Personalized cancer diagnosis

1. Business Problem

1.1. Description

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/

Data: Memorial Sloan Kettering Cancer Center (MSKCC)

Download training_variants.zip and training_text.zip from Kaggle.

Context:

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/discussion/35336#198462

Problem statement :

Classify the given genetic variations/mutations based on evidence from text-based clinical literature.

1.2. Source/Useful Links

Some articles and reference blogs about the problem statement

1. https://www.forbes.com/sites/matthewherper/2017/06/03/a-new-cancer-drug-helped-almost-everyone-who-took-it-almost-heres-what-it-teaches-us/#2a44ee2f6b25
2. https://www.youtube.com/watch?v=UwbuW7oK8rk
3. https://www.youtube.com/watch?v=qxXRKVompl8

1.3. Real-world/Business objectives and constraints.

- No low-latency requirement.
- Interpretability is important.
- Errors can be very costly.
- Probability of a data-point belonging to each class is needed.

2. Machine Learning Problem Formulation

2.1. Data

2.1.1. Data Overview

- Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment/data
- We have two data files: one conatins the information about the genetic mutations and the other contains the clinical evidence (text) that human experts/pathologists use to classify the genetic mutations.
- Both these data files are have a common column called ID
- Data file's information:
  - training_variants (ID , Gene, Variations, Class)
  - training_text (ID, Text)

2.1.2. Example Data Point

training_variants

ID,Gene,Variation,Class
0,FAM58A,Truncating Mutations,1
1,CBL,W802*,2
2,CBL,Q249E,2
...

training_text

ID,Text
0||Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin. Cyclin M, an orphan cyclin, is the product of FAM58A, whose mutations cause STAR syndrome, a human developmental anomaly whose features include toe syndactyly, telecanthus, and anogenital and renal malformations. We show that STAR syndrome-associated cyclin M mutants are unable to interact with CDK10. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and in conferring tamoxifen resistance to breast cancer cells. CDK10/cyclin M phosphorylates ETS2 in vitro, and in cells it positively controls ETS2 degradation by the proteasome. ETS2 protein levels are increased in cells derived from a STAR patient, and this increase is attributable to decreased cyclin M levels. Altogether, our results reveal an additional regulatory mechanism for ETS2, which plays key roles in cancer and development. They also shed light on the molecular mechanisms underlying STAR syndrome.Cyclin-dependent kinases (CDKs) play a pivotal role in the control of a number of fundamental cellular processes (1). The human genome contains 21 genes encoding proteins that can be considered as members of the CDK family owing to their sequence similarity with bona fide CDKs, those known to be activated by cyclins (2). Although discovered almost 20 y ago (3, 4), CDK10 remains one of the two CDKs without an identified cyclin partner. This knowledge gap has largely impeded the exploration of its biological functions. CDK10 can act as a positive cell cycle regulator in some cells (5, 6) or as a tumor suppressor in others (7, 8). CDK10 interacts with the ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2) transcription factor and inhibits its transcriptional activity through an unknown mechanism (9). CDK10 knockdown derepresses ETS2, which increases the expression of the c-Raf protein kinase, activates the MAPK pathway, and induces resistance of MCF7 cells to tamoxifen (6). ...

2.2. Mapping the real-world problem to an ML problem

2.2.1. Type of Machine Learning Problem

There are nine different classes a genetic mutation can be classified into => Multi class classification problem

2.2.2. Performance Metric

Source: https://www.kaggle.com/c/msk-redefining-cancer-treatment#evaluation

Metric(s):

- Multi class log-loss
- Confusion matrix

2.2.3. Machine Learing Objectives and Constraints

Objective: Predict the probability of each data-point belonging to each of the nine classes.

Constraints:

- Interpretability
- Class probabilities are needed.
- Penalize the errors in class probabilites => Metric is Log-loss.

- No Latency constraints.

2.3. Train, CV and Test Datasets

Split the dataset randomly into three parts train, cross validation and test with 64%,16%, 20% of data respectively

3. Exploratory Data Analysis

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import re
import time
import warnings
import numpy as np
from nltk.corpus import stopwords
# from sklearn.decomposition import TruncatedSVD
from sklearn.preprocessing import normalize
from sklearn.feature_extraction.text import CountVectorizer
# from sklearn.manifold import TSNE
import seaborn as sns
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics.classification import accuracy_score, log_loss
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import SGDClassifier
# from imblearn.over_sampling import SMOTE
from collections import Counter
from scipy.sparse import hstack
# from sklearn.multiclass import OneVsRestClassifier
from sklearn.svm import SVC
# from sklearn.cross_validation import StratifiedKFold
from collections import Counter, defaultdict
from sklearn.calibration import CalibratedClassifierCV
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
import math
# from sklearn.metrics import normalized_mutual_info_score
from sklearn.ensemble import RandomForestClassifier
warnings.filterwarnings("ignore")

from mlxtend.classifier import StackingClassifier

from sklearn import model_selection
from sklearn.linear_model import LogisticRegression

from tqdm import tqdm_notebook
```

3.1. Reading Data

3.1.1. Reading Gene and Variation Data

In [2]:

```
data = pd.read_csv('training_variants')
print('Number of data points : ', data.shape[0])
print('Number of features : ', data.shape[1])
print('Features : ', data.columns.values)
data.head()
```

```
Number of data points :  3321
Number of features :  4
```

```
Features :  ['ID' 'Gene' 'Variation' 'Class']
```

Out[2]:

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2
3	3	CBL	N454D	3
4	4	CBL	L399V	4

training/training_variants is a comma separated file containing the description of the genetic mutations used for training.
Fields are

- **ID :** the id of the row used to link the mutation to the clinical evidence
- **Gene :** the gene where this genetic mutation is located
- **Variation :** the aminoacid change for this mutations
- **Class :** 1-9 the class this genetic mutation has been classified on

3.1.2. Reading Text Data

In [3]:

```
# note the seprator in this file
data_text =pd.read_csv("training_text",sep="\|\|",engine="python",names=["ID","TEXT"],skiprows=1)
print('Number of data points : ', data_text.shape[0])
print('Number of features : ', data_text.shape[1])
print('Features : ', data_text.columns.values)
data_text.head()
```

```
Number of data points :  3321
Number of features :  2
Features :  ['ID' 'TEXT']
```

Out[3]:

	ID	TEXT
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

3.1.3. Preprocessing of text

In [4]:

```
# loading stop words from nltk library
stop_words = set(stopwords.words('english'))

def nlp_preprocessing(total_text, index, column):
    if type(total_text) is not int:
        string = ""

        # replace every special char with space
        total_text = re.sub('[^a-zA-Z0-9\n]', ' ', total_text)

        # replace multiple spaces with single space
```

```
        total_text = re.sub('\s+',' ', total_text)

        # converting all the chars into lower-case.
        total_text = total_text.lower()

        for word in total_text.split():
        # if the word is a not a stop word then retain that word from the data
            if not word in stop_words:
                string += word + " "

        data_text[column][index] = string
```

In [5]:

```
#text processing stage.
start_time = time.clock()
for index, row in tqdm_notebook(data_text.iterrows()):
    if type(row['TEXT']) is str:
        nlp_preprocessing(row['TEXT'], index, 'TEXT')
    else:
        print("there is no text description for id:",index)
print('Time took for preprocessing the text :',time.clock() - start_time, "seconds")
```

```
there is no text description for id: 1109
there is no text description for id: 1277
there is no text description for id: 1407
there is no text description for id: 1639
there is no text description for id: 2755

Time took for preprocessing the text : 154.3329649 seconds
```

In [6]:

