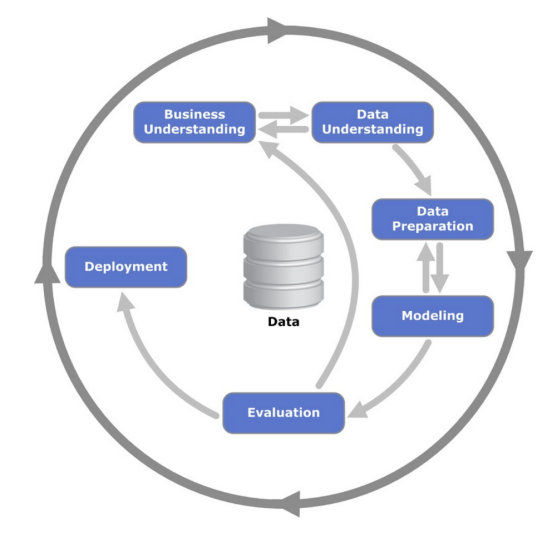
LECTURE 1

CRISP -DM

6 major phases



STEP 1 - BUSINESS UNDERSTANDING

purpose of data mining study

STEP 2 - DATA UNDERSTANDING

identify the relevant data

STEP 3 - DATA PREPARATION

STEP 4 - MODEL BUILDING

STEP 5 - TESTING AND EVALUATION

STEP 6 - DEPLOYMENT

SIMPLE LINEAR REGRESSIONYX

Y» *b*0 + *b*1X

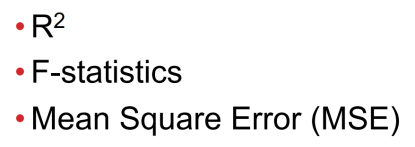
least squares criterion

least squares fit

OLS(Ordinary Least Square)

P value < 0.05

MEASURE MODEL PERFORMANCE



MULTI LINEAR REGRESSION

Difficult to visualize

CODING SCHEME FOR CATEGORICAL VARIABLES

QUALITATIVE PREDICTORS

1 -> female 0 -> male

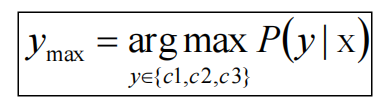
One MUST be ZERO

If k >=3 , create k-1 dummy value

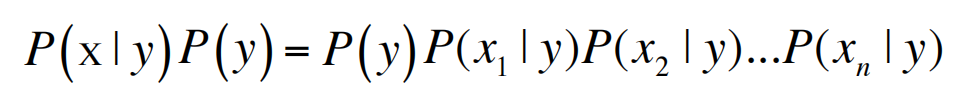
PREDICTIVE ANALYTICS

CLASSIFICATION METHODS

◎BAYESIAN METHODS



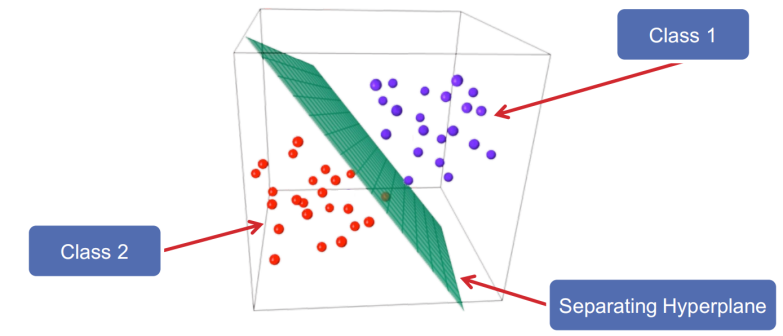
◎NAIVE BAYES CLASSIFIERS



◎KNN

◎Decision Tree

◎Support Vector Machine(SVM)



