Machine Learning Project 2: Unsupervised Analysis of the Human Development Report

1 Goal of the Project

In the "unsupervised learning" part of the course, you learned what unsupervised learning is. In particular, you studied two types of unsupervised problems: clustering and visualization. In this project, you will use the k-means algorithm to cluster countries according to their characteristics in the Human Development Report and t-SNE to visualize the resulting clusters. You are expected to handle out a (concise) report of max. 4 pages, including plots. Concision will be considered in the grades. For implementation, you will use the `scikit-learn` Python framework (`http://scikit-learn.org`). A Python implementation of t-SNE is also provided. Note that the provided implementations are written in Python2, which means that their behavior in Python3 is not determined. You are free to choose the implementation of your choice (in Python2 or Python3) to complete this assignment. Your code has to be sent with your report.

2 Data Extraction

Each year, the Human Development Report Office of the United Nations Development Program publishes the Human Development Report on `http://hdr.undp.org`. You will use the data released in 2007 which consists of 45 indicators for 138 countries. Cleaning has already been done, by replacing missing values by mean feature values and removing some outlying/abnormal countries/indicators. The country names and indicators are described in Tables 1 and 2.

The dataset is stored in the file `hdr_data.dat` as a Python dictionary with fields `X`, `country_names`, `indicator_names` and `indicator_descriptions`. This dictionary can be loaded with the `load_HDR_data` function in the `utils.py` file. In this project, there is no distinction between training and test data. Data should be normalized (indicators are very different) with the tools described here:

- `http://scikit-learn.org/stable/modules/preprocessing.html#standardization-or-mean-removal-and-variance-scaling`
- `http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

001	Norway	036	Bahrain	071	Tunisia	106	Ghana
002	Iceland	037	Estonia	072	Fiji	107	Bangladesh
003	Australia	038	Lithuania	073	Paraguay	108	Nepal
004	Ireland	039	Slovakia	074	Turkey	109	Papua New Guinea
005	Sweden	040	Uruguay	075	Sri Lanka	110	Sudan
006	Canada	041	Croatia	076	Dominican Republic	111	Uganda
007	Japan	042	Latvia	077	Belize	112	Togo
008	United States	043	Costa Rica	078	Iran, Islamic Rep. of	113	Zimbabwe
009	Switzerland	044	United Arab Emirates	079	Georgia	114	Madagascar
010	Netherlands	045	Mexico	080	Azerbaijan	115	Cameroon
011	Finland	046	Bulgaria	081	El Salvador	116	Swaziland
012	Luxembourg	047	Trinidad and Tobago	082	Algeria	117	Yemen
013	Belgium	048	Panama	083	Guyana	118	Kenya
014	Austria	049	Oman	084	Jamaica	119	Gambia
015	Denmark	050	Romania	085	Cape Verde	120	Senegal
016	France	051	Malaysia	086	Syrian Arab Republic	121	Rwanda
017	Italy	052	Mauritius	087	Indonesia	122	Nigeria
018	United Kingdom	053	Russian Federation	088	Viet Nam	123	Guinea
019	Spain	054	Macedonia, TFYR	089	Kyrgyzstan	124	Angola
020	New Zealand	055	Belarus	090	Egypt	125	Tanzania, U. Rep. of
021	Germany	056	Brazil	091	Nicaragua	126	Benin
022	Israel	057	Colombia	092	Moldova, Rep. of	127	Côte d'Ivoire
023	Greece	058	Venezuela, RB	093	Bolivia	128	Zambia
024	Singapore	059	Albania	094	Mongolia	129	Malawi
025	Korea, Rep. of	060	Thailand	095	Honduras	130	Mozambique
026	Slovenia	061	Saudi Arabia	096	Guatemala	131	Burundi
027	Portugal	062	Ukraine	097	South Africa	132	Ethiopia
028	Cyprus	063	Lebanon	098	Morocco	133	Chad
029	Czech Republic	064	Kazakhstan	099	Gabon	134	Central African Republic
030	Malta	065	Armenia	100	Namibia	135	Burkina Faso
031	Kuwait	066	China	101	India	136	Mali
032	Hungary	067	Peru	102	Cambodia	137	Sierra Leone
033	Argentina	068	Ecuador	103	Botswana	138	Niger
034	Poland	069	Philippines	104	Lao People's Dem. Rep.		
035	Chile	070	Jordan	105	Pakistan		

Table 1: Country names in the HDR 2007 dataset.

Your first task is to load the HDR dataset. Quickly look into the dictionary fields to see what kind of data is provided. Then, normalize the data for your unsupervised analysis. This task does not need to be described in the report.