```
#merging both gene_variations and text data based on ID
result = pd.merge(data, data_text,on='ID', how='left')
result.head()
```

Out[6]:

	ID	Gene	Variation	Class	TEXT
0	0	FAM58A	Truncating Mutations	1	cyclin dependent kinases cdks regulate variety...
1	1	CBL	W802*	2	abstract background non small cell lung cancer...
2	2	CBL	Q249E	2	abstract background non small cell lung cancer...
3	3	CBL	N454D	3	recent evidence demonstrated acquired uniparen...
4	4	CBL	L399V	4	oncogenic mutations monomeric casitas b lineag...

In [7]:

```
result[result.isnull().any(axis=1)]
```

Out[7]:

	ID	Gene	Variation	Class	TEXT
1109	1109	FANCA	S1088F	1	NaN
1277	1277	ARID5B	Truncating Mutations	1	NaN
1407	1407	FGFR3	K508M	6	NaN
1639	1639	FLT1	Amplification	6	NaN
2755	2755	BRAF	G596C	7	NaN

In [8]:

```
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result['Variation']
```

In [9]:

```
result[result['ID']==1109]
```

Out[9]:

	ID	Gene	Variation	Class	TEXT
1109	1109	FANCA	S1088F	1	FANCA S1088F

3.1.4. Test, Train and Cross Validation Split

3.1.4.1. Splitting data into train, test and cross validation (64:20:16)

In [10]:

```
y_true = result['Class'].values
result.Gene = result.Gene.str.replace('\s+', '_')
result.Variation = result.Variation.str.replace('\s+', '_')

# split the data into test and train by maintaining same distribution of output varaible 'y_true' [stra
tify=y_true]
X_train, test_df, y_train, y_test = train_test_split(result, y_true, stratify=y_true, test_size=0.2, ra
ndom_state=1)

# split the train data into train and cross validation by maintaining same distribution of output varai
ble 'y_train' [stratify=y_train]
train_df, cv_df, y_train, y_cv = train_test_split(X_train, y_train, stratify=y_train, test_size=0.2, ra
ndom_state=1)
```

We split the data into train, test and cross validation data sets, preserving the ratio of class distribution in the original data set

In [11]:

```
print('Number of data points in train data:', train_df.shape[0])
print('Number of data points in test data:', test_df.shape[0])
print('Number of data points in cross validation data:', cv_df.shape[0])
```

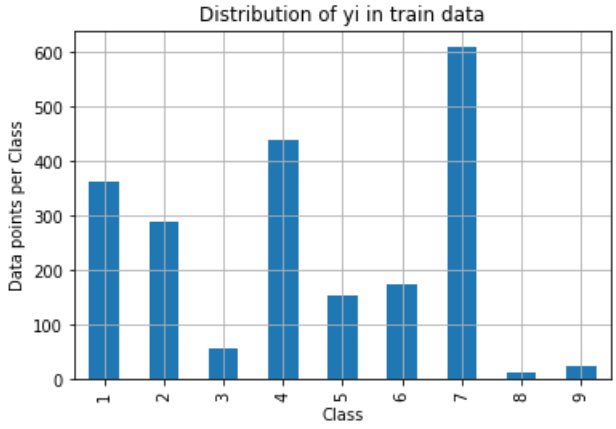
```
Number of data points in train data: 2124
Number of data points in test data: 665
Number of data points in cross validation data: 532
```

3.1.4.2. Distribution of y_i's in Train, Test and Cross Validation datasets

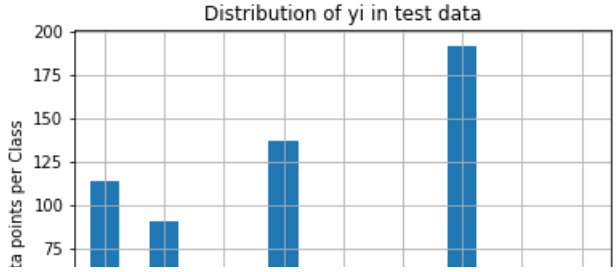
In [12]:

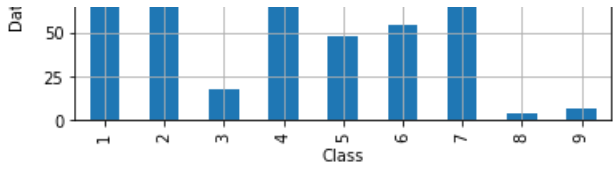
```
# it returns a dict, keys as class labels and values as the number of data points in that class
train_class_distribution = train_df['Class'].value_counts().sort_index()
test_class_distribution = test_df['Class'].value_counts().sort_index()
cv_class_distribution = cv_df['Class'].value_counts().sort_index()

my_colors = 'rgbkymc'
train_class_distribution.plot(kind='bar')
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in train data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-train_class_distribution.values)
for i in sorted_yi:
```

```
    print('Number of data points in class', i+1, ':',train_class_distribution.values[i], '(', np.round(
(train_class_distribution.values[i]/train_df.shape[0]*100), 3), '%)')


print('-'*80)
my_colors = 'rgbkymc'
test_class_distribution.plot(kind='bar')
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in test data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-test_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',test_class_distribution.values[i], '(', np.round((
test_class_distribution.values[i]/test_df.shape[0]*100), 3), '%)')

print('-'*80)
my_colors = 'rgbkymc'
cv_class_distribution.plot(kind='bar')
plt.xlabel('Class')
plt.ylabel('Data points per Class')
plt.title('Distribution of yi in cross validation data')
plt.grid()
plt.show()

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html
# -(train_class_distribution.values): the minus sign will give us in decreasing order
sorted_yi = np.argsort(-train_class_distribution.values)
for i in sorted_yi:
    print('Number of data points in class', i+1, ':',cv_class_distribution.values[i], '(', np.round((cv
_class_distribution.values[i]/cv_df.shape[0]*100), 3), '%)')
```

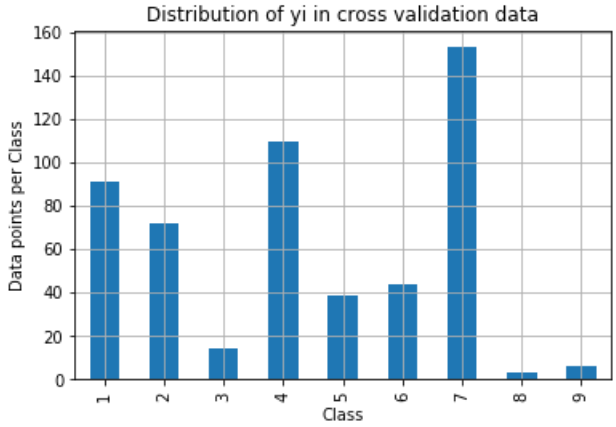



```
Number of data points in class 7 : 609 ( 28.672 %)
Number of data points in class 4 : 439 ( 20.669 %)
Number of data points in class 1 : 363 ( 17.09 %)
Number of data points in class 2 : 289 ( 13.606 %)
Number of data points in class 6 : 176 ( 8.286 %)
Number of data points in class 5 : 155 ( 7.298 %)
Number of data points in class 3 : 57 ( 2.684 %)
Number of data points in class 9 : 24 ( 1.13 %)
Number of data points in class 8 : 12 ( 0.565 %)
--------------------------------------------------------------------------------
```






```
Number of data points in class 7 : 191 ( 28.722 %)
Number of data points in class 4 : 137 ( 20.602 %)
Number of data points in class 1 : 114 ( 17.143 %)
Number of data points in class 2 : 91 ( 13.684 %)
Number of data points in class 6 : 55 ( 8.271 %)
Number of data points in class 5 : 48 ( 7.218 %)
Number of data points in class 3 : 18 ( 2.707 %)
Number of data points in class 9 : 7 ( 1.053 %)
Number of data points in class 8 : 4 ( 0.602 %)
------------------------------------------------------------------------------------------
```




