**DISCUSSION 14/01/2021**

|  |  |  |
| --- | --- | --- |
| DEVICE ID | CATEGORY 1 | CATEGORY 2 |
| A | [A\_X1, A\_Y1, A\_ Z1, A\_W1] | [A\_X2, A\_Y2, A\_ Z2, A\_W2] |
| B | ... | ... |
| C | ... | ... |
| .... | ... | ... |

Eg: If A visits POI\_A (which is in Category1) 10 times in 1st day and spends 1hr on each visit and 20 times in 2nd day and spends 20min in each visit, and POI\_B (which is in Category1) 10 times in 1st day and spends 30min in each visit. Assume distance between POI\_A and home is 10km and POI\_B and home is 20km

V – Average visits to a category

(Total no of visits to all POIs in a category/No of days)

V = (10 + 20 + 10)/2 = 20

**Interpretation:** On an average, a person visits this category for 20 times a day

**Doubt:** There can be cases that on one day the person may visit 100 times and never visit the place in the month. In this example, V = 100/30 -> 3 which we will interpret as that person visiting 3 times/day to that category. So, we might assume a strong affinity which is not true.

So, can we exclude V?

X - Frequency to a category

(No of days a person visited all POIs in the category/No of days)

X = (1 + 1)/2 = 1

**Interpretation:** X ranges from [0,1] with 1 being a regular visitor and 0 being never visited

Though A visited 2 POIs belonging to same category on same day, it’s counted as 1 thus avoiding double counting

Y – Avg Distance travelled from home to a category

(Total distance travelled from home to all POIs in this category/No of days)

Eg: (10\*10km (POI\_A) + 20\*10km (POI\_A) + 10\*20km (POI\_B))/2 = 250km

**Interpretation:** On an average, this user travels 250km/day to visit this category

**Doubt:** It’s said that more the distance travelled by the user, more will be affinity towards that category. But the less the distance travelled by the user also can have strong affinity towards that category (assume a near POI).

Z – Avg time spent in a category

(Total time spent on all POIs in all category/No of days)

Eg: (10\*1hr (POI\_A) + 20\*20min (POI\_A) + 10\*30min (POI\_B))/2 ~ 10hr/day

**Interpretation**: On an average, this user spends 10hrs/day in this category

**Doubt**:

1. How can we get this parameter? (Sai, Manideep)

1. Obviously, the more Z, the better. But again we can see examples where a user visits a place for one day, stays for entire 24hrs and didn’t visit in the whole month. So, there Z = 24hrs/30days ~ 0.8hrs/day which implies a strong affinity towards that category.

W – Time gap between consecutive visits to same category

**Doubt:** Is X carrying similar info as W?

**Algorithm**

1. Assume we computed the features and have a table as above. How can we assign the top category to each user?

Because assigning the top category to user implies classification, but we don’t have labelled data to update weights/do gradients.

So, is trees, forests, xgboost etc; at all useful for this problem? Because we can’t simply classify without labelled data

1. If not, can we do some ranking based approach using these features and say for user A, the affinity of categories are

CATEGORY3, CATEGORY4, CATEGORY2 etc; with a probability value or something like that (see below table). And take the top 3 categories to which the user is affined

**Refer:** <https://towardsdatascience.com/ranking-algorithms-know-your-multi-criteria-decision-solving-techniques-20949198f23e>

**Input:** Similar to above table

|  |  |  |  |
| --- | --- | --- | --- |
| USER1 | X | Y | Z |
| CATEGORY1 | 0.1 | 20km | 20min |
| CATEGORY2 | 0.5 | 80km | 2hr |
| CATEGORY3 | 0.9 | 2km | 8hr |

(\*imagine CATEGORY3 is office which is very close)

**Output:**

|  |  |
| --- | --- |
| CATEGORY1 | 4 |
| CATEGORY2 | 3 |
| CATEGORY3 | 1 |
| CATEGORY4 | 2 |

Can we use xgboost, decision tree, random forest for ranking?