Lower Dimensional Representations of City Neighbourhoods

Marzieh Saeidi
University College London
m.saeidi@cs.ucl.ac.uk

Sebastian Riedel
University College London
s.riedel@cs.ucl.ac.uk

Licia Capra
University College London
l.capra@ucl.ac.uk

Abstract

We aim to profile characteristics of areas of variant units across a district, city or a country. Studying attributes of areas can be very useful in several situations. In the past, research has focused mainly on studying specific characteristics of areas using a few selected attributes. In this paper we propose an alternative view on neighbourhood profiles. Instead of characterising a neighbourhood through a set of attributes such as those collected by the census, we propose use of a low-dimensional feature representation, or embedding, created from one or more input sources. The purpose of the embeddings is having a generic representation for entities that can do well across several downstream tasks such as regression for attributes prediction.

Introduction

Profiling city neighbourhoods are crucial for many reasons; For example it is important for policy makers to recognise areas with higher degree of deprivation or areas with communities that are less integrated. Equally when opening a new business it is important to identify neighbourhoods in which the business can potentially do well. For instance opening a coffee shop can be more profitable in a hipster area or a nursery in areas with higher intensity of families who have young kids. On a personal level, we are often faced with making decisions based on the attributes of areas or cities. For instance when buying or renting a place or going out, it is necessary to have access to information about areas.

Today the only source of data for some of these attributes is the census. Not only is census data expensive for the government to obtain, it is also limited in topic coverage. For example, it is unlikely for the government to run a poll on the population of homosexuals or hipsters. Currently most research on profiling cities and neighbourhoods focus on predicting specific attributes such as deprivation or crime.

We propose a shift of paradigm in profiling areas: creating low dimensional representations that contain necessary information about the areas and are capable of performing well across a range of downstream tasks. These embeddings are generated using one or more sources of information such as census data, text from social media, etc. and by applying representation learning techniques.

In this paper we create embeddings from census data using data compression methods such as principal components analysis and stacked auto-encoders. We test performance of those representations on regression tasks across London neighbourhoods using Gaussian Processes as a non-linear spatial regression model.

Method

The aim of this paper is to investigate generating low-dimensional embeddings for geographical units which is known as representation learning. To test the performance of embeddings we perform regression tasks.

Learning Representations

When the input data is of high dimension and there is possibly a redundancy or relation between input dimensions, it is preferred to transform data into lower dimension representation. This transformation is often called feature extraction or representation learning. A good transformation should learn the relevant information from the data in order to do well in downstream tasks. In this work, we investigate two dimensionality reduction techniques.

PCA : We use Principal component analysis (PCA) (Jolliffe 2005) to obtain a linear transformation of our data. PCA is the most common transformation technique. It transforms the data to be represented in terms of its principal components rather than the axis of the input space. Principal components are the directions where there is the most variance.

Stacked Autoencoder Inspired by the recent successes of deep learning methods in obtaining unsupervised representations for entities (Bengio, Courville, and Vincent 2013), we use stacked autoencoders (SAE) to generate embeddings for neighbourhoods using the census data. Embeddings are the features learned by the last hidden layer in the network. An autoencoder can contain one hidden layer or multiple hidden layers (stacked autoencoder). Often a deeper architecture can help to better learn the underlying factors of separation.

Regression

To test the performance of the learned representations, we use them in prediction tasks. We investigate linear regression and Gaussian Processes regression models to study the suitably of them for our framework. Ideally such regression model is able to make use of the spatial smoothness of attributes of neighbourhoods, e.g. areas with geographical closeness have similar characteristics. Moreover, a model that does better with less training points is preferable.

Linear regression is the most common regression model, however it can only capture the linear relationship between the entities. Gaussian Processes or GPs (Rasmussen 2006) not only support standard linear regression as a special case, they are also capable of modelling non-linear regression problems. Moreover GPs are particularly useful when we have access to only a few labelled samples. This is extremely useful since collecting data about the population or characteristics of an area is an expensive task.

Experiments

We use census data for England and Wales to generate the embeddings. Collectively we have 500 attributes across all selected categories of the census. We then try to predict several attributes such as crime that is not in the census data and also some of the attributes within the census data. However when predicting a specific attribute that exists in the census, we remove the category in which that attribute belongs to from the input data and for the generation of its lower dimensional embeddings. For instance if we predict the population of Asian, we remove all the data about ethnic groups from the census data and then learn the representations. The regression task is performed for the following attributes: crime, population of Asian, population of Jewish, percentage of population with the highest social grade, lowest social grade, and the number of people travel to work by bicycle in an area.