```
Number of data points in class 7 : 153 ( 28.759 %)
Number of data points in class 4 : 110 ( 20.677 %)
Number of data points in class 1 : 91 ( 17.105 %)
Number of data points in class 2 : 72 ( 13.534 %)
Number of data points in class 6 : 44 ( 8.271 %)
Number of data points in class 5 : 39 ( 7.331 %)
Number of data points in class 3 : 14 ( 2.632 %)
Number of data points in class 9 : 6 ( 1.128 %)
Number of data points in class 8 : 3 ( 0.564 %)
```

3.2 Prediction using a 'Random' Model

In a 'Random' Model, we generate the NINE class probabilites randomly such that they sum to 1.

In [13]:

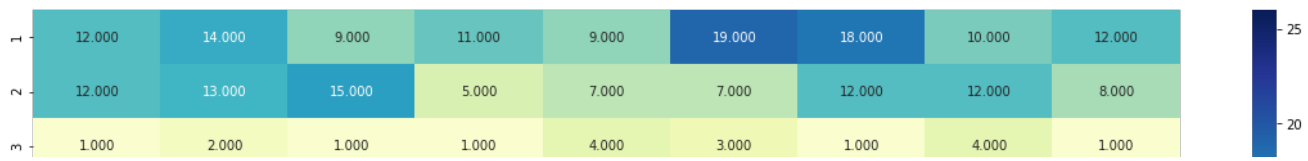
```
# This function plots the confusion matrices given y_i, y_i_hat.
def plot_confusion_matrix(test_y, predict_y):
    C = confusion_matrix(test_y, predict_y)
    # C = 9,9 matrix, each cell (i,j) represents number of points of class i are predicted class j

    A =(((C.T)/(C.sum(axis=1))).T)
    #divid each element of the confusion matrix with the sum of elements in that column

    # C = [[1, 2],
    #      [3, 4]]
    # C.T = [[1, 3],
    #        [2, 4]]
    # C.sum(axis = 1)  axis=0 corresonds to columns and axis=1 corresponds to rows in two diamensional
array
    # C.sum(axix =1) = [[3, 7]]
    # ((C.T)/(C.sum(axis=1))) = [[1/3, 3/7]
    #                            [2/3, 4/7]]

    # ((C.T)/(C.sum(axis=1))).T = [[1/3, 2/3]
```

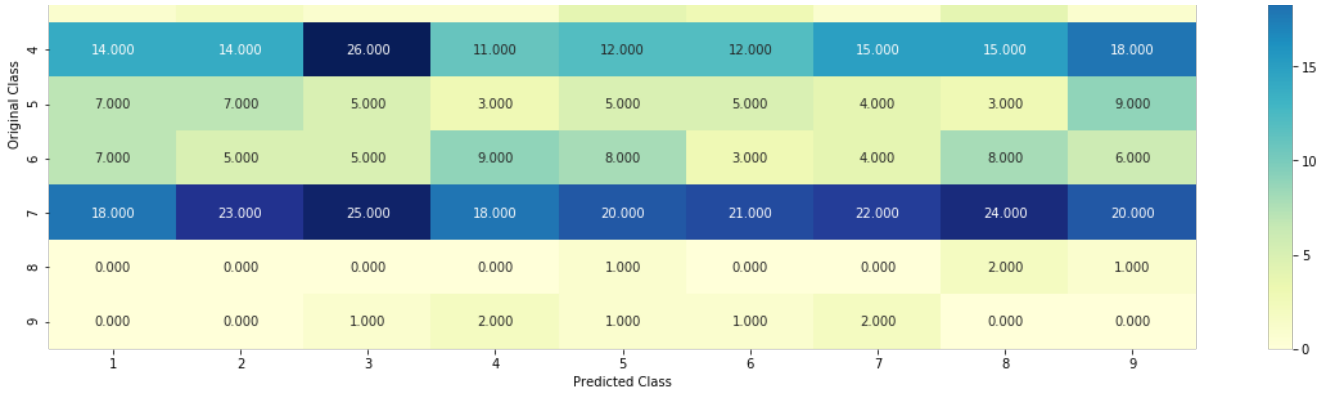
```
    #                              [3/7, 4/7]]
    # sum of row elements = 1

    B =(C/C.sum(axis=0))
    #divid each element of the confusion matrix with the sum of elements in that row
    # C = [[1, 2],
    #     [3, 4]]
    # C.sum(axis = 0)  axis=0 corresonds to columns and axis=1 corresponds to rows in two diamensional
array
    # C.sum(axix =0) = [[4, 6]]
    # (C/C.sum(axis=0)) = [[1/4, 2/6],
    #                      [3/4, 4/6]]

    labels = [1,2,3,4,5,6,7,8,9]
    # representing A in heatmap format
    print("-"*20, "Confusion matrix", "-"*20)
    plt.figure(figsize=(20,7))
    sns.heatmap(C, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

    print("-"*20, "Precision matrix (Columm Sum=1)", "-"*20)
    plt.figure(figsize=(20,7))
    sns.heatmap(B, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

    # representing B in heatmap format
    print("-"*20, "Recall matrix (Row sum=1)", "-"*20)
    plt.figure(figsize=(20,7))
    sns.heatmap(A, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()
```

In [14]:

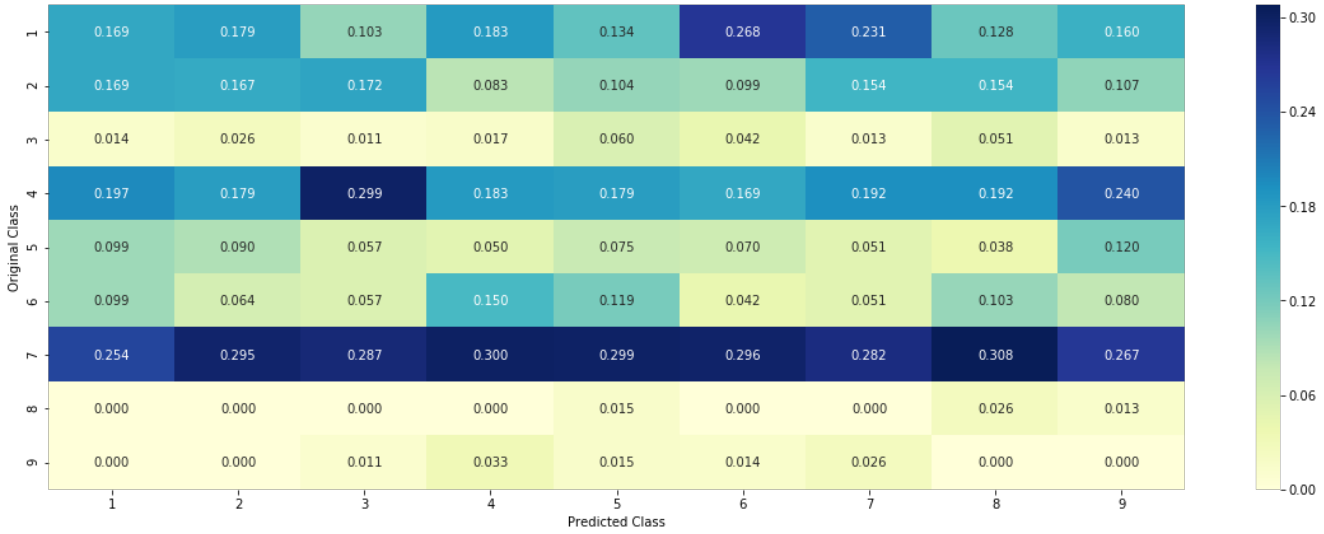
```
# we need to generate 9 numbers and the sum of numbers should be 1
# one solution is to genarate 9 numbers and divide each of the numbers by their sum
# ref: https://stackoverflow.com/a/18662466/4084039
test_data_len = test_df.shape[0]
cv_data_len = cv_df.shape[0]

