING AN STRUCTURE TO DATA**

DECISION TREES, RANDOM FOREST, BOOSTING
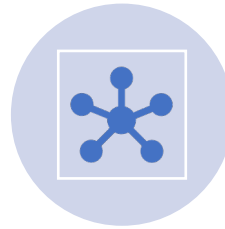
FEATURE SELECTION

LOGISTIC REGRESSION

OVERFITTING & UNDERFITTING

CS5056 Data Analytics
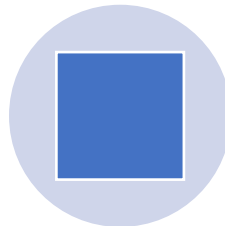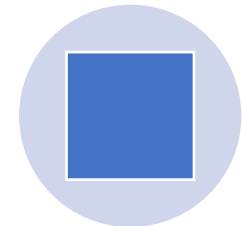## March 17, 24 2021
Class 6,7/16

Agenda

PROBABILISTIC MODELLING

BAYES CLASSIFIER

BAYESIAN NETWORKS
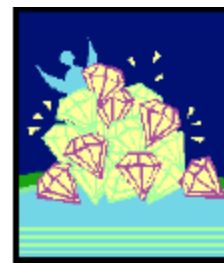
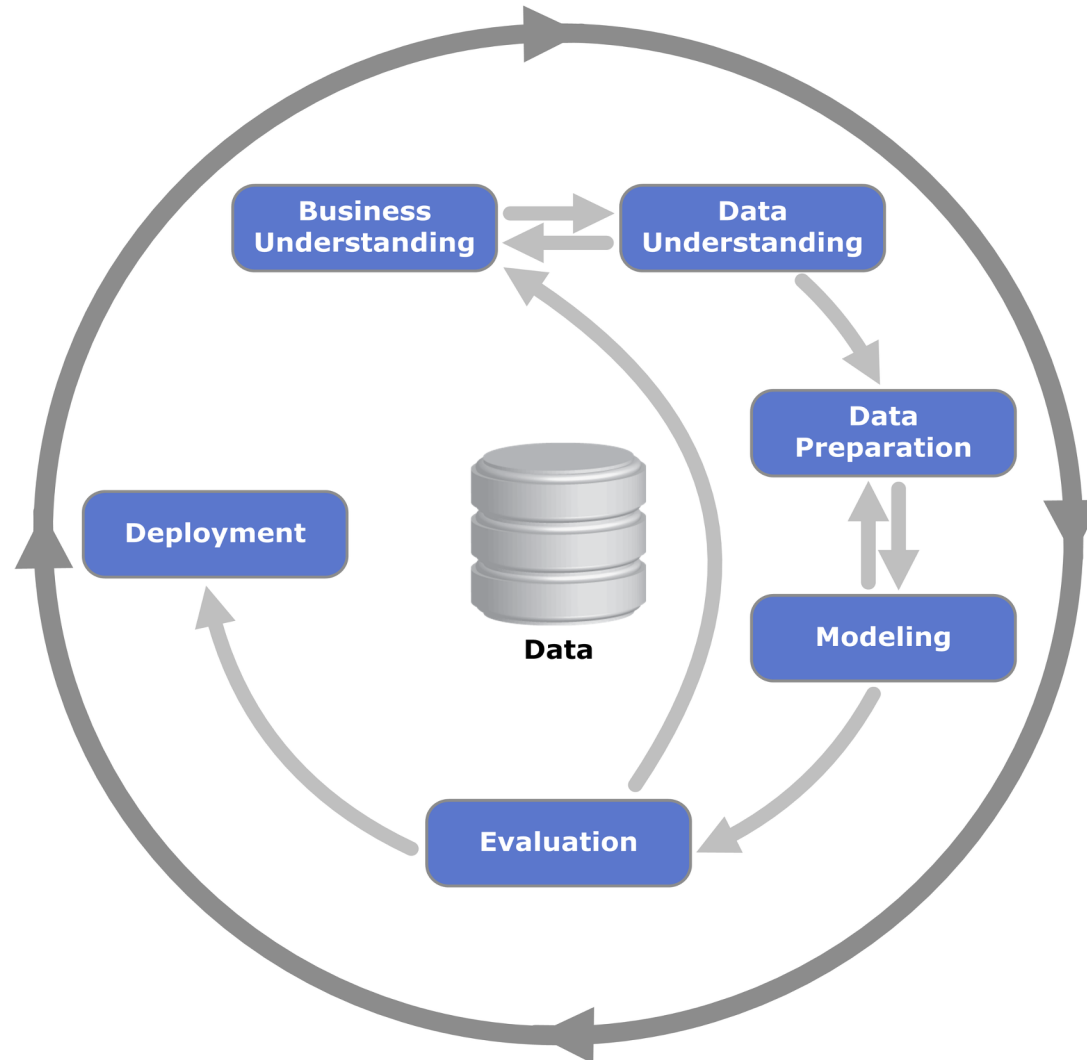FEATURE & MODEL SELECTION

