Kathy Byun

Raye Liu

Population segmentation for health system management

List of Questions

* How to know if we have good group, how to know if group make sense
* JHU ACG, can we have access? Run algorithm on mimic iii

Mimic-III clinical database consists of health data of forty thousand patients in Beth Israel Deaconess Medical Center in Boston from 2001 to 2012. The dataset includes information such as diagnosis, demographics, prescriptions, ICU stays, procedures, vital signs and mortality.

International Statistical Classification of Diseases and Related Health Problems, ICD-9 code, is a list of codes used to classify disease, symptoms and external causes of disease. ICD-9 code was used to categorize diseases.

Diabetes Matrix Description

The matrix contains diabetes patients administered for ICU. The matrix has 10,310 rows × 931 columns. Each row represents a patient.

Patients.csv, D\_ICD\_DIAGNOSES.csv, DIAGNOSES\_ICD.csv, ADMISSIONS.csv and ICUSTAYS.csv are joined to create the matrix with diagnoses matched to patients, ICD9 code, disease categories and demographic information.

Numerical columns:

\* LOS is the length of the ICU stay of a patient. The length of stay is measured in days.

\* Hospitalization is the number of hospitalization of each patient.

Categorical/ Binary columns:

\* Admission type columns: elective, emergency, newborn, urgent

\* Marital status columns: divorced, life partner, married, NaN, separated, single, unknown (default), widowed

\* Insurance columns: government, Medicaid, Medicare, private, self-pay

\* Ethnicity columns: there are 40 ethnicity columns. ASIAN, WHITE, BLACK/AFRICAN AMERICAN, HISPANIC OR LATINO, etc.

\* Gender: 1 for males. 0 for females

\* Dead: 1 for death. 0 otherwise

\* ICD9 code columns: Column names are 3-4 digit alphanumerics. 1 if a patient has a disease with the ICD9 code, 0 otherwise

Ordinal column:

\* ORDINAL\_AGE column: the following 10 age groups are assigned to 1-10 in ascending order. AGE: 0-10, AGE: 11-20, AGE: 21-30, AGE: 31-40, AGE: 41-50, AGE: 51-60, AGE: 61-70, AGE: 71-80, AGE: 81-89, AGE: 90+

Diabetes\_Numerical.csv and Diabetes\_Category.csv are the same matrices.

Diabetes\_Numerical.csv column datatypes are all numbers. Diabetes\_Category.csv column datatypes are numbers for numerical columns and categories for categorical columns

## Low Rank Model

K-means clustering is used to find clusters of similar patients. Using low rank model, the diabetes patient matrix was fitted to k-means algorithm with different k values. Low rank model is used in GLRMfinal.ipynb

1.K means with different loss functions (hinge)

2.K means with different loss function and param (inner iteration)

3.Keamns with different weighted loss function, param and (weighted loss function)

Objective function and fcross validation error

Performance on prediction

1.K means with different loss functions (hinge) → glrm\_ final

k=6

iteration 30: objective value = 457844.11504416

Train error = 0.048182, test error = 0.0478441

|  |  |  |
| --- | --- | --- |
|  | Train Error | Test Error |
|  | 0.048182 | 0.0478441 |
|  | 0.0478393 | 0.0481871 |

k=5

Iteration 30: objective value = 458976.40695412154

Train error = 0.0482869, test error = 0.0482868

|  |  |  |
| --- | --- | --- |
|  | Train Error | Test Error |
|  | 0.0482773 | 0.0478441 |
|  | 0.0478393 | 0.0481871 |

k=4

Iteration 30: objective value = 458055.4999498729

Train error = 0.0482435 test error = 0.0481639

k=3

Iteration 30: objective value = 457943.0327361399

Train error = 0.048235, test error = 0.0482053

2.K means with different loss functions and inner iteration 40 → glrm\_kmeans

K = 7 objective value = 455586.3300457709

k= 6 objective value = 457239.5370158994

K = 5 objective value = 458976.40695412154

|  |  |  |
| --- | --- | --- |
|  | Train Error | Test Error |
|  | 0.0482773 | 0.048291 |
|  | 0.0482869 | 0.0482868 |

K = 4 objective value = 458055.4999498729

|  |  |  |
| --- | --- | --- |
|  | Train Error | Test Error |
|  | 0.0481591 | 0.0482537 |
|  | 0.0482435 | 0.0481639 |

K = 3 objective value = 457943.0327361399

|  |  |  |
| --- | --- | --- |
|  | Train Error | Test Error |
|  | 0.048235 | 0.0482053 |
|  | 0.0482114 | 0.048249 |

K = 6 has the smallest objective value of 457239.5370158994. k =3 has the second smallest objective value of 457943.0327361399.