# we create a output array that has exactly same size as the CV data
cv_predicted_y = np.zeros((cv_data_len,9))
for i in range(cv_data_len):
    rand_probs = np.random.rand(1,9)
    cv_predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
print("Log loss on Cross Validation Data using Random Model",log_loss(y_cv,cv_predicted_y, eps=1e-15))


# Test-Set error.
#we create a output array that has exactly same as the test data
test_predicted_y = np.zeros((test_data_len,9))
for i in range(test_data_len):
    rand_probs = np.random.rand(1,9)
    test_predicted_y[i] = ((rand_probs/sum(sum(rand_probs)))[0])
print("Log loss on Test Data using Random Model",log_loss(y_test,test_predicted_y, eps=1e-15))

predicted_y =np.argmax(test_predicted_y, axis=1)
plot_confusion_matrix(y_test, predicted_y+1)
```

```
Log loss on Cross Validation Data using Random Model 2.5104111490430268
Log loss on Test Data using Random Model 2.5158449593068393
-------------------- Confusion matrix --------------------
```

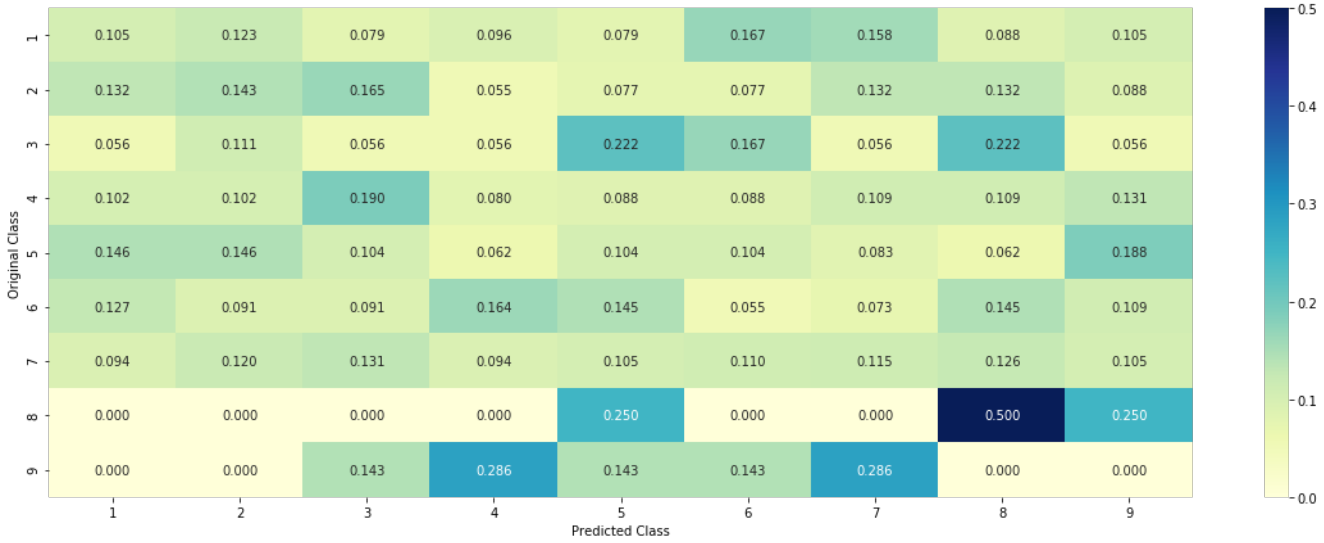




---------- Precision matrix (Columm Sum=1) ----------



---------- Recall matrix (Row sum=1) ----------



3.3 Univariate Analysis

In [15]:

```
# code for response coding with Laplace smoothing.
# alpha : used for laplace smoothing
# feature: ['gene', 'variation']
# df: ['train_df', 'test_df', 'cv_df']
# algorithm
```

```
# ----------
# Consider all unique values and the number of occurances of given feature in train data dataframe
# build a vector (1*9) , the first element = (number of times it occured in class1 + 10*alpha / number
of time it occurred in total data+90*alpha)
# gv_dict is like a look up table, for every gene it store a (1*9) representation of it
# for a value of feature in df:
# if it is in train data:
# we add the vector that was stored in 'gv_dict' look up table to 'gv_fea'
# if it is not there is train:
# we add [1/9, 1/9, 1/9, 1/9,1/9, 1/9, 1/9, 1/9, 1/9] to 'gv_fea'
# return 'gv_fea'
# ----------------------

# get_gv_fea_dict: Get Gene varaition Feature Dict
def get_gv_fea_dict(alpha, feature, df):
    # value_count: it contains a dict like
    # print(train_df['Gene'].value_counts())
    # output:
    #        {BRCA1      174
    #         TP53       106
    #         EGFR        86
    #         BRCA2       75
    #         PTEN        69
    #         KIT         61
    #         BRAF        60
    #         ERBB2       47
    #         PDGFRA      46
    #         ...}
    # print(train_df['Variation'].value_counts())
    # output:
    # {
    # Truncating_Mutations                     63
    # Deletion                                 43
    # Amplification                            43
    # Fusions                                  22
    # Overexpression                            3
    # E17K                                      3
    # Q61L                                      3
    # S222D                                     2
    # P130S                                     2
    # ...
    # }
    value_count = train_df[feature].value_counts()

    # gv_dict : Gene Variation Dict, which contains the probability array for each gene/variation
    gv_dict = dict()

    # denominator will contain the number of time that particular feature occured in whole data
    for i, denominator in value_count.items():
        # vec will contain (p(yi==1/Gi) probability of gene/variation belongs to perticular class
        # vec is 9 diamensional vector
        vec = []
        for k in range(1,10):
            # print(train_df.loc[(train_df['Class']==1) & (train_df['Gene']=='BRCA1')])
            #         ID   Gene             Variation  Class
            # 2470  2470  BRCA1               S1715C      1
            # 2486  2486  BRCA1               S1841R      1
            # 2614  2614  BRCA1                  M1R      1
            # 2432  2432  BRCA1               L1657P      1
            # 2567  2567  BRCA1               T1685A      1
            # 2583  2583  BRCA1               E1660G      1
            # 2634  2634  BRCA1               W1718L      1
            # cls_cnt.shape[0] will return the number of rows

            cls_cnt = train_df.loc[(train_df['Class']==k) & (train_df[feature]==i)]

            # cls_cnt.shape[0](numerator) will contain the number of time that particular feature occur
ed in whole data
            vec.append((cls_cnt.shape[0] + alpha*10)/ (denominator + 90*alpha))