◎Neural Network

ACCESSING MODEL PERFORMANCE

ACCURACY

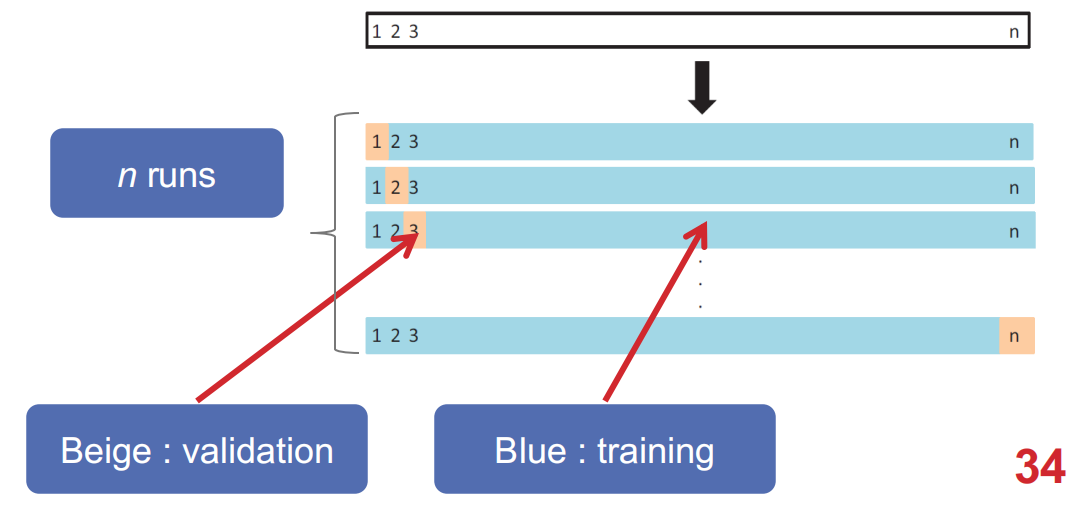
MODEL EVALUATION

Resampling Methods (running the experiment multiple times using the same training set)

Validation Set Approach

randomly split

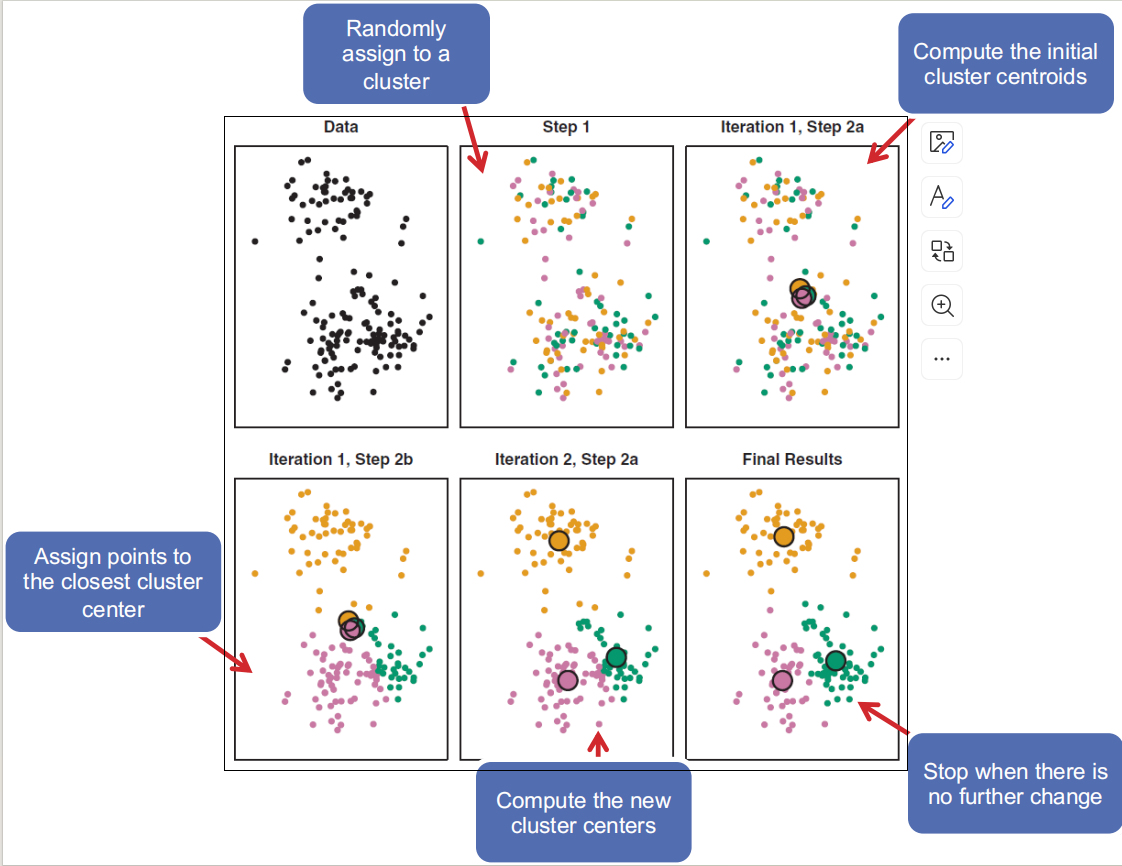
Leave-One-Out Cross Validation(LOOCV)