Train test errors are similar across the different k values

For param 100 glrm

K = 4 objective value = 458896.69003507175

|  |  |  |
| --- | --- | --- |
|  | **train\_error** | **test\_error** |
|  | **Float64** | **Float64** |
| **1** | 0.0480938 | 0.0485145 |
| **2** | 0.0484729 | 0.0480695 |

Predictive Performance

K = 6 hinge6 decision tree

Cluster 1 Score: 91.79407462701668

Cluster 2 Score: 108.23295410287105

Cluster 3 Score: 211.3686790483595

Cluster 4 Score: 135.3733607617661

Clutter 5 Score: 205.66740605137792

Cluster 6 Score: 82.73244305555494  
Max score: 211.3686790483595

k=6 param6 inner iteration = 40 decision tree

Cluster 1 Score: 124.37333457434498

Cluster 2 Score: 71.99592873953884

Cluster 3 Score: 118.676595133789

Cluster 4 Score: 143.7226363661737

Cluster 5 Score: 140.8101663410238

Cluster 6 Score: 98.7513866921253

Max score: 143.7226363661737

K = 6 param100\_6 inner iteration= 100 decision tree

Cluster 1 Score: 118.60164252778323

Cluster 2 Score: 98.09954845387755

Cluster 3 Score: 103.82216194321512

Cluster 4 Score: 170.67940930799213

Cluster 5 Score: 76.38225196635935

Cluster 6 Score: 110.70070526238044

Max score: 170.67940930799213

3 clusters score greater than and 3 clusters score smaller than param 6

K = 4 hinge4 prediction performance

Cluster 1 Decision Tree Score: 71.07275675466538

Cluster 2 Decision Tree Score: 126.62211856517415

Cluster 3 Decision Tree: Score: 114.20062898677885

Cluster 4 Decision Tree Score: 136.53558997669276

Max score: 136.53558997669276

k=4 param100\_4

Cluster 1 Score: 95.92851193960264

Cluster 2 Score: 127.72917139157506

Cluster 3 Score: 92.4829874758904

Cluster 4 Score: 121.85875849432557

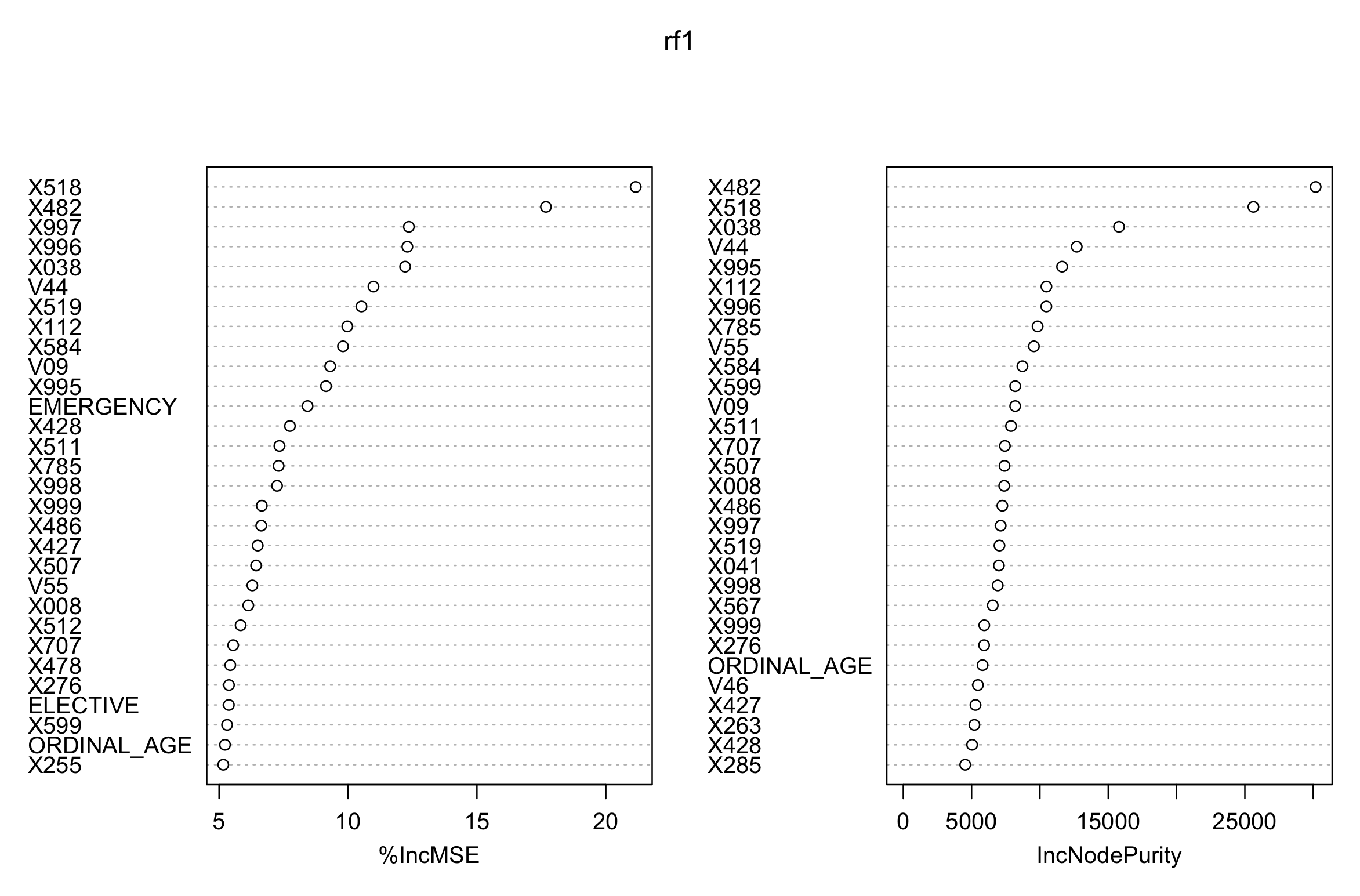
Max score: 127.72917139157506

Cluster summary statistics

Characteristics of clusters: all columns of each cluster has similar mean to columns of the original dataframe

1. k=7 hinge100\_7, printcluster(param100\_7, df\_pat, 7)
2. means of variables in each cluster
   1. Cluster 1 LOS: 6.411285, Hospitalization: 1.456417, ordinal age: 7.357584, most of them emergency: 0.868909
   2. Cluster 2 LOS: 7.163089, Hospitalization: 1.545215, ordinal age: 7.322772
   3. For All clusters feature mean for each cluster similar to total dataframe feature mean
3. k=6 hinge100\_6: For All clusters feature mean for each cluster similar to total dataframe feature mean
4. k=4
5. Glrm kmeans with modified loss function weight

Figure: random forest to predict ICU, IncMSE



Important features are features with highest %incMSE