        # we are adding the gene/variation to the dict as key and vec as value
        gv_dict[i]=vec
    return gv_dict

# Get Gene variation feature
def get_gv_feature(alpha, feature, df):
    # print(gv_dict)
```

```
    # print(gv_dict)
    #     {'BRCA1': [0.20075757575757575, 0.03787878787878788, 0.068181818181818177, 0.1363636363636363
5, 0.25, 0.19318181818181818, 0.03787878787878788, 0.03787878787878788, 0.03787878787878788],
    #      'TP53': [0.32142857142857145, 0.061224489795918366, 0.061224489795918366, 0.2704081632653061
5, 0.061224489795918366, 0.066326530612244902, 0.051020408163265307, 0.051020408163265307, 0.0561224489
79591837],
    #      'EGFR': [0.056818181818181816, 0.21590909090909091, 0.0625, 0.068181818181818177, 0.06818181
8181818177, 0.0625, 0.34659090909090912, 0.0625, 0.056818181818181816],
    #      'BRCA2': [0.13333333333333333, 0.060606060606060608, 0.060606060606060608, 0.078787878787878
782, 0.1393939393939394, 0.34545454545454546, 0.060606060606060608, 0.060606060606060608, 0.06060606060
6060608],
    #      'PTEN': [0.069182389937106917, 0.062893081761006289, 0.069182389937106917, 0.465408805031446
55, 0.075471698113207544, 0.062893081761006289, 0.069182389937106917, 0.062893081761006289, 0.062893081
761006289],
    #      'KIT': [0.066225165562913912, 0.25165562913907286, 0.072847682119205295, 0.07284768211920529
5, 0.066225165562913912, 0.066225165562913912, 0.27152317880794702, 0.066225165562913912, 0.06622516556
2913912],
    #      'BRAF': [0.066666666666666666, 0.17999999999999999, 0.073333333333333334, 0.0733333333333333
34, 0.093333333333333338, 0.080000000000000002, 0.29999999999999999, 0.066666666666666666, 0.0666666666
66666666],
    #      ...
    #     }
    gv_dict = get_gv_fea_dict(alpha, feature, df)
    # value_count is similar in get_gv_fea_dict
    value_count = train_df[feature].value_counts()

    # gv_fea: Gene_variation feature, it will contain the feature for each feature value in the data
    gv_fea = []
    # for every feature values in the given data frame we will check if it is there in the train data t
hen we will add the feature to gv_fea
    # if not we will add [1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9] to gv_fea
    for index, row in df.iterrows():
        if row[feature] in dict(value_count).keys():
            gv_fea.append(gv_dict[row[feature]])
        else:
            gv_fea.append([1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9])
#             gv_fea.append([-1,-1,-1,-1,-1,-1,-1,-1,-1])
    return gv_fea
```

when we caculate the probability of a feature belongs to any particular class, we apply laplace smoothing

- (numerator + 10*alpha) / (denominator + 90*alpha)

3.2.1 Univariate Analysis on Gene Feature

Q1. Gene, What type of feature it is ?

Ans. Gene is a categorical variable

Q2. How many categories are there and How they are distributed?

In [16]:

```
unique_genes = train_df['Gene'].value_counts()
print('Number of Unique Genes :', unique_genes.shape[0])
# the top 10 genes that occured most
print(unique_genes.head(10))
```

```
Number of Unique Genes : 235
BRCA1     168
TP53      111
PTEN       86
EGFR       86
BRCA2      80
BRAF       65
KIT        60
ALK        44
ERBB2      39
PDGFRA     38
Name: Gene, dtype: int64
```

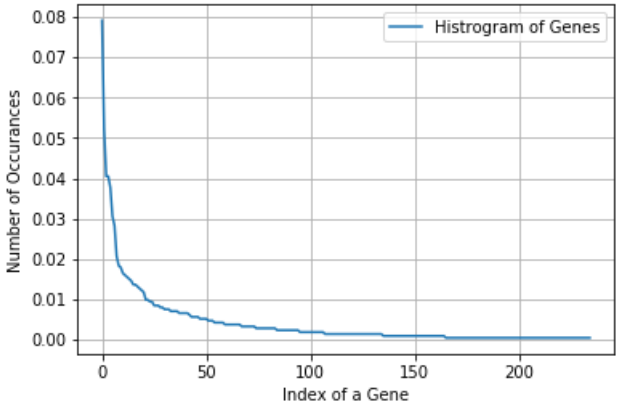
In [17]:

```
print("Ans: There are", unique_genes.shape[0] ,"different categories of genes in the train data, and th
ey are distibuted as follows",)
```

```
Ans: There are 235 different categories of genes in the train data, and they are distibuted as follows
```

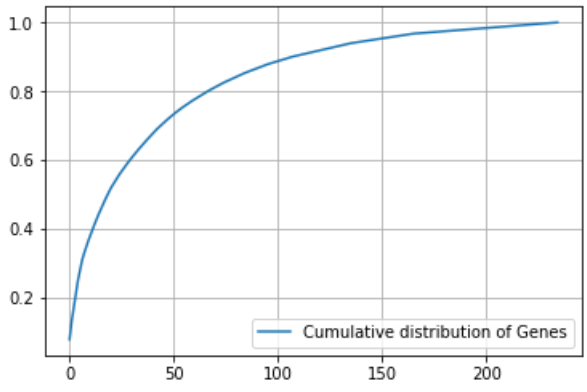
In [18]:

```
s = sum(unique_genes.values);
h = unique_genes.values/s;
plt.plot(h, label="Histrogram of Genes")
plt.xlabel('Index of a Gene')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```



In [19]:

```
c = np.cumsum(h)
plt.plot(c,label='Cumulative distribution of Genes')
plt.grid()
plt.legend()
plt.show()
```



Q3. How to featurize this Gene feature ?

Ans.there are two ways we can featurize this variable check out this video:
https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/

1. One hot Encoding
2. Response coding

We will choose the appropriate featurization based on the ML model we use. For this problem of multi-class classification with categorical features, one-hot encoding is better for Logistic regression while response coding is better for Random Forests.

In [20]:

```
# Response-coding of the Gene feature

# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", train_df))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", test_df))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", cv_df))
```

In [21]:

```
print("train_gene_feature_responseCoding is converted feature using respone coding method. The shape of
gene feature:", train_gene_feature_responseCoding.shape)
```

```
train_gene_feature_responseCoding is converted feature using respone coding method. The shape of gene f
eature: (2124, 9)
```

In [22]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = CountVectorizer()
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(train_df['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(test_df['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(cv_df['Gene'])
```

In [23]:

```
gene_vectorizer.get_feature_names()
```

Out[23]:

```
['abl1',
 'acvr1',
 'ago2',
 'akt1',
 'akt2',
 'akt3',
 'alk',
 'apc',
 'ar',
 'araf',
 'arid1a',
 'arid1b',
 'arid2',
 'arid5b',
 'asxl1',
 'asxl2',
 'atm',
 'atr',
 'atrx',
 'aurka',
 'axin1',
 'axl',
 'b2m',
 'bap1',
 'bard1',
 'bcl10',
 'bcl2l11',
 'bcor',
 'braf',
 'brca1',
 'brca2',
 'brd4',
 'brip1',
 'btk',
 'card11',
 'carm1',
```

```
'casp8',
'cbl',
'ccnd1',
'ccnd2',
'ccnd3',
'ccne1',
'cdh1',
'cdk12',
'cdk4',
'cdk6',
'cdkn1a',
'cdkn1b',
'cdkn2a',
'cdkn2b',
'cdkn2c',
'cebpa',
'chek2',
'cic',
'crebbp',
'ctcf',
'ctla4',
'ctnnb1',
'ddr2',
'dicer1',
'dnmt3a',
'dnmt3b',
'egfr',
'elf3',
'ep300',
'epas1',
'epcam',
'erbb2',
'erbb3',
'erbb4',
'ercc2',
'ercc3',
'ercc4',
'erg',
'errfi1',
'esr1',
'etv1',
'etv6',
'ewsr1',
'ezh2',
'fanca',
'fat1',
'fbxw7',
'fgf4',
'fgfr1',
'fgfr2',
'fgfr3',
'fgfr4',
'flt1',
'flt3',
'foxa1',
'foxl2',
'foxp1',
'fubp1',
'gata3',
'gli1',
'gnaq',
'gnas',
'h3f3a',
'hist1h1c',
'hla',
'hras',
'idh1',
'idh2',
'igf1r',
'ikbke',
'ikzf1',
'jak1',
'jak2',
'jun',
'kdm5c',
'kdm6a',
'kdr',
```