3 Clustering and Visualization

Once the data are normalized, you can cluster the instances into several groups. However, you do not know *how many* clusters can be found in the data. Hence, you need to try several number of clusters and choose the clustering which makes the most sense for you. Use the k-means algorithm which is described here:

- `http://scikit-learn.org/stable/modules/clustering.html#k-means`
- `http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

In order to choose the right number of clusters, you need to visualize the clustering itself. First, you can look at the cluster centers. Since they are obtained with k-means, they do not correspond to a particular country (each cluster center is the mean of the countries in the corresponding cluster). However, you can look at the name/indicators of the countries which are the closest to the cluster centers. These countries can be obtained with the `find_closest_instances_to_kmeans`

Name	Interpretation
Pop growth	Annual population growth rate (%) 1975-2004
Pop growth 2004	Annual population growth rate (%) 2004
Price index	Average annual change in consumer price index (%) 1990-2004
Carbon Dioxide 2003	CO2 emissions - per capita (mertic ton) 2003
Export 1990	Export of goods and services (% of GDP) 1990
Export 2004	Export of goods and services (% of GDP) 2004
Elec 2003	Electricity comsumption per capita (kW/h) 2003
GDP	GDP (US$ billions) 2004
GDP PPP	GDP (PPP US$ billions) 2004
GDP pc	GDP per capita (US$) 2004
GDP pc growth rate	GDP per capita growth rate (%) 1990-2004
Fem Econo Rate	Female economic activity rate (% age 15 and older) 2004
Fem Econo 1990	Female economic activity rate (index, 1990=100, % age 15 and older) 1990
Fem Econo 2004	Female economic activity rate (index, 1990=100, % age 15 and older) 2004
Health Exp	Health expenditure per capita (PPP US$) 2003
Babies	Infant with low birth weight (%) 1996-2004
Internet 1990	Internet users (per 1,000 people) 1990
Import 1990	Import of goods and services (% of GDP) 1990
Import 2004	Import of goods and services (% of GDP) 2004
Tertiary female ratio	Gross tertiary enrolment - ratio of female to male 2004
Babies immunized	One-years-olds fully immunized against measles (%) 2004
Manufactured Exp 2004	manufactured exports (% of merchandise exports) 2004
Foreign invest 2004	Net foreign direct investment inflows (% GDP) 2004

Name	Interpretation
Military 2004	Military expenditure (% GDP) 2004
Public Health 2003	Public health expenditure (% of GDP) 2003
Private Health 2003	Private health expenditure (% of GDP) 2003
Primary export 2004	Primary exports (% of meschandise exports) 2004
Public Health	Public expenditure on health (% of GDP) 2003-2004
Refugees asylum	Refugees by country of asylum (thousands) 2005
Refugees origin	Refugees by country of origin (thousands) 2005
Armed forces	Total armed forces (thousands) 2006
Parliament Women	Seats in parliament held by women (% total)
Female Male income	Ratio estimated female to male earned income
House women 2006	Seats in lower house or single house held by women (% total)
Pop 1975	Total population 1975 (millions)
Pop 2004	Total population 2004 (millions)
Pop 2015	Total population 2015 (millions)
Tuberculosis detected	Tubreculosis cass detected under DOTS (%) 2004
Tuberculosis cured 2004	Tubreculosis cass cured under DOTS (%) 2004
Trad fuel	Traditional fuel consumption (% total energy requirements) 2003
ODA pc donnor 2004	ODA per capita of donor country (2004 US$) 2004
ODA to least dev 1990	ODA to least developed countries (% of total) 1990
ODA to least dev 2004	ODA to least developed countries (% of total) 2004
ODA received	Official developement assistance (ODA) received (net disbursements) Total (US $ millions)
ODA received pc	Official developement assistance (ODA) received (net disbursements) per capita (US $)

Table 2: Indicator names in the HDR 2007 dataset.

function in the `utils.py` file. Second, you need to reduce the dimensionality of the data in order to visualize them on your computer screen. This can be done with the `tsne` function in the `tsne.py` file. Eventually, the data can be visualized (using the 2D coordinates given by t-SNE and the country names) with the `show_annotated_clustering` function in the `utils.py` file.

> Your second task is to cluster the HDR data with the k-means algorithm for $k = 2 \ldots 10$ clusters. Choose one clustering which contains not enough clusters, one clustering which contains enough clusters and one clustering which contains too many clusters. Show the t-SNE visualization of these clusterings (and only those) in the report. In each case, give the names of the countries which are the closest to the centers. Explain/discuss why you chose each of these three clusterings (not enough clusters, enough clusters, too many clusters). In each case, can you give an interpretation of the clusters? Notice that your choice (e.g. of the "good number" k of clusters) is subjective and may be different from the one of other students working with the same methods on the same data.

Tip: the quality of the visualizations obtained with t-SNE may be sensitive to its meta-parameter: the perplexity. The right perplexity (interpreted as a soft number of neighbors) must be chosen. Notice that the t-SNE coordinates do not depend on the clustering, i.e. they can be precomputed before each clustering with $k = 2 \ldots 10$ clusters. This way, you will be able to compare the clusterings on similar visualizations

Tip: if you need to open several figures to display the clusterings, you can use the function `matplotlib.pyplot.figure`. Also, the `find_closest_instances_to_kmeans` function returns two sets of values, which are retrieved as:
`closest_instances, closest_indices = find_closest_instances_to_kmeans(...)`.