518: Other diseases of lung. A disorder characterized by the collapse of part or the entire lung. Absence of air in the entire or part of a lung, such as an incompletely inflated neonate lung or a collapsed adult lung

0.327536

482: Other bacterial pneumonia.

0.062112

997 : Complications affecting specified body system not elsewhere classified.

* [**997.0** Nervous system complications](https://icd.codes/icd9cm/9970)
* [**997.1** Cardiac complications, not elsewhere classified](https://icd.codes/icd9cm/9971)
* [**997.2** Peripheral vascular complications, not elsewhere classified](https://icd.codes/icd9cm/9972)
* [**997.3** Respiratory complications not elsewhere classified](https://icd.codes/icd9cm/9973)
* [**997.4** Digestive system complications not elsewhere classified](https://icd.codes/icd9cm/9974)
* [**997.5** Urinary complications, not elsewhere classified](https://icd.codes/icd9cm/9975)
* [**997.6** Amputation stump complication](https://icd.codes/icd9cm/9976)
* [**997.7** Vascular complications of other vessels](https://icd.codes/icd9cm/9977)
* [**997.9** Complications affecting other specified body systems not elsewhere classified](https://icd.codes/icd9cm/9979)

996: Complications peculiar to certain specified procedures.

* Specific code [996.00](http://www.icd9data.com/2012/Volume1/800-999/996-999/996/996.00.htm) Mechanical complication of unspecified cardiac device, implant, and graft [convert 996.00 to ICD-10-CM](http://www.icd10data.com/Convert/996.00)
* Specific code [996.01](http://www.icd9data.com/2012/Volume1/800-999/996-999/996/996.01.htm) Mechanical complication due to cardiac pacemaker (electrode) [convert 996.01 to ICD-10-CM](http://www.icd10data.com/Convert/996.01)
* Specific code [996.02](http://www.icd9data.com/2012/Volume1/800-999/996-999/996/996.02.htm) Mechanical complication due to heart valve prosthesis [convert 996.02 to ICD-10-CM](http://www.icd10data.com/Convert/996.02)
* Specific code [996.03](http://www.icd9data.com/2012/Volume1/800-999/996-999/996/996.03.htm) Mechanical complication due to coronary bypass graft [convert 996.03 to ICD-10-CM](http://www.icd10data.com/Convert/996.03)
* Specific code [996.04](http://www.icd9data.com/2012/Volume1/800-999/996-999/996/996.04.htm) Mechanical complication of automatic implantable cardiac defibrillator [convert 996.04 to ICD-10-CM](http://www.icd10data.com/Convert/996.04)
* Specific code [996.09](http://www.icd9data.com/2012/Volume1/800-999/996-999/996/996.09.htm) Other mechanical complication of cardiac device, implant, and graft [convert 996.09 to ICD-10-CM](http://www.icd10data.com/Convert/996.09)

038: Septicemia: blood poisoning, especially that caused by bacteria or their toxins.

