Info 290: Data Mining In Intelligent Business Systems

Group members: Faye Ip, Chalenge Masekera, Ramit Malhotra
Git handles: @fayeip, @ramitmalhotra, @chalenge
Project: Analyzing Public Health Data
April 9, 2014

Objective: to study patterns and outcomes of patient hospital visits at a Massachusetts hospital from 2006 to 2008.

Data set:

	A	B	C	D	E	F	G	H	I	J
1	ICD_9_CODE	DISPO	AGE	SEX	RACE	ZIP	Insur	INSURANCE	DOS	YEAR
34	274.0	1	43	M	Irish	01104	BLUE CROSS O MA-HM(	4	1-Oct-07	3
35	724.3	1	23	M	White	01071	BLUE CROSS O MA-HM(	4	1-Oct-07	3
36	913.0	1	15	M	White	01151	BLUE CROSS O MA-HM(	4	1-Oct-07	3
37	288.00	1	57	F	White	01060	BLUE CROSS O MA-HM(	4	1-Oct-07	3
38	789.00	1	35	F	White	06082	BLUE CROSS O MA-HM(	4	1-Oct-07	3
39	812.09	1	15	F	White	01085	BLUE CROSS O MA-HM(	4	1-Oct-07	3
40	786.2	1	32	M	White	01104	BLUE CROSS O MA-HM(	4	1-Oct-07	3
41	786.01	1	18	M	American	01056	BLUE CROSS O MA-HM(	4	1-Oct-07	3
42	789.00	1	31	F	White	01020	BLUE CROSS O MA-HM(	4	1-Oct-07	3
43	493.92	1	26	M	Puerto Rican	01104	BLUE CROSS- PPO	4	1-Oct-07	3
44	320.1	0	51	F	White	06082	BLUE CROSS- PPO	4	1-Oct-07	3
45	311	1	19	M	American	01104	BLUE CROSS- PPO	4	1-Oct-07	3
46	873.43	1	18	F	American	01028	BLUE CROSS- PPO	4	1-Oct-07	3
47	780.2	0	33	M	Unknown/Not Speci	01033	BLUE CROSS- PPO	4	1-Oct-07	3
48	784.0	1	18	F	White	01040	BLUE CROSS- PPO	4	1-Oct-07	3
49	847.0	1	39	M	African American	01104	BLUE CROSS- PPO	4	1-Oct-07	3
50	786.59	0	64	M	White	01095	BLUE CROSS- PPO	4	1-Oct-07	3
51	813.44	1	12	F	White	01028	BLUE CROSS- PPO	4	1-Oct-07	3
52	845.00	1	28	M	White	01077	BLUE CROSS- PPO	4	1-Oct-07	3
53	719.46	1	48	M	White	06103	BLUE CROSS- PPO	4	1-Oct-07	3
54	786.50	1	42	F	Unknown/Not Speci	01107	BLUE CROSS- PPO	4	1-Oct-07	3
55	845.00	1	18	M	White	01119	BLUE CROSS- PPO	4	1-Oct-07	3
56	789.09	1	84	F	Unknown/Not Speci	01104	Blue Care 65	4	1-Oct-07	3
57	486	0	79	M	White	01118	Blue Care 65	4	1-Oct-07	3
58	428.0	0	94	F	White	01001	Blue Care 65	4	1-Oct-07	3
59	434.91	0	80	F	White	01013	Blue Care 65	4	1-Oct-07	3
60	728.85	1	71	F	White	01108	Blue Care 65	4	1-Oct-07	3
61	913.0	1	5	M	White	01106	BLUE CROSS- CT ANTH	4	1-Oct-07	3
62	796.3	1	36	M	White	06082	BLUE CROSS- CT ANTH	4	1-Oct-07	3
63	826.0	1	15	F	White	01151	HARVARD PILGRIM HM(	4	1-Oct-07	3

The dataset contains ~300,000 records, each representing one hospital visit. The data is currently in a spreadsheet which we intend to export to csv and bring into python. The ICD-9 code identifies the illness for which the patient was treated. The Disposition is the outcome, whether the patient was discharged, etc.

The data is mostly consistent except for the “Race” column, where there are some “Unknown” values and some overlapping categories, i.e. some people put “White” which is a skin color and others put “Spanish” which is a nationality.

Related studies:
A previous study done using the same data set is “Emergency Department Utilization After the Implementation of Massachusetts Health Reform” by Smulowitz et al. The abstract can be found

here: http://www.ncbi.nlm.nih.gov/pubmed/21570157. That study focused on the use of the emergency room before and after health care reform passed in Massachusetts for conditions potentially amenable to primary care.

Preliminary data analysis:
- Find top illnesses that result in hospital visits
- Average age and demographics of hospital patients
- Percentage of hospital visits which are emergency room visits.

Algorithms:
1. Clustering - We intend to use the KNN algorithm to do clustering of the data. Some potential interesting clusters might be, for example, that people of a certain age group and ethnicity might be more susceptible to a certain illness, or time of year and emergency room visits. To calculate the distance, we intend to try different similarity measures such as the Jaccard and Manhattan.

2. Frequent Patterns - We can treat each dimension as an item in a bucket, and use the Apriori algorithm to see which buckets appear frequently together. For example, maybe people who live in a zip code between 01101 and 01200 and who are between 50-60 years old tend to be people who have a certain illness.

3. SVM - Classification and regression to predict illnesses that a person of certain characteristics (age, location, race and sex) may have.

4. Outliers - We can look at the attributes and to search for rare occurrences of patient treatments. For example, a patient under the age of 10 being treated for a disease normally associated with old age or a patient being treated for a disease not common in their zip code.

Questions for Jimmy/Shreyas:

1. What python libraries should we use, if any?

2. Are the four algorithms we outlined too extensive? Can you comment on the extent of our proposed scope?