We heuristically choose the dimension of embeddings for both PCA and stacked autoencoder. In the case of the stacked autoencoder, we choose number of hidden layers and the hidden units so that they minimise the reconstruction error. Regression tasks are also performed on different sized embeddings., e.g. PCA 50 and PCA 20.

The units of analysis that are chosen for these task are called lower layer super output area(LSOA). LSOAs were created for the purpose of aggregation of census data. Population of an LSOA is around 3,000 people or 1,200 households. There are 34744 LSOAs across England and Wales. When creating the embeddings we use all the data across England and Wales. For the regression task, we only select units across Greater London. There are 4630 LSOAs in London. Regression is trained over subsets of 10 shuffles of the training data and validated over a fixed set of 800 units. The error of prediction for each experiment is measured in terms of normalized root mean squared error (NRMSE). Normalized mean squared error is RMSE normalized by the range of output variable.

We use Theano (Bergstra et al. 2010) for implementing a SAE and GPy (the GPy authors 2014) for Gaussian Processes regression.

Results

In this section we present the results of regression tasks to compare the performance of our embeddings. In particular we are interested to know if lower-dimensional representations are doing better than the raw high-dimensional census data. We also discuss which regression models are more appropriate for the task. Moreover to understand the behaviour of SAE, in the last section we present a map of England and Wales which is clustered using the embeddings from an SAE.

Model.

Preliminary regression results indicate that a GPs model does considerably better than a linear regression model especially when we have only a few data points. Moreover a GPs regression model with a non-linear kernel (Radial Basis Function) performs better than a GPs with a linear kernel.[1] Figure 1 shows the difference in performance between these two kernels. Therefore for the reminder of the tasks, we focus on using GPs with an RBF kernel.

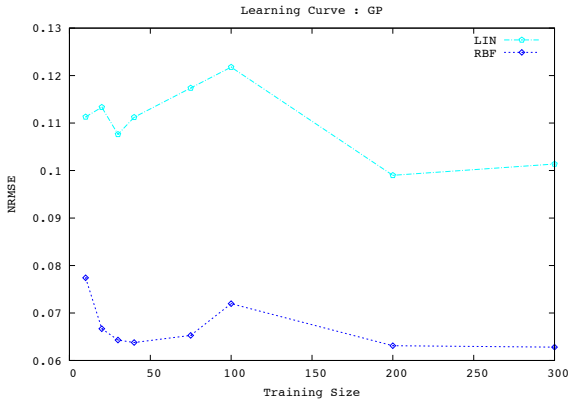

Figure 1: RBF kernel vs linear in GPs regression.

Embeddings. Figures 2 and 3 show that in general embeddings perform better than census data. Error bars indicate the variation of the errors is greater across different shuffles when predicting using census data. As results indicate, on average higher dimensional raw census data does poorly compare to lower dimensional embeddings. This is present across all the tasks. Census data performs better only for the task of predicting the number of people who travel to work by bicycle which is shown in figure 4.

As shown in Figure 5, none of the lower-dimensional embeddings is superior to others. One can argue that in this settings, PCA should be preferable since given the similar performance, it is more efficient to obtain the embeddings using PCA than a stacked autoencoder in terms of performance and the number of hyper-parameters that are needed to be tuned.

[1] A GPs with a linear kernel corresponds to a Bayesian Linear Regression.

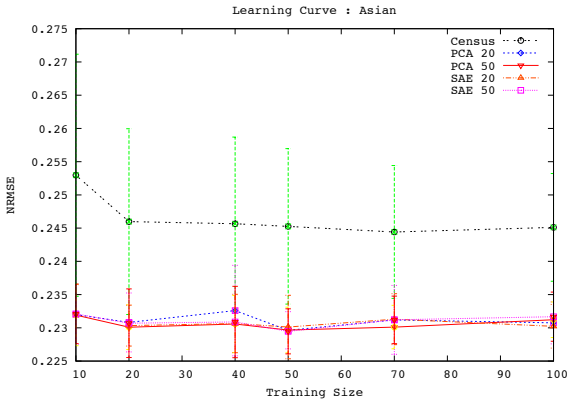

Figure 2: Performance of different embeddings for the task of predicting population of Asian