Mean: 0.170600

112:

Mean 0.051760

Candidiasis. A condition in which candida albicans, a type of yeast, grows out of control in moist skin areas of the body.

Related to diabetes

Predictive Performance of Low Rank Model

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Since the objective value of our prediction is quantitive variable LOS, therefore we use the regression tree. We do separate trainings for different clusters produced by the general low rank model training, as well as for the whole data set in Diabetes\_Numerical.csv. The train-test set split is 75%-25% of the target dataframe, and the mean square errors are:

1.K means with different loss functions (hinge)

K = 6

Cluster 1 Score: 91.79407462701668

Cluster 2 Score: 108.23295410287105

Cluster 3 Score: 211.3686790483595

Cluster 4 Score: 135.3733607617661

Clutter 5 Score: 205.66740605137792

Cluster 6 Score: 82.73244305555494  
Max score: 211.3686790483595

k=5

Cluster 1 Score: 115.16327501139055

Cluster 2 Score:131.60936177178976

Cluster 3 Score: 92.46493992424074

Cluster 4 Score: 110.83058711641635

Clutter 5 Score: 110.06729997609855  
Max score: 131.60936177178976

k=4

Cluster 1 Decision Tree Score: 71.07275675466538

Cluster 2 Decision Tree Score: 126.62211856517415

Cluster 3 Decision Tree: Score: 114.20062898677885

Cluster 4 Decision Tree Score: 136.53558997669276

Max score: 136.53558997669276

k=3

Cluster 1 Decision Tree Score: 87.1130393644

Cluster 2 Decision Tree Score: 165.41786262812693

Cluster 3 Decision Tree: Score: 117.60504860109836

Max score: 165.41786262812693

The R^2 for each training set is:

(Placeholder for screenshot of the training k=7).

(Placeholder for screenshot of the training k=6).

(Placeholder for screenshot of the training k=5).

(Placeholder for screenshot of the training on the whole dataset).

2.K means with different loss functions and inner iteration 40

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The training is similar as that of the decision tree, however, we use 500 trees with 10 variables at each split in the random forest training and obtain the mean square errors and the R^2 values.

(Placeholder for screenshot of the training on diabetes numerical).

Random forest evaluates each factors’ contribution towards the mean square errors and puts weights on each of them. From the diagram below, we select the top 20 (or another number) of the factors and put them in training again for the general low rank model.

(Placeholder for screenshot of the training on the graph for factor weights).

(Need a new graph)

(Placeholder for screenshot of the training on the factor weights coefficients).

Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

We use linear regression models as a baseline evaluation for the performances of other forthcoming models we are trying. The performance of the linear regression model is considerably poor on the diabetes patients’ subgroup dataset, even after tuning the model with non-linear terms.

1.K means with different loss functions (hinge)

k = 6

Cluster 1: 6.4553564721143735e+22

Cluster 2: 2.6163306751921842e+22

Cluster 3: 4.5454524460558475e+19

Cluster 4: 4.554160043204208e+20

Clutter 5: 6.192207418609172e+22

Cluster 6: 1.9742893196383015e+22  
Max score: 6.4553564721143735e+22

k=5

Cluster 1: 3.133211643586004e+20

Cluster 2: 1.5483866834806618e+21

Cluster 3: 2.920527363935486e+21

Cluster 4:5.451606501473484e+22

Clutter 5: 3.448475511420542e+23

Max score: 3.448475511420542e+23

k=4

Cluster 1: 1.5978054899701247e+21

Cluster 2: 8.714307290317095e+21

Cluster 3: 1.6016118399737103e+23

Cluster 4: 6.465389092682245e+19

Max score: 1.6016118399737103e+23

k=3

Cluster 1: 2.8919513891721223e+22

Cluster 2: 1.6551507799146446e+23

Cluster 3: 2.0447264347063947e+21

Max score: 1.6551507799146446e+23

2.K means with different loss functions and inner iteration 40

(Placeholder for screenshot of the training on the whole dataset).

Logistic Regression

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

In our training approach, we define the top 5% percent of the patients with large LOS values as the “1” group, standing for high emergency health care resources consumers, the other 95% percent of patients are “0” group. This approach resembles the HRUPoRT of Canadian population survey data from the paper *“Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System”* by L. C. Rosella.

Training result: to be out.

(Placeholder for screenshot of the training on the whole dataset).

(Placeholder for screenshot of the R^2 on the whole dataset).

Gaussian Copula Model

John’s Hopkins ACG

Clustered diseases into ACG based on medical knowledge, severity

Patients in MCG

Improvement in Predicting the cost and length of stay with ACG

Predicting High Health Care Resource Utilization