```
'kdr',
'keap1',
'kit',
'kmt2a',
'kmt2b',
'kmt2c',
'kmt2d',
'knstrn',
'kras',
'lats2',
'map2k1',
'map2k2',
'map2k4',
'map3k1',
'mapk1',
'mdm2',
'mdm4',
'med12',
'mef2b',
'men1',
'met',
'mlh1',
'mpl',
'msh2',
'msh6',
'mtor',
'myc',
'mycn',
'myd88',
'ncor1',
'nf1',
'nf2',
'nfe2l2',
'nfkbia',
'nkx2',
'notch1',
'notch2',
'nras',
'ntrk1',
'ntrk2',
'ntrk3',
'nup93',
'pak1',
'pax8',
'pbrm1',
'pdgfra',
'pdgfrb',
'pik3ca',
'pik3cb',
'pik3cd',
'pik3r1',
'pik3r2',
'pik3r3',
'pim1',
'pms1',
'pms2',
'pole',
'ppm1d',
'ppp2r1a',
'ppp6c',
'prdm1',
'ptch1',
'pten',
'ptpn11',
'ptprd',
'ptprt',
'rab35',
'rac1',
'rad21',
'rad50',
'rad51b',
'rad51d',
'raf1',
'rara',
'rasa1',
'rb1',
'rbm10',
'ret',
```

```
 'rheb',
 'rhoa',
 'rictor',
 'rit1',
 'rnf43',
 'ros1',
 'rras2',
 'runx1',
 'rybp',
 'sdhb',
 'sdhc',
 'setd2',
 'sf3b1',
 'shoc2',
 'smad2',
 'smad3',
 'smad4',
 'smarca4',
 'smarcb1',
 'smo',
 'sox9',
 'spop',
 'src',
 'stag2',
 'stat3',
 'stk11',
 'tcf7l2',
 'tert',
 'tet1',
 'tet2',
 'tgfbr1',
 'tgfbr2',
 'tmprss2',
 'tp53',
 'tp53bp1',
 'tsc1',
 'tsc2',
 'u2af1',
 'vegfa',
 'vhl',
 'whsc1',
 'whsc1l1',
 'xpo1',
 'xrcc2',
 'yap1']
```

In [24]:

```
print("train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of gene feature:", train_gene_feature_onehotCoding.shape)
```

```
train_gene_feature_onehotCoding is converted feature using one-hot encoding method. The shape of gene feature: (2124, 235)
```

Q4. How good is this gene feature in predicting y_i?

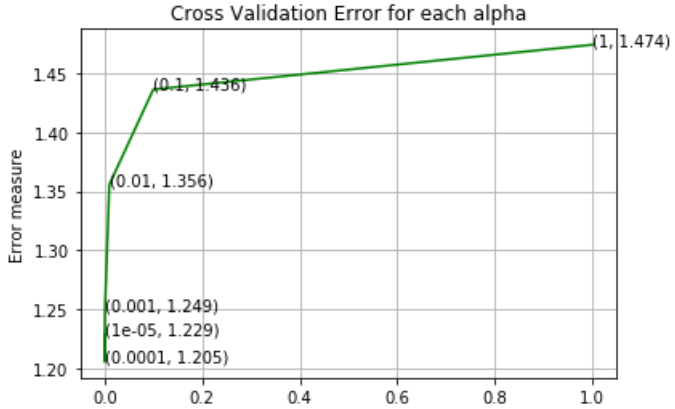
There are many ways to estimate how good a feature is, in predicting y_i. One of the good methods is to build a proper ML model using just this feature. In this case, we will build a logistic regression model using only Gene feature (one hot encoded) to predict y_i.

In [25]:

```
alpha = [10 ** x for x in range(-5, 1)] # hyperparam for SGD classifier.

# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0,
```

```
power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, …]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_gene_feature_onehotCoding, y_train)

    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_gene_feature_onehotCoding, y_train)

    predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_
, eps=1e-15))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)

clf.fit(train_gene_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_gene_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, pred
ict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_
cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_gene_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predic
t_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.2285570397154246
For values of alpha =  0.0001 The log loss is: 1.205397678433504
For values of alpha =  0.001 The log loss is: 1.2491650009150361
For values of alpha =  0.01 The log loss is: 1.3556362010078462
For values of alpha =  0.1 The log loss is: 1.4361429435939435
For values of alpha =  1 The log loss is: 1.4740301432719325
```



Alpha i's

```
For values of best alpha =  0.0001 The train log loss is: 1.004691587658098
For values of best alpha =  0.0001 The cross validation log loss is: 1.205397678433504
For values of best alpha =  0.0001 The test log loss is: 1.1962093819056512
```

Q5. Is the Gene feature stable across all the data sets (Test, Train, Cross validation)?

Ans. Yes, it is. Otherwise, the CV and Test errors would be significantly more than train error.

In [26]:

```
print("Q6. How many data points in Test and CV datasets are covered by the ", unique_genes.shape[0], "
genes in train dataset?")

test_coverage=test_df[test_df['Gene'].isin(list(set(train_df['Gene'])))].shape[0]
cv_coverage=cv_df[cv_df['Gene'].isin(list(set(train_df['Gene'])))].shape[0]

print('Ans\n1. In test data',test_coverage, 'out of',test_df.shape[0], ":",(test_coverage/test_df.shape
[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',cv_df.shape[0],":" ,(cv_coverage/cv_df.shape
[0])*100)
```

```
Q6. How many data points in Test and CV datasets are covered by the  235  genes in train dataset?
Ans
1. In test data 644 out of 665 : 96.84210526315789
2. In cross validation data 511 out of  532 : 96.05263157894737
```

3.2.2 Univariate Analysis on Variation Feature

Q7. Variation, What type of feature is it ?

Ans. Variation is a categorical variable

Q8. How many categories are there?

In [27]:

```
unique_variations = train_df['Variation'].value_counts()
print('Number of Unique Variations :', unique_variations.shape[0])
# the top 10 variations that occured most
print(unique_variations.head(10))
```

```
Number of Unique Variations : 1920
Truncating_Mutations    58
Amplification           51
Deletion                45
Fusions                 25
G12V                     4
Overexpression           3
Y64A                     2
C618R                    2
Q61R                     2
Q61L                     2
Name: Variation, dtype: int64
```

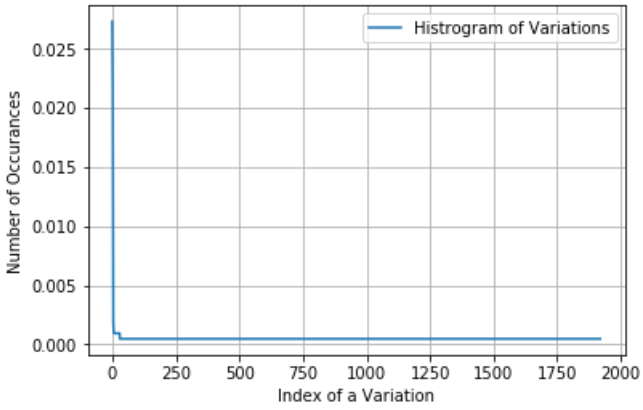
In [28]:

```
print("Ans: There are", unique_variations.shape[0] ,"different categories of variations in the train da
ta, and they are distibuted as follows",)
```

```
Ans: There are 1920 different categories of variations in the train data, and they are distibuted as fo
llows
```

In [29]:

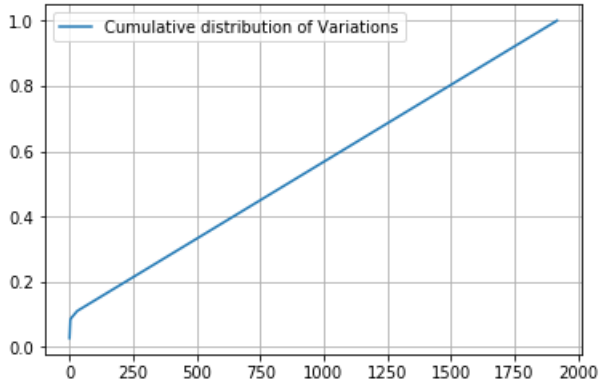
```
s = sum(unique_variations.values);
h = unique_variations.values/s;
plt.plot(h, label="Histrogram of Variations")
plt.xlabel('Index of a Variation')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```



In [30]:

```
c = np.cumsum(h)
print(c)
plt.plot(c,label='Cumulative distribution of Variations')
plt.grid()
plt.legend()
plt.show()
```

```
[0.02730697 0.05131827 0.07250471 ... 0.99905838 0.99952919 1.        ]
```



Q9. How to featurize this Variation feature ?

