$10,000 • 122 teams

ICDM 2015: Drawbridge Cross-Device Connections

Merger and 1st Submission Deadlir

Mon 1 Jun 2015 Mon 24 Aug 2015 (56 days to go)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline

Forum

Scripts

Leaderboard

Leaderboard

1. Dave Mullen
2. gk
3. YcdoiT
4. idle_speculation
5. agdavis
6. Bayesian Boat
7. Joshua Havelka
8. RingRing
9. Bryan Johnson
10. Milton Bose

46 Scripts

Sample Rows From Each SQLite Table
2 Votes / 5 days ago / SQLite

Exploring the Drawbridge Data
12 Votes / 13 days ago / RMarkdown

t-SNE Visualization of Devices
7 Votes / 15 days ago / R

quasirandom sequence
0 Votes / 9 days ago / Python

Reading Bad CSV Files
10 Votes / 16 days ago / RMarkdown

Fixing Bad CSV Files (With Download)
5 Votes / 13 days ago / Python

Forum (38 topics)

understanding problem
15 hours ago

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
dev_train_basic.csv	.zip (2.14 mb)
cookie_all_basic.csv	.zip (34.08 mb)
ipagg_all.csv	.zip (112.23 mb)
dev_test_basic.csv	.zip (713.96 kb)
property_category.csv	.zip (2.97 mb)
sampleSubmission.csv	.zip (126.38 kb)
id_all_ip.csv	.zip (225.42 mb)
id_all_property.csv	.zip (356.89 mb)
database.sqlite	.zip (3.35 gb)

See this script for a quick exploration of the data

This competition asks you to determine which cookies belong to an individual using a device. You are provided with relational information about users (represented by the id column drawbridge_handle), devices (device_id), cookies (cookie_id), as well as other information on IP addresses and behavior. For each device in the test set (dev_test_basic.csv), you must provide a list of cookie ids (from cookie_all_basic.csv) which you believe belong to the person using the given device_id. As you will see, the drawbridge_handle column is missing, denoted by the value -1, in the test set.

Training set (semi-)supervised learning methods

If you want to construct the training set and apply supervised learning, you can take the training data (dev_train_basic.csv), and find those cookies in the file cookie_all_basic.csv with the same drawbridge_handle. You could use device and cookie pairs with different drawbridge_handles as negative training data. Please note that some of the cookies have drawbridge_handle = -1, which means the drawbridge_handle for that cookie is unknown.

The same set of cookies that will be used for both training and testing purposes.

Data types

There are four different types of data attributes: *Index, Categorical, Boolean and Int*. Index and Categorical are both enumerated type. Index has bigger set of elements (e.g. device_id or cookie_id), while Categorical has smaller set of elements (e.g. all the device types, or all the desktop browser version). Boolean applies to those attributes with only 2 possible values, and Int describe the count of the attribute in a continuous way.

looking for partner
19 hours ago

relation of a device_ID with drawbridge handle?
2 days ago

More trees in random forest lead to overfitting?
3 days ago

SQLite Database
3 days ago

Reading Bad CSV Files
4 days ago

teams

players

entries

Meaning of the attributes

For some attributes, the meanings are publicly available, as specified in the table schema. For some other attributes, the meanings are anonymous.

Data table schema and meaning

1) Device basic information table (dev_train_basic.csv and dev_test_basic.csv) and cookie basic information table (cookie_all_basic.csv). Basic information tables provide high-level summary information regarding the device and cookie. For devices, the data is split into train and test parts. For cookies, there is one table that has the basic information for all the cookies. The schema of basic table is as below:

Device basic Info tables (device_train_basic.csv and device_test_basic.csv)

	Feature Name	Type	Meaning
1	Drawbridge Handle	Index	Drawbridge identifier, uniquely identify a person behind device and cookie. Device and cookie belong to the same person will have the same handle. Different handles represent different users.
2	Device ID	Index	Index of each device. Uniquely identify each device
3	Device type	Categorical	Device type, iphone, android phone, ipad, android pad, etc.
4	Device OS version	Categorical	Device OS version. e.g. ios 8.0,
5	Device Country Info	Categorical	Which country this device belongs to
6	Anonymous_c0	Boolean	Drawbridge anonymous feature to describe device. If the value is unknown, it will be -1
7	Anonymous_c1	Categorical	Drawbridge anonymous feature to describe device. Categorical value. Will be -1 for unknown value.
8	Anonymous_c2	Categorical	Drawbridge anonymous feature to describe device. Categorical value. Will be -1 for unknown value.
9	Anonymous_5	Int	Drawbridge anonymous feature to describe device. Integer value.
10	Anonymous_6	Int	Drawbridge anonymous feature to describe device. Integer value.
11	Anonymous_7	Int	Drawbridge anonymous feature to describe device. Integer value.

cookie basic info table (cookie_basic.csv)