K-Fold Cross Validation

CLUSTERING

K-MEANS CLUSTERING



MINING WEB CONTENT I

WEB SCRAPER

WEB BASICS

HTML = Hyper Text Markup Language

TECHNIQUES FOR MINING

WEB SCRAPER

^ indicates the start of the string

( ) defines a new capturing group

. indicates any character

\* indicates 0 or more times

.\* indicates a sequence of any character occurring 0 or more times

$ indicates the end of the string

MINING WEB CONTENT II

FLAW of string-based approach of web scraping : too easily affected

TECHNIQUES:

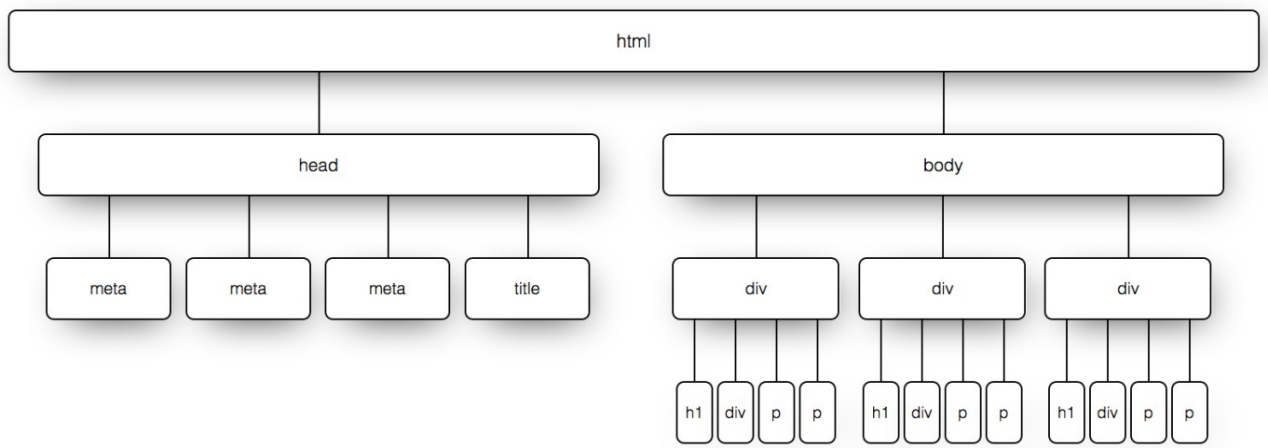
• Extracting content using a HTML parser

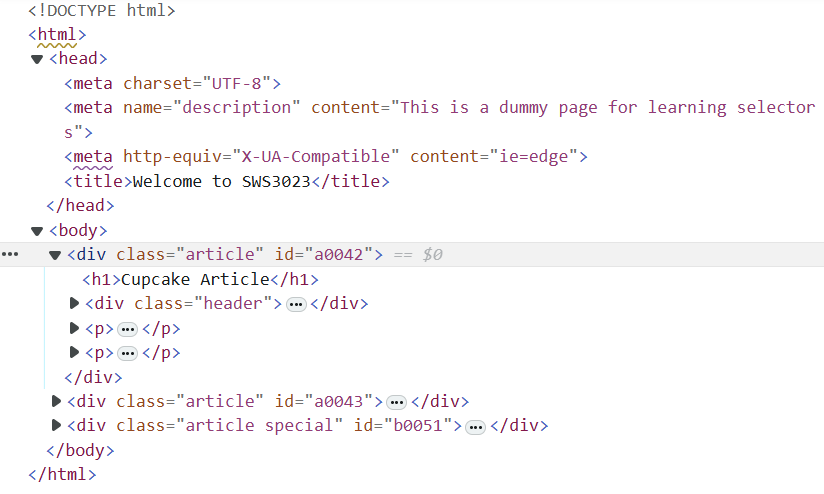
• Web Scraping using API

• Scraping using an actual browser/headless browser

DOM(Document Object Model)