Ans. There are two ways we can featurize this variable check out this video:
https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/handling-categorical-and-numerical-features/

1. One hot Encoding
2. Response coding

We will be using both these methods to featurize the Variation Feature

In [31]:

```
# Response coding on 'Variation'

# alpha is used for laplace smoothing
alpha = 1
# train gene feature
```

```
# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", train_df))
# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", test_df))
# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", cv_df))
```

In [32]:

```
print("train_variation_feature_responseCoding is a converted feature using the response coding method. The shape of Variation feature:", train_variation_feature_responseCoding.shape)
```

```
train_variation_feature_responseCoding is a converted feature using the response coding method. The shape of Variation feature: (2124, 9)
```

In [33]:

```
# one-hot encoding of variation feature.
variation_vectorizer = CountVectorizer()
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(train_df['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(test_df['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(cv_df['Variation'])
```

In [34]:

```
print("train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The shape of Variation feature:", train_variation_feature_onehotCoding.shape)
```

```
train_variation_feature_onehotEncoded is converted feature using the onne-hot encoding method. The shape of Variation feature: (2124, 1948)
```

Q10. How good is this Variation feature in predicting y_i?

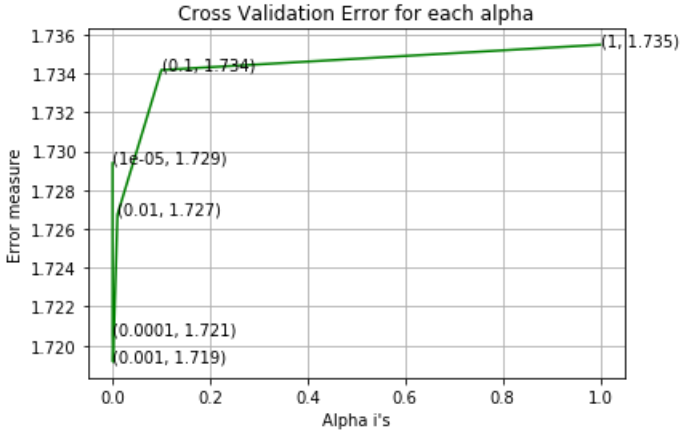
Let's build a model just like the earlier!

In [35]:

```
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_variation_feature_onehotCoding, y_train)

    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_variation_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)

    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.classes
```

```
, eps=1e-15))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_variation_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_variation_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, pred
ict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_
cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_variation_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predic
t_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.7293899332490221
For values of alpha =  0.0001 The log loss is: 1.7205800598416228
For values of alpha =  0.001 The log loss is: 1.719154795433691
For values of alpha =  0.01 The log loss is: 1.7267443626734562
For values of alpha =  0.1 The log loss is: 1.7341851235193924
For values of alpha =  1 The log loss is: 1.7354863846993358
```




```
For values of best alpha =  0.001 The train log loss is: 1.0984554907261728
For values of best alpha =  0.001 The cross validation log loss is: 1.719154795433691
For values of best alpha =  0.001 The test log loss is: 1.7066671376163376
```

Q11. Is the Variation feature stable across all the data sets (Test, Train, Cross validation)?

Ans. Not sure! But lets be very sure using the below analysis.

In [36]:

```
print("Q12. How many data points are covered by total ", unique_variations.shape[0], " genes in test an
d cross validation data sets?")
test_coverage=test_df[test_df['Variation'].isin(list(set(train_df['Variation'])))].shape[0]
cv_coverage=cv_df[cv_df['Variation'].isin(list(set(train_df['Variation'])))].shape[0]
print('Ans\n1. In test data',test_coverage, 'out of',test_df.shape[0], ":",(test_coverage/test_df.shape
[0])*100)
print('2. In cross validation data',cv_coverage, 'out of ',cv_df.shape[0],":" ,(cv_coverage/cv_df.shape
[0])*100)
```

Q12. How many data points are covered by total 1920 genes in test and cross validation data sets?
Ans
1. In test data 61 out of 665 : 9.172932330827068
2. In cross validation data 58 out of 532 : 10.902255639097744

3.2.3 Univariate Analysis on Text Feature

1. How many unique words are present in train data?
2. How are word frequencies distributed?
3. How to featurize text field?
4. Is the text feature useful in predicitng y_i?
5. Is the text feature stable across train, test and CV datasets?

In [37]:

```
# cls_text is a data frame
# for every row in data fram consider the 'TEXT'
# split the words by space
# make a dict with those words
# increment its count whenever we see that word

def extract_dictionary_paddle(cls_text):
    dictionary = defaultdict(int)
    for index, row in cls_text.iterrows():
        for word in row['TEXT'].split():
            dictionary[word] +=1
    return dictionary
```

In [38]:

```
import math
#https://stackoverflow.com/a/1602964
def get_text_responsecoding(df):
    text_feature_responseCoding = np.zeros((df.shape[0],9))
    for i in range(0,9):
        row_index = 0
        for index, row in df.iterrows():
            sum_prob = 0
            for word in row['TEXT'].split():
                sum_prob += math.log(((dict_list[i].get(word,0)+10 )/(total_dict[i+1]+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_prob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [39]:

```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = CountVectorizer(min_df=3, ngram_range=(1,2))
train_text_feature_onehotCoding = text_vectorizer.fit_transform(train_df['TEXT'])
# getting all the feature names (words)
train_text_features= text_vectorizer.get_feature_names()

# train_text_feature_onehotCoding.sum(axis=0).A1 will sum every row and returns (1*number of features)
vector
train_text_fea_counts = train_text_feature_onehotCoding.sum(axis=0).A1

# zip(list(text_features),text_fea_counts) will zip a word with its number of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_counts))


print("Total number of unique words in train data :", len(train_text_features))
```

```
Total number of unique words in train data : 780378
```

In [40]:

```
dict_list = []
# dict_list =[] contains 9 dictoinaries each corresponds to a class
for i in range(1,10):
    cls_text = train_df[train_df['Class']==i]
    # build a word dict based on the words in that class
    dict_list.append(extract_dictionary_paddle(cls_text))
    # append it to dict_list

# dict_list[i] is build on i'th  class text data
# total_dict is buid on whole training text data
total_dict = train_df.groupby('Class').count().to_dict().get('ID')


confuse_array = []
for i in train_text_features:
    ratios = []
    max_val = -1
    for j in range(0,9):
        ratios.append((dict_list[j][i]+10 )/(total_dict[j+1]+90))
    confuse_array.append(ratios)
confuse_array = np.array(confuse_array)
```

In [41]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(train_df)
test_text_feature_responseCoding  = get_text_responsecoding(test_df)
cv_text_feature_responseCoding  = get_text_responsecoding(cv_df)
```

In [42]:

```
# https://stackoverflow.com/a/16202486
# we convert each row values such that they sum to 1
train_text_feature_responseCoding = (train_text_feature_responseCoding.T/train_text_feature_responseCod
ing.sum(axis=1)).T
test_text_feature_responseCoding = (test_text_feature_responseCoding.T/test_text_feature_responseCoding
.sum(axis=1)).T
cv_text_feature_responseCoding = (cv_text_feature_responseCoding.T/cv_text_feature_responseCoding.sum(a
xis=1)).T
```

