Data Innovations

Mert Nuhoglu

August 23, 2015

Trailer

Introduction

- New and shiny
- Widely varying applications
- Different problems
- Different domains
- Marketing, healthcare, biotechnology, utilities

Table of Contents

- Classification
- Regression
- Clustering
- Affinity
- Profiling
- Dimension reduction
- Graph mining
- Text mining

Customer Churn Prediction

- Customer churn prediction
  - Which customers will churn?

customer	will churn
John	?
Lisa	?

-> classification ->

customer	will churn
John	Yes
Lisa	No

Classification: Customer Churn

- Existing customer database

name	city	age	sex	profession	edu
Adams John	NY	30	M	programmer	undergrad
Lisa Meyer	LA	40	F	pianist	high school
Bruce Elm	SF	20	M	teacher	undergrad

- Use this as input to classification model

How do classification algorithms know in advance?

- Needs historical data
- Historical data is already labeled

Historical data:

name	city	age	sex	profession	churned
Adams John	NY	30	M	programmer	No
Lisa Meyer	LA	40	F	pianist	Yes
Bruce Elm	SF	20	M	teacher	No

- “churned” is the label/class

Terminology of learning

- Learning from labeled historical data
  - Existing data: Training data and test data
  - Output: Model for classification
  - Train the model
- Use the model to predict the class of new data

How does classification work in whole?

- Learn from historical data (rules/model)
- Apply those rules to new data

What is a model?

- Model $=$ set of rules
- Rules?
  - If the customer is female and younger than 30, she will churn.

Why is it called classification?

customer	will churn
John	?
Lisa	?

Other Churn Problems

- Which customers will cancel their subscription?
- Which gamers won't buy the game?
- Which web visitors will end session?
- MegaTelco: telecom company
  - 20% of customers leave when contracts expire

Which gamers won't buy the game?

- Gaming company
- Uses paid marketing campaigns in several channels - Wants to improve efficiency in real time

Which web visitors will end session?

- News web site
- Wants to keep visitors on site
- Show interesting stuff to visitors that will end session

Real time vs. batch classification

- Real time classification
    - Classify entities at the moment
- Batch classification
    - Classify entities at every night

MegaTelco: telecom company

- 20% of customers leave when contracts expire
- Attractive offers to customers who will churn

Classification uses in marketing campaigns

- Which customers will respond to an offer
  - Direct marketing campaigns
  - Select people who are likely to respond

Classification uses in anomaly detection

- Detecting diseases
- Detecting frauds
    - Credit card
    - Intrusions to computer networks
    - Spam emails
- Detecting life style change
    - Expecting a baby?

Detecting or preventing diseases

- Quanttus: Preventing heart attacks
- Growsafe: Detecting sick cattle

Detecting fraud

- Credit card fraud
- Fraud in public social help

Detecting fraud in computer networks

- Network intrusion

Detecting life style change

- Target stores: Predicting pregnant customers

Risk classification in insurance

- Dynamic risk management

Risk classification in insurance

- Probability of a claim

Risk classification in consumer credits

- Signet Bank 1990s
- The risk level of a consumer credit to default
- Customize the credit conditions by risk level

Risk classification in buildings

- NYC Fire Department: risk score of buildings

Risk classification in healthcare

- Efficacy of treatments

Risk classification in higher education

- University admissions
- Will the admitted student accept the offer or not?

Risk classification in product manufacturing

- Manufacturing companies
- Will the next product lead to warranty claim?

Predicting demand level

- What will be the demand for our clothes next season?
- What will be the demand for our cars next season?
- Classification: qualitative variable
- Regression: quantitative variable

Predicting production level

- Potato yield prediction
  - The crop is underground
- Groundcover

Predicting customer's purchase level

- How much calls will a telecom customer make?
- How much payment will a consumer make with his credit card?
- How much virtual products will a gamer buy?

Customer Segmentation Problems

name	spending
john	100
lisa	200
eva	180
...	...

-> clustering ->

cluster	range
high spenders	> 500
middle spenders	$100 < x < 500$
low spenders	< 100

Call Usage Patterns

- Different groups of customers by
  - Calls
  - Sms messages
  - Data utilization

Common patterns among patients

- Root causes of diseases
- Is the disease related to some location?
- Is the disease related to some specific range of values in different variables?