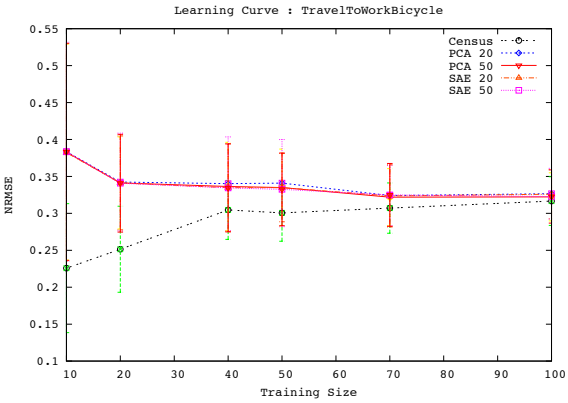

Figure 4: Census data does better than embeddings for the task of number of people travel to work by bicycle.

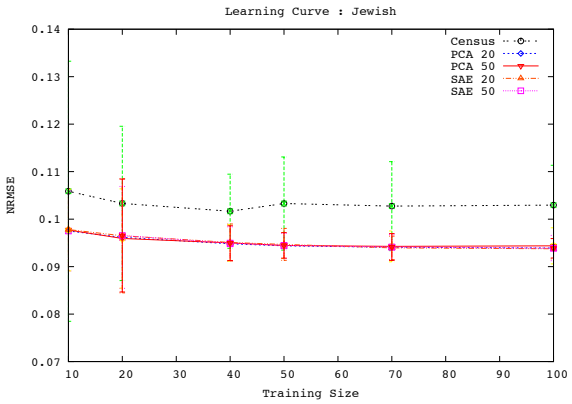

Figure 3: Performance of different embeddings for the task of predicting population of Jewish

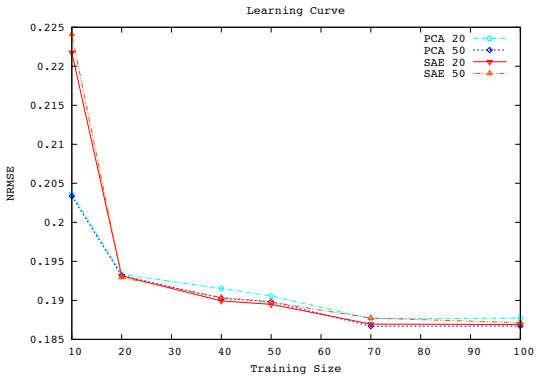

Figure 5: Performance of all low-dimensional embeddings across all tasks.

Effect of Locality. We expected that most attributes show a smoothness property across geographical regions, i.e. areas that are in proximate closeness are similar in terms of characteristics. For this reason we argue if it is necessary for the location information of a unit to be part of its embedding. To investigate this, we ran two experiments for each task: In the first experiments we only use the embeddings obtained from PCA or stacked autoencoder or the raw census data. In the second experiment, we add two attributes to the embeddings, the latitude and longitude of the centroid of each unit. This is to capture the geographical closeness of units to each other. Results however show that adding coordinates did not result in a major improvement. This should be investigated further in the future work.

Clustering Effect. Using the embeddings obtained from an SAE we cluster our input areas across England and Wales. This is done by assigning to each hidden unit those areas which cause the highest activation for the unit. Moreover, we interpret each hidden unit through input attributes by assigning to it the input units that are connected with the highest weights (Erhan et al. 2009). Figure 6 shows the clustering obtained from a 50-unit hidden layer in a SAE. Clusters with the orange colour (with arrows pointing to them) are the ones that are associated with attributes that represent population of migrants. We can see that cities such as London, Birmingham and Leeds contain areas characterised by this cluster.

Discussion

In this work, we introduced the notion of generating lower dimensional embeddings of neighbourhoods where these embeddings manifest the underlying factors of separation. Such embeddings are generated so that they perform well across several downstream tasks. We use census data for generating such embeddings however other sources of data can be added such as information about amenities in an area or the data from social media platforms.

