Where are you going?
An overview on machine learning models for human mobility

Gianni Barlacchi
@giannibarlacchi
About me

• MSc in Computer Engineering

• PhD @University of Trento & TIM
  Machine learning and natural language processing

• Having fun with Databeers Tuscany
Outline

• Introduction to human mobility
• Mobility data
• Human mobility characterization
• Machine learning models
• Conclusions
Goal of the talk

• Introduce human mobility

• Show how to use mobility data within machine learning models

• Show how to set the machine learning problem for next place prediction
Human Mobility

• Understanding the laws of human motion.
• Long-time and long-distance trips are rare.
• Short-time trips mostly consist of intracity.

Human trajectories typically present [4]:

• a high degree of spatial and temporal regularity
• significant probability to return to a few highly frequented locations

[4] Understanding individual human mobility patterns (Gonzalez et al.)
Nearly the entire population can be described with 17 different network motifs [5]

[5] Unravelling daily human mobility motifs (Schneider et al.)
Understanding Human Mobility

Collective human mobility
  • Fluxes of people between places

Individual human mobility
  • Individual movements paths
Why do we want to know how people move?

Statistical mobility properties of group of people can be used for:

**Urban Planning**
- Discover functional regions
  
  [1] *Discovering regions of different functions in a city using human mobility and POIs* (Yuan et al.)

**Traffic Forecasting**
- Improving public transportantion system
  

**Migration analysis**
- Refugee flows in a crisis region
  
Refugee flows in a crisis region

Text, weather and open-data to predict the flow of refugees with machine learning [3].

Refugee flows in a crisis region

Flows of refugees through the balkan route.
Individual Human Mobility

The mobility behaviours of a user can be used for:

### Sharing Economy
Improving bike-sharing system
- [1] Measuring the impact of opening the London shared bicycle scheme to casual users (Lathia et al.)

### Economy & Society
Credit scoring system
- [2] MobiScore: Towards Universal Credit Scoring from Mobile Phone Data (San Pedro et al.)

### Human Behaviour Understanding
Mobility data to assess user physical health conditions
Human mobility to predict flu-like symptoms

Individual mobility behaviours is used to predict the future presence of flu-like symptoms (e.g. cold, fever and cough).
Human mobility to predict flu-like symptoms

Flu-like Symptoms

Mobility Features → Feature Selection → Parameters optimization → Model selection

Classification Model

Flu-like symptom presence?
{ yes - no}
Human Mobility Data

Nowadays, there are almost **6 billions** of mobile phone users worldwide.

- The world coverage has raised from 12% of the world population in 2000 to 96% in 2014

Mobile devices enable massive data collection
Human Mobility Data

Mobility data can be collected through smartphone

- From the device, installing an ad-hoc application
- From the network, cell tower you are connected

**GPS**: best accuracy, high battery drain

**WLAN**: low accuracy, low battery drain

**GSM**: very low accuracy, very low battery drain
Human Mobility Data

Call Detail Records (CDRs)
CDRs contain information about how, when, from where and with whom users communicate [1].

CDR can contain one of the following activities:

- Received SMS
- Sent SMS
- Incoming Call
- Outcoming Call
- Internet

[1] A multi-source Dataset of Urban Life in the City of Milan and the Province of Trentino. (Barlacchi et al.)
Human Mobility Data

Call Detail Records (CDRs)
CDRs contain information about how, when, from where and with whom users communicate [1].

<table>
<thead>
<tr>
<th>Caller ID</th>
<th>Caller Cell ID</th>
<th>Receiver ID</th>
<th>Receiver Cell ID</th>
<th>Datetime</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>111023</td>
<td>123</td>
<td>222012</td>
<td>543</td>
<td>142778572</td>
<td>123</td>
</tr>
<tr>
<td>111023</td>
<td>123</td>
<td>334021</td>
<td>213</td>
<td>142778121</td>
<td>32</td>
</tr>
<tr>
<td>110321</td>
<td>345</td>
<td>320123</td>
<td>321</td>
<td>143213432</td>
<td>20</td>
</tr>
</tbody>
</table>

[1] A multi-source Dataset of Urban Life in the City of Milan and the Province of Trentino. (Barlacchi et al.)
Human Mobility Data

Location history

$L = abce$
Space segmentation

The space has to be divided into regions.
Point of interests and Census Data
Location-based social networks (LBSNs) and government data reveal the function of certain areas in the city.

