Evaluation of a Neighborhood Weighted Graph between Products

CONFIDENTIAL

Nicolas Drizard

Advisors: David Bessis, Artem Kozhevnikov, Victor Mazzeo
Supervisor: Michalis Vazirgiannis
July, 9^{th} 2015

Outline

Table of Contents

Objective

Workflow of the targeting Application:

1. Choosing the campaign content
2. Choosing the campaign volume
3. Learning the model
4. Scoring the user base
5. Obtaining the target

Data Tables

- **user**: socio demographic characteristics and contact data of each user
- **product**
- **purchase**
- **page-view**: information about the navigation of the user on the website, i.e. how he arrived, on which content he clicked...
- **email**: data about the marketing email sent, in particular if the user opened it or clicked inside

Example

domain	gmail.com
zipcode	93400
user_id	420050933
contactable	True
firstname	Roger
title	MISTER
Lastname	Dupond
yob	1978
first_purchase_date	2001-02-16
dob	1978-12-05
gender	M
country	France
login	bernarddupond

Figure: User Table

Example

product_id	188752258
date	2015-06-30
user_id	420050933
basket_id	391290000
price	10.4

Figure: Purchase Table

product_id	156820100
genre	Livre
categories_1	Litterature
categories_2	Littérature française
name	un bel morir
price	5.00

Figure: Product Table

Method

supervised.png

Figure: Supervised Learning Workflow

Source: http://www.astroml.org/sklearn_tutorial

Method

- **Dimension Reduction**: embeds each category of data in a low dimensional vector space
- **Regularized Logistic Regression**

$$\min_{w} \frac{1}{2n_{sample}} ||Xw - y||^2 + \alpha ||w||$$

Targeted Products

target.png

Performance

gain_curve_ex.png

Extract Info

sex_distribution.png

age.png

Extract Info

geographic.png

domain.png

Extracts Comparison

sex_comp.png

ex_overlap.png

Filter

Issue

Should we apply a filter before the learning or after directly to the scored table?

product	in idf	not in idf
theatre	141	71
orsay	241	227
veles	517	411

Figure: Original Data

filter	product	robustness	lift 10%
pre	orsay	0.996	5.708
post	orsay	+6%	+14
pre	theatre	0.947	4.412
post	theatre	-7%	0
pre	veles	0.888	5.174
post	veles	-3%	+6

Figure: Comparison Results

Minimum of Positive Events

Issue

How could we target the products with not enough purchase?

Neighborhood Weighted Graph or Similarity Mapping

Neighborhood Weighted Graph is a mapping where each structure is mapped to a ranked list of similar products, called buddies.

Target Extension

Extension of the targeted products list with the most similar products to increase the number of positive events considered for the learning until a given threshold.

Table of Contents

Global Model Performance

- AUC
- Robustness
- Lift

gain_curve.png

Figure: Gain Curve

Graph Homogeneity & Stability

Homogeneity

Evaluates the homogeneity of the buddies for any targeted product

$$overlap_{homogeneity} = Extract_{even}^{5\%} \cap Extract_{odd}^{5\%}$$

Stability

Evaluates the stability of the construc- tion of the graph

$$overlap_{stability} = Extract_{train}^{5\%} \cap Extract_{test}^{5\%}$$

Remark

References needed: $overlap_{extratag} < ... < overlap_{intratag}$

Buddies Specificity

Buddies Specificity Overlap

Overlap among the buddies lists truncated at a given number (here 50) for different entries weighted by contribution.

Overlap_Buddies.png

Figure: Gain Curve

Aggregated Report

auc	robustness	$overlap_{intratag}$	$overlap_{extratag}$	$overlap_{homogeneity}$	lift_5	lift_10
0.69	0.91	0.93	0.77	0.93	8.90	5.99

Figure: Model Performance Metrics

positive_events	length	entropy	std	max	argmax	mean	argmean	argmedian
182	150	2.99	4.61	35.4	15	1.10	60	50

Figure: Detailed Report

curves.png

Figure: Gain Curves

pressure.png

Figure: Pressure Distribution

campaign_overlap.png

Figure: Campaing Overlap

Applications

Possible applications of the reporting mission:

- Features evaluation
- Benchmark of the positive events minimum required
- Setting dynamically targeting parameters for each client
- Set a trust threshold

Table of Contents

Let M_{id} be the matrix of purchase occurrence over the user_id

$$M_{id} = \begin{array}{c c} & \begin{array}{ccc} user_1 & \dots & user_n \end{array} \\ \begin{array}{c} product_1 \\ \vdots \\ product_n \end{array} & \begin{pmatrix} 1 & \dots & 0 \\ \dots & \dots & \dots \\ 1 & \dots & 1 \end{pmatrix} \end{array}$$

Data

Let M_{sd} be the matrix of purchase occurrence over the socio demo characteristics: (**year of birth** (**yob**), **sex**, **firstname**)

$$M_{sd} =$$

$$\begin{array}{c c}
 & \begin{array}{cccccccc} yob_1 & \dots & yob_n & firstname_1 & \dots & firstname_n & M & F \end{array} \\
\begin{array}{c} product_1 \\ \vdots \\ product_n \end{array} & \begin{pmatrix} 1 & \dots & 0 & 1 & \dots & 0 & 1 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \dots & 1 & 1 & \dots & 1 & 0 & 1 \end{pmatrix}
\end{array}$$

Methods

- Jaccard Index
- Singular Value Decomposition (SVD)
- Non Negative Matrix Factorization (NMF)

Top Buddies Intersection

Remark

Evaluation of the intersection between the buddies lists from two graphs:
x axis : number of products
y axis : number of common buddies

jaccard.png

Figure: Jaccard Index

Top Buddies Intersection

svd.png

Figure: SVD

nmf.png

Figure: NMF

Buddies Specificity Overlap: Jaccard

jac_id.png

Figure: User id

jac_sd.png

Figure: Socio Demo

Buddies Specificity Overlap: SVD

svd_id.png

Figure: User id

svd_sd.png

Figure: Socio Demo

Buddies Specificity Overlap: NMF

nmf_id.png

Figure: User id

nmf_sd.png

Figure: Socio Demo

Other Approaches

Possible applications of the reporting mission:

- Distance-based metrics between the distribution of the socio-demo characteristics in the occurrence profile
- Combining Different Methods