Determining a location to set up a medical laboratory using geolocation data and SVM models

SERGIO NAVARRO

1.Introduction

- Business problem:
  - Dynacare Medical Laboratories would like to expand in Calgary, but do not know where would be a suitable location to place a laboratory

- Approach to the solution:
  - Analyze geolocation data with a machine learning model to suggest potential locations to set up a laboratory

2. Data – the data sets

Positive-only

- Neighborhoods where Dynacare has already set up laboratories

Positive-unlabelled

- All the neighborhoods of cities with laboratories

To analyze

- All the neighborhoods of Calgary

2. Data – source of the data

1. Obtain postal codes of the neighborhoods of interest using wikipedia or Dynacare's website

2. Obtain latitudes and longitudes using latlong.net

3. Obtain venues of neighborhoods using Foursquare API

2. Data – examples with coordinates

Postal code	Neighborhood	Latitude	Longitude
G1C	example lab 1	46.881771	-71.189369
G1E	example lab 2	46.860130	-71.194054
G1M	example lab 3	46.817230	-71.269836
G6W	example lab 4	46.757560	-71.225570
H1K	example lab 5	45.608180	-73.544520

2.Data - example with venues

Neighborhood name	Bakery	Bank	Bar	Breakfast Spot	Construction & Landscaping
Braeside, Cedarbrae, Woodbine	0.000000	0.000000	0.000000	0.000000	0.5
Brentwood, Collingwood, Nose Hill	0.000000	0.000000	0.000000	0.333333	0.000000
Bridgeland, Greenview, Zoo, YYC	0.000000	0.000000	0.000000	0.000000	0.000000
City Centre, Calgary Tower	0.022222	0.022222	0.066667	0.000000	0.000000

3.Methodology

▶ Models:
  o For positive-only data: one-class suport vector machine
  o For positive-unlabelled data: Elkanoto support vector machine

▶ Visualization:
 □ Labeled map with follium

4. Results – model evaluation

Positive only

Test size	Jaccard score	F1 Score
0.05	0.400000	0.571429
0.10	0.400000	0.571429
0.15	0.533333	0.695652
0.20	0.578947	0.733333
0.25	0.541667	0.702703
0.30	0.517241	0.681818
0.35	0.558824	0.716981
0.40	0.552632	0.711864
0.45	0.465116	0.634921
0.50	0.520833	0.684932

Positive-unlabelled

Test size	Jaccard score	F1 Score
0.05	0.500000	0.666667
0.10	0.733333	0.846154
0.15	0.916667	0.956522
0.20	0.714286	0.833333
0.25	0.965517	0.982456
0.30	0.900000	0.947368
0.35	1.000000	1.000000
0.40	0.973684	0.986667
0.45	0.560976	0.718750
0.50	1.000000	1.000000

4. Results – suggested neighborhoods
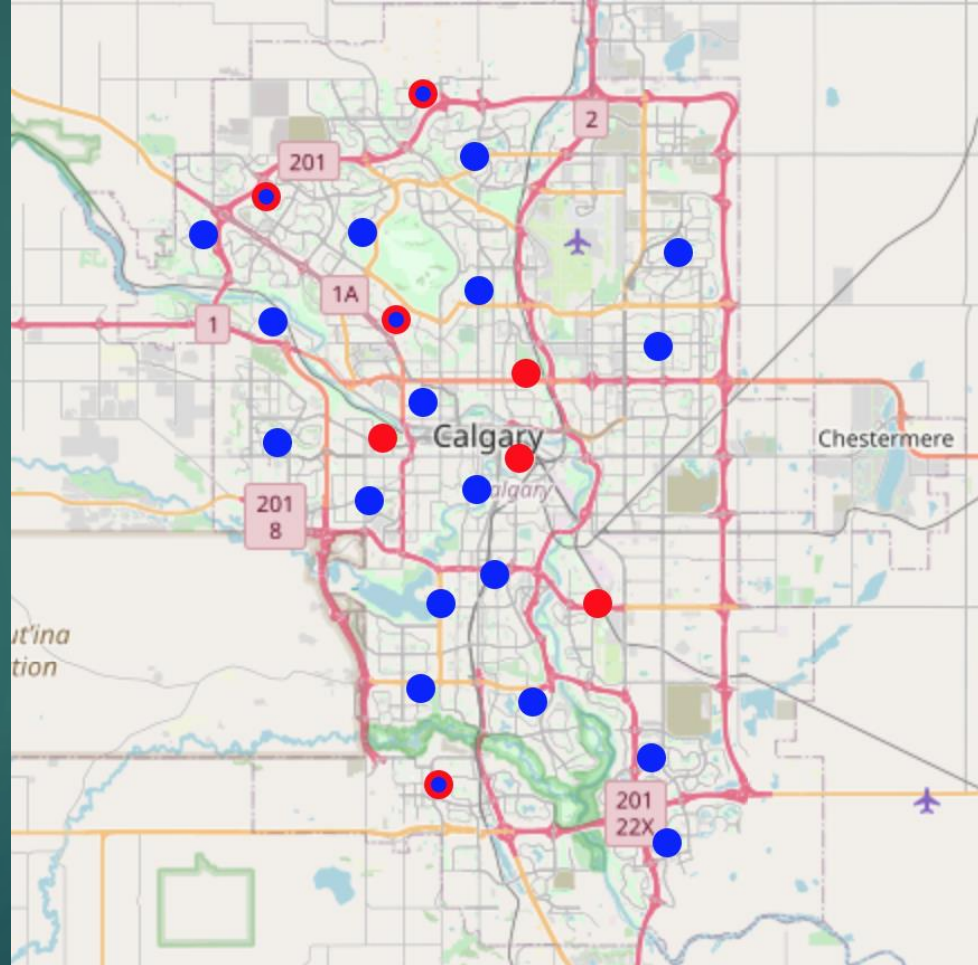
5. Discussion

- Models results
  - PU SVM was more numerous than positive-only
  - Most of the results were in the western side of the city
  - Both models suggest positions in the north and south of the city
- Assumptions
  - Locations with labs represent positive locations
  - Locations without labs represent unlabelled locations
- Limitations and recommendations
  - Try with other classification models
  - Other evaluation methods specific to PU learning

6. Consclusion

- SVM can be used with geolocation data to suggest places to set up new businesses
- PU learning is aplicable in many real-life situations