DOM tree

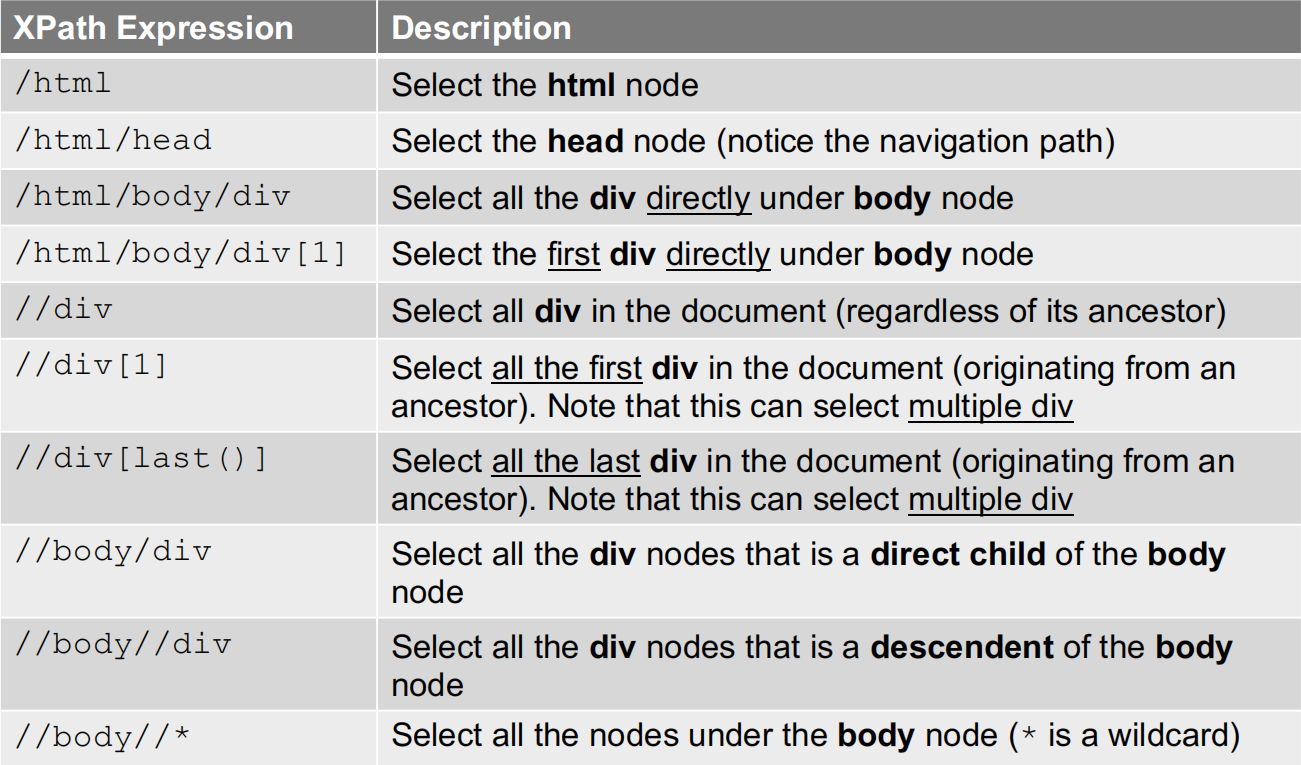


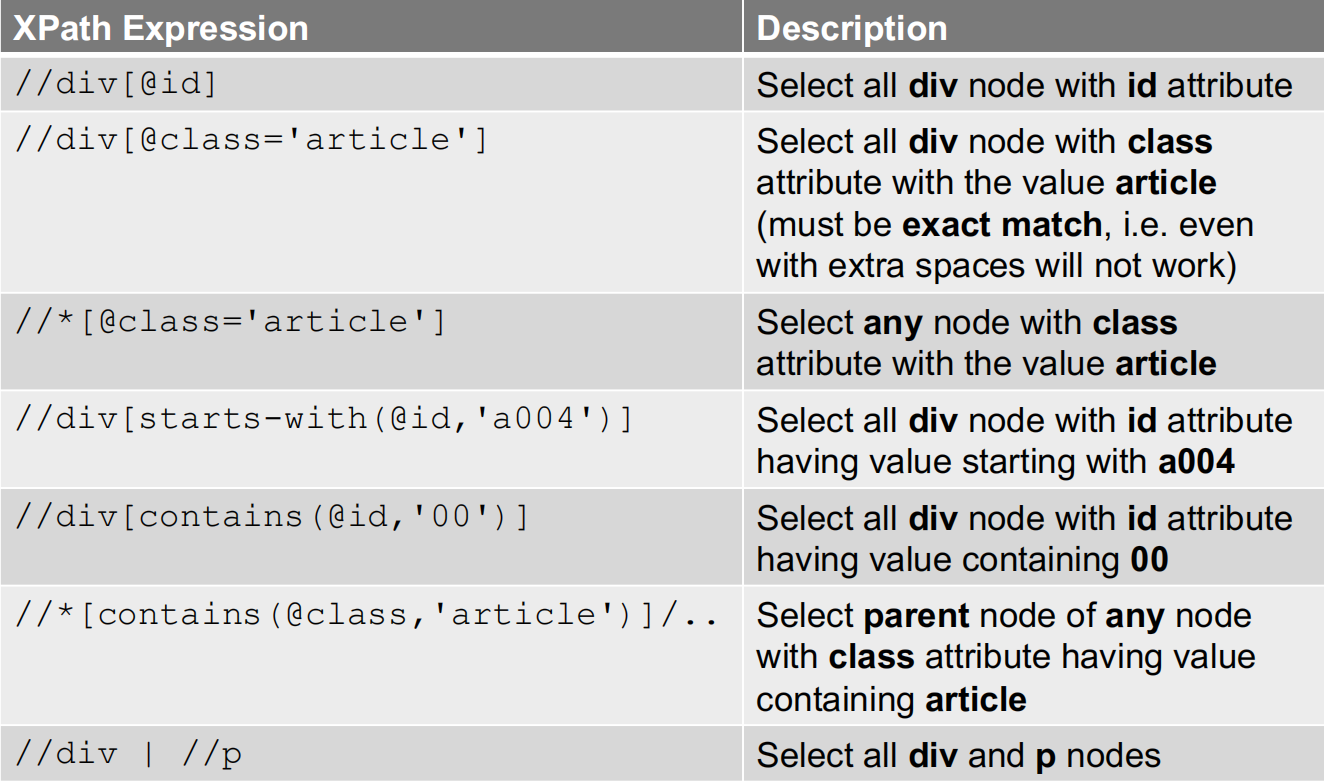


NAVIGATE THE DOM TREE

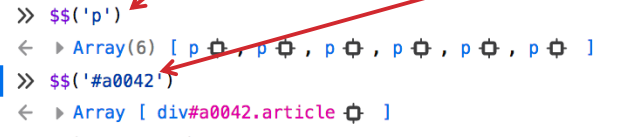
XPATH

Locate in XML(eXtensible Markup Language) document





CSS SELECTORS



MINING WEB CONTENT III

Web Scraping Using API

Web application design

Client : Web Browser , Mobile App

Server : Web Server

Private & public API

RECOMMENDER SYSTEMS

