**The things that I have done:**

1. still doing the research on article "HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on genetic defect(find out that it analyzed HRDetect predicts BRCA1/2 gene function defects, it utilized lasso logistic regression for classification, the validation procedure is done by using sensitivity and specificity, the feature that they used for classification are (**microhomology-mediated deletions (2.398), base-substitution signature 3 (1.611), rearrangement signature 3 (1.153), rearrangement signature 5 (0.847), HRD index (0.667) and base-substitution signature 8 (0.091)**)
2. I implemented the classification procedure made by author of the article “HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures” and tried to analyze the work that has been done by him and I realize that he used the tool to generated the existing file（contains features）for classification from other files and I realized that I also need to generate the data file from other files using matrix factorization.

**This is the code implantation of how he did to classify if it’s under the effect of BRCA1 or BRCA2**

*"""  
 using lasso regression train model on original dataset  
"""*import numpy as np  
import pandas as pd  
  
from sklearn.model\_selection import KFold  
from sklearn.preprocessing import StandardScaler  
  
from sklearn.linear\_model import Lasso  
  
from sklearn.metrics import accuracy\_score  
from sklearn.metrics import confusion\_matrix  
from sklearn.metrics import classification\_report  
from sklearn.metrics import roc\_auc\_score, roc\_curve  
  
import warnings  
import sklearn.exceptions  
warnings.filterwarnings("ignore", category=sklearn.exceptions.UndefinedMetricWarning)  
  
  
  
# \*\*\*\*\*\*\*\*\*config\*\*\*\*\*\*\*\*\*\*  
  
# Donor Age,ER status,Gene,isBrcaMonoallelic,isKnownGermline,isNewGermline,IsSomaticMeth,  
using\_columns = ['SV%d' % i for i in range(1, 7)]  
  
# Donor Age,ER status,Gene,isBrcaMonoallelic,isKnownGermline,isNewGermline,  
# IsSomaticMeth,'ins', 'del.mh.prop', 'del.rep.prop', 'del.none.prop', 'hrd'  
using\_columns.extend(['ins', 'del.mh.prop', 'del.rep.prop', 'del.none.prop', 'hrd'])  
  
# set the y here  
label\_column = 'Gene'  
  
sub\_columns = ['e.%d' % item for item in [1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26]]  
using\_columns.extend(sub\_columns)  
  
# 'SV5', 'SV6', 'SV2', 'e.8', 'e.2', 'e.17', 'e.1', 'e.18', 'e.13',  
# 'hrd', 'e.20', 'ins', 'e.26', 'SV4', 'SV1', 'e.5', 'e.6', 'e.3', 'SV3'  
normalize\_columns = list(set(using\_columns) - set(['del.mh.prop', 'del.rep.prop', 'del.none.prop']))  
  
path\_data = '../../data/raw/b\_dataset.csv'  
dataset = pd.read\_csv(path\_data)  
  
# MARK The data contains the processed data of the original data. Let's try this unprocessed first here  
# process the data  
  
dataset[sub\_columns] = np.log(dataset[sub\_columns] + 1)  
  
  
# standardization -- fitting in same scale  
ss = StandardScaler()  
dataset[normalize\_columns] = ss.fit\_transform(dataset[normalize\_columns])  
  
# lasso regression (560,22)  
x = dataset[using\_columns].fillna(0).to\_numpy()  
  
# Replace the label. Cases affected by BRCA1 / BRCA2 are 1, and those not affected are 0  
y = dataset[label\_column].map({'BRCA1': 1, 'BRCA2': 1}).fillna(0).to\_numpy()  
  
result = []  
# using k-fold cross validation  
kf = KFold(n\_splits=10)  
for train\_index, test\_index in kf.split(x):  
 train\_x, train\_y = x[train\_index], y[train\_index]  
 test\_x, test\_y = x[test\_index], y[test\_index]  
 lasso = Lasso(max\_iter=10000, alpha=0.9, fit\_intercept=True) # Use the default parameters for now  
 lasso.fit(train\_x, train\_y)  
 pred\_y = lasso.predict(test\_x)  
  
 pred\_y[pred\_y > 0.5] = 1  
 pred\_y[pred\_y <= 0.5] = 0  
 acc = accuracy\_score(pred\_y, test\_y)  
 cp = classification\_report(test\_y, pred\_y)  
 confusion\_mat = confusion\_matrix(test\_y, pred\_y)  
 roc\_c = roc\_curve(test\_y, pred\_y)  
 roc\_acc\_s = roc\_auc\_score(test\_y, pred\_y)  
 # *TODO store the data ，draw the graph* print(acc)  
 #print(lasso.coef\_)

1. .I did research analyze on Landscape of somatic mutations in 560 breast cancer whole- genome sequences and Deciphering signatures of mutational processes operative in human cancer to understand some of the terminology that has been used in the classification of the procedure above.

**The questions that I have:**

**Open the file**

**There are 4 features have been used for classification**

(i) counts of mutations associ-ated with each signature of single-base substitutions: signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20 and 26 (signature 30 was excluded as it involved only 1 sample),**(already obtained with xinyu’s help) (e.1…..etc)**

(ii) indels with microhomology at the indel breakpoint junction, indels

at polynucleotide-repeat tracts and other complex indels as proportions, **through this tool to generate http://cancerit.github.io/cgpPindel/ Indels (ins)**

(iii) counts of rearrangements associated with each signature of rearrangements RS1–RS6 and **obtained by using http://cancerit.github.io/CaVEMan/ somatic substitutions (sv1…..etc) which involves matrix factorization.**

(iv) HRD index. **(hrd) haven’t figured out how to obtain, any idea?**

**Question: are those the features used for our classification as they used in article?**

1. **Are we categorizing different cancers following the method of the paper?**
2. **Are we going to apply same method above for classifying every cancers?**

**The plan that I have:**

1. try to get the features from file downloaded using tools and performing classification on it