Land use
Describe the primary use of an area (e.g. commercial)

Foursquare Point of interest
It is a collection of information regarding a venue (e.g. school, restaurant or police department).
Point of Interests
Geographical Data

01 One & Two Family Buildings
02 Multi-Family Walk-Up Buildings
03 Multi-Family Elevator Buildings
04 Mix Residential & Commercial Buildings
05 Commercial & Office Buildings
06 Industrial & Manufacturing Buildings
07 Transportation & Utility
08 Public Facilities & Institutions
09 Open Space & Outdoor Recreation
10 Parking Facilities
11 Vacant Land
Challenges of data collection

• How to engage people to collect data?
• How to protect privacy of people?
• How to deal with missing or fake data?

Reality Mining
95 users, 9 Months
Location, calls, sms, WLAN and Bluetooth connections, application usage

GeoLife
178 users for 4 years.
GPS trajectory of with sequence of time-stamped points (latitude, longitude and altitude).
Deal with mobility dataset

GeoPandas adds a **spatial geometry data** type to Pandas and enables spatial operations on these types, using shapely.

```python
import geopandas as gpd
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from shapely.geometry import Polygon
import math

# Uncomment this part if you run the script from command line
meters = 2000
inputfile = "ny_boros.geojson"
outputfile = "grid.geojson"
```

Code and data are available at https://github.com/gbarlacchi/urbandatascience
Deal with mobility dataset

```python
# Load the area
boros = gpd.GeoDataFrame.from_file(inputfile)

# select the epsg considering your area, http://www.spatialreference.org/ref/epsg/nad83-new-york-long-island-ftus/
bosos.geometry = bosos.geometry.to_crs({'init': 'epsg:2236', 'units': 'm'})
bosos.geometry.plot(figsize=[6,6])
```

![Map of New York City showing different areas.](image)
Deal with mobility dataset

```python
In [16]:
    # Obtain the boundaries of the grid
gps = gpd.GeoSeries(boros['geometry'])

    boundaries = dict({'min_x':gps.total_bounds[0], 'min_y':gps.total_bounds[1], 'max_x':gps.total_bounds[2], 'max_y':gps.total_bounds[3]})

    # Find number of square for each side
    x_squares = int(math.ceil(math.hypot(boundaries['max_x'] - boundaries['min_x'], 0) / meters))
    y_squares = int(math.ceil(math.hypot(0, boundaries['min_y'] - boundaries['max_y']) / meters))

    print "x-axis number of squares: " + str(x_squares)
    print "y-axis number of squares: " + str(y_squares)

    print("Boundaries")
    print(boundaries)
	x-axis number of squares: 23
    y-axis number of squares: 25
    Boundaries
    {'min_x': 771815.88502796332, 'min_y': 1814271.9497595134, 'max_x': 816810.35775163397, 'max_y': 1863152.1449739903}
```
Deal with mobility dataset

```python
In [19]:
polygons = []

for i in range(0, x_squares):
    # increment x
    x1 = boundaries['min_x'] + (meters * i)
    x2 = boundaries['min_x'] + (meters * (i+1))

for j in range(0, y_squares):
    # increment y
    y1 = boundaries['min_y'] + (meters*j)
    y2 = boundaries['min_y'] + (meters*(j+1))
    polygon_desc = {}

    # Create shape (polygon)
    p = Polygon([(x1,y1),(x2,y1),(x2,y2),(x1,y2)])

    # Compute centroid coordinates and check if it's inside the area of interest
    centroid = p.centroid
    s = boros.geometry.intersects(centroid)
    t = pd.concat([boros,s],axis=1)

    if(True in s.values):
        polygon_desc['id_x'] = i
        polygon_desc['id_y'] = j
        polygon_desc['geometry'] = p
        polygon_desc['areas'] = ','.join(t[t[0] == True]['BoroName'].values).encode('UTF8')
polygons.append(polygon_desc)
```
Deal with mobility dataset