OVERFITTING & UNDERFITTING

Data Mining Cycle: CRISP-DM

Modelling and Evaluation

- Exploratory
- Predictive
- Classification
- Monitoring
- Diagnosis
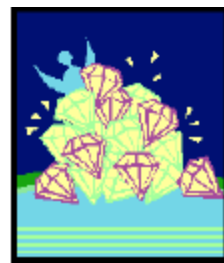- Control

Your Dataset

- How well do you know your data?

- Does it contain text, images, speech, combinations?

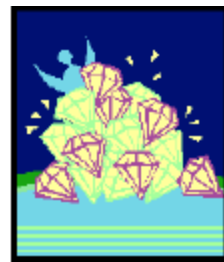- What type of attributes is in data?

Your Dataset

- It is a supervised or non-supervised data science problem?

- If it is a supervised problem, have you identified a target attribute?

- What type of attribute the target and how does it relate with the rest of the variables?

Fitting a Model to Data

- Fitting a mathematical equation
  - Select the type of equation
  - Select the features
  - Learn the parameters of the equation
  - Assess the model learned

- Fitting a mathematical structure
  - Select the type of structure
  - Select the features
  - Learn the parameters of the structure
 - Assess the model learned
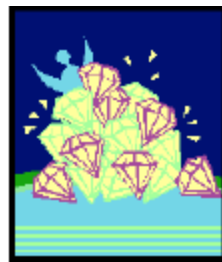
Modelling

- It consists in building a model that best fits the data using Statistical or Machine Learning techniques

- How do we select the best model and how do we know it is the best one?

- What are the best attributes to use and how do we select them?
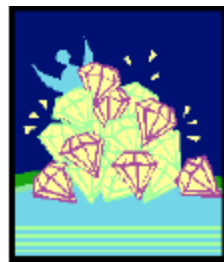
Modelling

- What Statistical or Machine Learning techniques produce the best model?

- What are the optimal parameters of the chosen model?

- How do we assess the chosen model and determine how good it is?
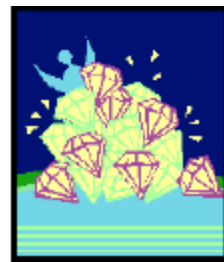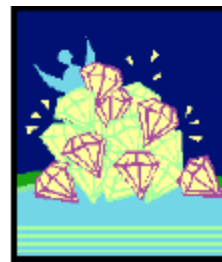
Supervised Learning

- Numerical Target Variable
  - Regression Analysis
  - Support Vector Machine
  - Panel Data
  - Decision Tree
  - Random Forest
  - Boosting
  - Probabilistic modelling

- Categorical Variable
  - Logistic Regression
  - Class-probability Estimation

Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence
Tecnológico de Monterrey
Enago-Academy Advisor for Strategic Alliances

E-mail: fcantu@tec.mx, fjcantor@gmail.com

Cel: +52 81 1050 8294, SNI-2 CVU: 9804

Personal Page: http://semtech.mty.itesm.mx/fcantu/

Facebook: fcantu; Twitter: @fjcantor; Skype: fjcantor

Orcid: 0000-0002-2015-0562

Scopus ID:6701563520

Researcher ID: B-8457-2009

https://www.researchgate.net/profile/Francisco_Cantu-Ortiz

https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ

https://itesm.academia.edu/FranciscoJavierCantuOrtiz

Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México