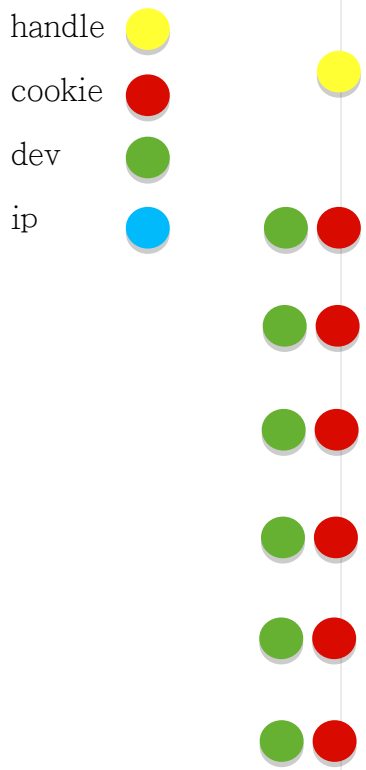
	Feature Name	Type	Meaning
1	Drawbridge handle	Index	Drawbridge identifier, uniquely identify a person behind device and cookie. Device and cookie belong to the same person will have the same handle. Different handles represent different users.
2	cookie ID	Index	Index of each cookie. Uniquely identify each cookie
3	computer OS type	Categorical	cookie computer operation system type(e.g. window xp)
4	Browser version	Categorical	cookie browser version (e.g. Safari-6.0)
5	cookie country info	Categorical	Which country this cookie belongs to

6	Anonymous_c0	Boolean	Drawbridge anonymous feature to describe device. Same meaning as Anonymous_c0 feature in device_basic table.
7	Anonymous_c1	Categorical	Drawbridge anonymous feature to describe device. Categorical value. Will be -1 for unknown value. Same meaning as Anonymous_c1 feature in device_basic table.
8	Anonymous_c2	Categorical	Drawbridge anonymous feature to describe device. Categorical value. Will be -1 for unknown value. Same meaning as Anonymous_c2 feature in device_basic table.
9	Anonymous_5	Int	Drawbridge anonymous feature to describe device. Integer value. Same meaning as Anonymous_5 feature in device_basic table
10	Anonymous_6	Int	Drawbridge anonymous feature to describe device. Integer value. Same meaning as Anonymous_6 feature in device_basic table
11	Anonymous_7	Int	Drawbridge anonymous feature to describe device. Integer value. Same meaning as Anonymous_7 feature in device_basic table

2) IP table (id_all_ip.csv) describes the joint behavior of device or cookie on IP address. All the info of devices and cookies are merged into one single table, and we can use column 2, a boolean type, to differentiate if it's a device or cookie. One device or cookie may appear on multiple IPs, and we put all the IPs where we have seen a device/cookie into a bag. The data from column 3 to column 9 in the table makes up a tuple inside the bag, and that tuple describes the behavior of device/cookie on that particular IP. The table schema is below:

device/cookie ip table (id_all_ip.csv)

	Feature name	Type	Meaning
1	Device/cookie ID	Index	ID of the device or cookie
2	Device or Cookie	Boolean	specify if it's a device or cookie. 0 for device and 1 for cookie. Only has two possible values
3	IP	Index	IP address (First field in the tuple)
4	Freq count	Int	How many times have we seen dev or cookie in column 1 appear on the IP in column 3
5	Anonymous Count 1	Int	Anonymous number that describes the behavior of the specified device or cookie on the IP
6	Anonymous Count 2	Int	Anonymous number that describes the behavior of the specified device or cookie on the IP
7	anonymous Count 3	Int	Anonymous number that describes the behavior of the specified device or cookie on the IP
8	Anonymous Count 4	Int	Anonymous number that describes the behavior of the specified device or cookie on the IP
9	Anonymous Count 5	Int	Anonymous number that describes the behavior of the specified device



			or cookie on the IP (Last field in the tuple)

3) IP aggregation table (ipagg_all.csv). In general, we could see many different devices or cookies from a single IP. While the device/cookie IP table provides the information regarding the individual behavior of one device or one cookie on one IP, the IP aggregation table provides summary information that describe each IP across all the devices or cookies seen on that IP.

IP aggregation table (provide aggregated behavior of each IP) (ipagg_all.csv)

	Feature name	Type	Meaning
1	IP Address	Index	ip address
2	Is cell IP	Boolean	If IP is cellular IP or not. 1 for cellular and 0 for non cellular.
3	Total Freq	Int	Total number of observations seen on this IP (This number is the aggregated observation count on all the devices and cookies seen from this IP)
4	Anonymous count c0	Int	Anonymous count that describes the behavior of the IP
5	Anonymous count c1	Int	Anonymous count that describes the behavior of the IP
6	Anonymous count c2	Int	Anonymous count that describes the behavior of the IP

4) Property observation and property category tables provide the information regarding website (for cookie) and mobile app (for device) that user has visited before. "id_all_property.csv" table lists the specific name of the website or mobile app, and property_category.csv table lists the categorical information of the website/mobile app. They schemas are listed below:

Property observation table (id_all_property.csv)

	Feature name	type	meaning
1	Device/cookie ID	Index	ID of device or cookie
2	device or cookie indicator	Boolean	specify if it's a device or cookie. 0 for device and 1 for cookie. Only has two possible values
3	Property ID	Index	Website name for cookie, and mobile app name for the device
4	Property unique count	Int	How many times have we seen device or cookie on this property

Property category table (property_category.csv)

	Feature name	type	meaning
1	Property ID	Index	Website name for cookie, and mobile app name for the device
2	Property category	Categorical	Category of the website or the mobile app

How tojoin all the above tables?

Device or cookie ID can be used as persistent keys to join tables. For IP observation and aggregation tables, we can also use IP to join the tables. For property information, we can use property ID to join the id_all_property.csv and property_category.csv tables.