INTRODUCTION

Editorial and hand curated

•Product of the Week

•Staff’s favorites

• etc

Simple Aggregates

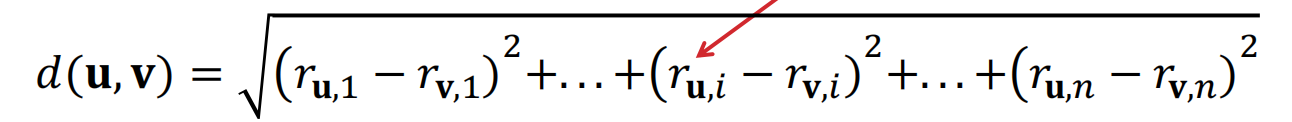
• Most popular, Top rated

Tailored to individual users

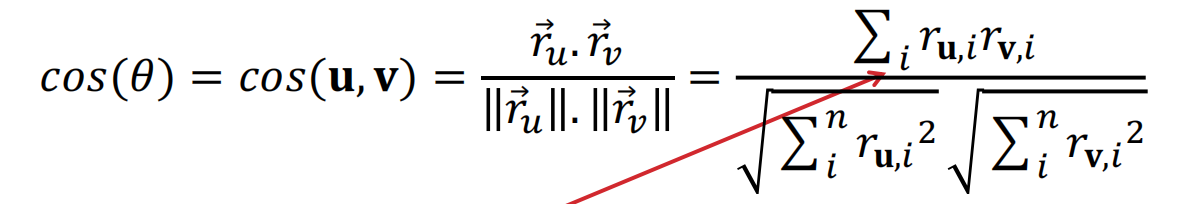
•Personalized recommendations

SIMILARITY MEASURES

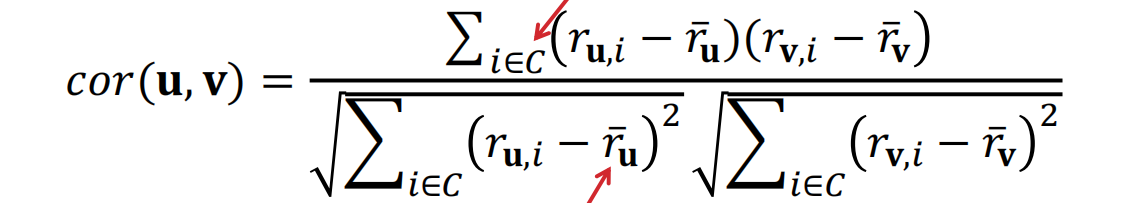
Euclidean Distance



Cosine Similarity



Correlation



Jaccard Similarity