In [28]:
# Create the geoDataFrame and save it into the file
gdf = gpd.GeoDataFrame(polygons)
base = boros.geometry.plot(color='white', figsize=[8,8])
gdf.plot(ax=base)

Out[28]: `<matplotlib.axes._subplots.AxesSubplot at 0x117bcd150>`
Human mobility characterization
Human mobility characterization

Each individual is represented with a mobility trace as the sequence of places visited by an individual in a given period of time:

$$MT(t_1, t_2) = (Pl_1, Pl_2, ..., Pl_{N(t_1, t_2)})$$

where a place is defined as:

$$Pl_n = [ID, t^a, t^d, C]$$
Human mobility characterization

We characterize the mobility of an individual by computing a set of features based on its movements [5]:

**Distance traveled related features**
- Total traveled distance, total displacements

**Movements related features**
- Radius of gyration, number of geo-locations points

**Visited places related features**
- Diversity of visited places, number of unique visited places

Preliminary conclusions

• Human mobility understanding is about studying the laws of human motion

• To deal with trajectories, the space has to be segmented into regions

• Human mobility can be characterized with measures based on distance, location etc.

• Geopandas allows the use of spatial data in a dataframe
Questions?
Machine learning settings
Classification
Classify the next square the user will visit
[8] Human Mobility Prediction based on Individual and Collective Geographical Preferences. (Calabrese et al.)
[9] DeepTransport: Prediction and Simulation of Human Mobility and Transportation Mode at a Citywide Level (Song et al.)
[10] Next Place Prediction using Mobility Markov Chains (Gambs et al.)

Regression
Predict along the axis the values of next latitude and logitude coordinates
Classification

Probabilistic model to predict the location of a person over time based on individual and collective behaviors [8].

\[ P(x_{k+1}(u) = j | x_k(u) = i) = (1 - \alpha(k)) P_I(x_{k+1}(u) = j | x_k(u) = i) + (\alpha(k)) P_C(x_{k+1}(u) = j | x_k = i) \]

Individual trajectories of the user

Collective mobility

[8] Human Mobility Prediction based on Individual and Collective Geographical Preferences. (Calabrese et al.)
Humans use memory to perform actions.
• In reading a book, the understanding of each word is based on the understanding of previous words.
Recurrent Neural Network

Connect previous information to the present task

\[ h_t \rightarrow A \rightarrow X_t \]

\[ h_0 \rightarrow A \rightarrow X_0 \rightarrow A \rightarrow X_1 \rightarrow A \rightarrow X_2 \rightarrow A \rightarrow \ldots \rightarrow A \rightarrow X_t \]

RNN present problem in learning long-time term dependencies

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
LSTM Networks
LSTM Networks to predict coordinates

Increment on latitude and longitude

Neural Network

Enriched input feature representation

Time and cell representation
Easiest input:

- Time
- Cell representation (one-hot encoding)
Input with embedding of the cell:

- **Time**
- **Cell embedding (Word2Vec)**
A more complete input:
- Time
- Cell embedding (Word2Vec)
- Geographical features
The idea is to predict the increment that has to be applied along the latitude and longitude axis.
Evaluation

Classification

• Accuracy respect the cell

Regression

• Distance between the predicted point (lat, longt) and the target point.
Conclusion

• Two possible settings: Classification and Regression

• Geographical features can be combined with trajectories

• LSTM is a promising model to work with trajectories
Thank you!
Questions?

Gianni Barlacchi
gianni.barlacchi@gmail.com
@giannibarlacchi