Using the textual data from social media platforms such as twitter, Foursquare and Facebook has been subject of many research in the recent years. It has been shown that some characteristics of neighbourhoods demonstrate high correlation with the geographically-related data (often geo-tagged) obtained from these platforms. In the absence of census data (census data is not very frequent and can be often stale), we can use frequently updated social media data for inferring

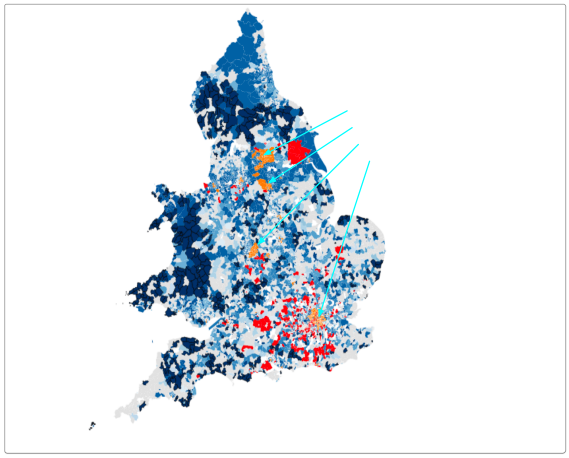

Figure 6: Clustering with respect to the hidden units of an autoencoder

attributes of areas. This data can also be used together with the census data to generate the embeddings.

Furthermore, we aim to profile cities in terms of attributes that are not covered in census such as hipster-ness or vibrancy. We show that a non-linear regression model such as GPs with an RBF kernel can do better with the availability of only few data points. This is crucial as collecting ground truth data for neighbourhoods is an expensive task.

Related Work

In this work, we investigate the notion of generating low-dimensional representations for geographical units, across a country or a city. Even though in this work, we only use census data for the creation of embeddings, we aim to make use of the massive amount of available textual data from geo-related social media in the future work. We can place our research within two different fields: creating generic low dimensional representation of entities and urban data mining.

Creating generic low dimensional embeddings of entities has been used pattern recognition, image processing, signal processing and more recently Natural Language Processing (NLP). For example Mikolov et al. (Mikolov et al. 2013) developed a tool (Mikolov et al.) that can provide embeddings for all the terms given a corpus. This embeddings can be used by the community in many downstream NLP tasks. We want to do a similar work but in the context of city neighbourhoods.

Census data has been used for predicting specific attributes; For example Hentschel et al. study spatial dimensions of poverty using census data (Hentschel et al. 2000). However recently research on urban data mining is inspired by the availability of location based social networks. Some only use the activities of users in such platforms (Wakamiya, Lee, and Sumiya 2011) (Noulas, Mascolo, and Frias-Martinez 2013) and some others take advantage of the textual contents provided by such platforms (Quercia, Séaghdha, and Crowcroft 2012) (Quercia et al. 2012). The results of these works are very interesting and show that social media data can be used in profiling urban area characteristics. These works however focus on profiling only one specific attribute. Whereas in our work we try to create generic embeddings that do well for profiling several attribute.

References

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35(8):1798–1828.

Bergstra, J.; Breuleux, O.; Bastien, F.; Lamblin, P.; Pascanu, R.; Desjardins, G.; Turian, J.; Warde-Farley, D.; and Bengio, Y. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*.

Erhan, D.; Bengio, Y.; Courville, A.; and Vincent, P. 2009. Visualizing higher-layer features of a deep network. *Dept. IRO, Université de Montréal, Tech. Rep*.

Hentschel, J.; Lanjouw, J. O.; Lanjouw, P.; and Poggi, J. 2000. Combining census and survey data to trace the spatial dimensions of poverty: A case study of ecuador. *The World Bank Economic Review* 14(1):147–165.

Jolliffe, I. 2005. *Principal component analysis*. Wiley Online Library.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. word2vec.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Noulas, A.; Mascolo, C.; and Frias-Martinez, E. 2013. Exploiting foursquare and cellular data to infer user activity in urban environments. In *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, volume 1, 167–176. IEEE.

Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2012. Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 965–968. ACM.

Quercia, D.; Séaghdha, D. Ó.; and Crowcroft, J. 2012. Talk of the city: Our tweets, our community happiness. In *ICWSM*.

Rasmussen, C. E. 2006. Gaussian processes for machine learning.

the GPy authors. 2014. GPy: A gaussian process framework in python. https://github.com/SheffieldML/GPy.

Wakamiya, S.; Lee, R.; and Sumiya, K. 2011. Urban area characterization based on semantics of crowd activities in twitter. In *GeoSpatial Semantics*. Springer. 108–123.