In [43]:

```
# don't forget to normalize every feature
train_text_feature_onehotCoding = normalize(train_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
test_text_feature_onehotCoding = text_vectorizer.transform(test_df['TEXT'])
# don't forget to normalize every feature
test_text_feature_onehotCoding = normalize(test_text_feature_onehotCoding, axis=0)

# we use the same vectorizer that was trained on train data
cv_text_feature_onehotCoding = text_vectorizer.transform(cv_df['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

In [44]:

```
#https://stackoverflow.com/a/2258273/4084039
sorted_text_fea_dict = dict(sorted(text_fea_dict.items(), key=lambda x: x[1] , reverse=True))
sorted_text_occur = np.array(list(sorted_text_fea_dict.values()))
```

In [45]:

```
# Number of words for a given frequency.
print(Counter(sorted_text_occur))
```

```
Counter({3: 150307, 4: 102413, 5: 64028, 6: 62907, 7: 48553, 8: 44824, 9: 38974, 12: 24979, 10: 21951,
11: 16250, 13: 13621, 14: 13301, 16: 13030, 15: 11945, 18: 9064, 17: 9024, 20: 6431, 19: 6367, 28: 5094
, 21: 5036, 22: 4997, 24: 4658, 30: 4317, 26: 4207, 23: 3710, 41: 3372, 25: 3300, 27: 3093, 42: 2894, 2
```

9: 2640, 32: 2565, 31: 2329, 44: 2235, 33: 2109, 36: 2016, 34: 1944, 35: 1777, 37: 1518, 40: 1469, 38: 1401, 39: 1387, 45: 1382, 68: 1370, 43: 1337, 48: 1141, 46: 1107, 63: 1104, 47: 1047, 49: 940, 50: 898, 52: 887, 51: 848, 56: 843, 60: 763, 54: 762, 53: 736, 57: 678, 55: 672, 69: 634, 58: 634, 61: 622, 64: 590, 59: 554, 65: 546, 66: 542, 62: 536, 72: 524, 70: 522, 67: 499, 71: 493, 88: 487, 84: 451, 76: 437, 73: 431, 74: 421, 75: 420, 77: 408, 80: 390, 78: 380, 82: 374, 81: 356, 90: 341, 79: 341, 85: 313, 86: 309, 83: 297, 87: 295, 89: 278, 95: 270, 91: 269, 92: 268, 93: 265, 96: 254, 99: 245, 98: 236, 100: 235, 102: 228, 94: 222, 108: 214, 104: 213, 97: 213, 126: 206, 103: 204, 105: 193, 112: 186, 101: 184, 106: 181, 110: 178, 109: 172, 115: 170, 136: 164, 116: 161, 111: 161, 114: 159, 132: 158, 118: 157, 130: 153, 125: 153, 107: 153, 119: 152, 123: 150, 121: 141, 120: 141, 124: 140, 113: 140, 140: 138, 122: 136, 117: 134, 133: 131, 131: 127, 137: 125, 134: 124, 127: 124, 128: 123, 138: 120, 141: 119, 135: 118, 145: 114, 144: 113, 129: 110, 150: 107, 147: 104, 139: 103, 153: 101, 148: 101, 146: 101, 163: 98, 156: 97, 142: 97, 151: 96, 168: 94, 149: 91, 152: 87, 176: 86, 159: 84, 157: 83, 160: 82, 158: 82, 154: 82, 162: 79, 173: 78, 165: 78, 143: 78, 180: 77, 171: 77, 170: 76, 155: 75, 164: 74, 161: 72, 190: 70, 172: 70, 179: 66, 178: 66, 177: 66, 166: 66, 195: 65, 174: 65, 186: 64, 192: 63, 183: 63, 184: 62, 182: 61, 208: 60, 185: 60, 181: 60, 207: 59, 196: 59, 169: 59, 210: 58, 167: 58, 188: 57, 209: 56, 191: 56, 175: 56, 226: 53, 198: 53, 194: 53, 243: 52, 220: 51, 234: 50, 205: 50, 201: 50, 197: 50, 187: 50, 213: 49, 189: 49, 216: 48, 204: 48, 245: 47, 222: 47, 211: 47, 202: 46, 249: 45, 236: 44, 228: 44, 230: 43, 212: 43, 273: 42, 229: 42, 225: 42, 287: 41, 232: 41, 199: 41, 193: 41, 248: 40, 219: 40, 206: 40, 242: 39, 224: 39, 221: 39, 218: 39, 203: 39, 264: 38, 223: 38, 215: 38, 214: 38, 200: 38, 294: 37, 247: 37, 239: 37, 227: 36, 217: 36, 291: 35, 290: 35, 258: 35, 257: 35, 235: 35, 280: 34, 240: 34, 277: 33, 252: 33, 250: 33, 238: 33, 233: 33, 251: 32, 231: 32, 293: 31, 283: 31, 279: 31, 276: 31, 326: 30, 272: 30, 270: 30, 259: 30, 255: 30, 241: 30, 281: 29, 269: 29, 267: 29, 266: 29, 265: 29, 303: 28, 299: 28, 288: 28, 268: 28, 260: 28, 254: 28, 253: 28, 328: 27, 308: 27, 300: 27, 297: 27, 284: 27, 261: 27, 237: 27, 289: 26, 246: 26, 244: 26, 301: 25, 295: 25, 285: 25, 278: 25, 274: 25, 262: 25, 337: 24, 271: 24, 348: 23, 325: 23, 302: 23, 263: 23, 361: 22, 355: 22, 342: 22, 338: 22, 334: 22, 315: 22, 309: 22, 305: 22, 296: 22, 286: 22, 256: 22, 344: 21, 317: 21, 282: 21, 409: 20, 359: 20, 320: 20, 316: 20, 307: 20, 418: 19, 377: 19, 372: 19, 366: 19, 336: 19, 335: 19, 331: 19, 323: 19, 318: 19, 435: 18, 429: 18, 420: 18, 362: 18, 352: 18, 345: 18, 333: 18, 329: 18, 319: 18, 314: 18, 298: 18, 275: 18, 400: 17, 382: 17, 356: 17, 349: 17, 347: 17, 343: 17, 330: 17, 321: 17, 576: 16, 463: 16, 451: 16, 441: 16, 421: 16, 403: 16, 401: 16, 395: 16, 378: 16, 373: 16, 371: 16, 363: 16, 351: 16, 332: 16, 324: 16, 312: 16, 311: 16, 310: 16, 306: 16, 304: 16, 462: 15, 442: 15, 419: 15, 404: 15, 388: 15, 380: 15, 379: 15, 376: 15, 365: 15, 357: 15, 340: 15, 339: 15, 292: 15, 557: 14, 542: 14, 528: 14, 508: 14, 443: 14, 424: 14, 422: 14, 402: 14, 398: 14, 397: 14, 396: 14, 392: 14, 390: 14, 387: 14, 386: 14, 368: 14, 353: 14, 341: 14, 327: 14, 322: 14, 313: 14, 555: 13, 496: 13, 480: 13, 466: 13, 457: 13, 452: 13, 423: 13, 410: 13, 385: 13, 384: 13, 383: 13, 374: 13, 370: 13, 367: 13, 364: 13, 350: 13, 1100: 12, 637: 12, 549: 12, 546: 12, 534: 12, 523: 12, 519: 12, 498: 12, 492: 12, 488: 12, 476: 12, 475: 12, 446: 12, 437: 12, 417: 12, 416: 12, 399: 12, 391: 12, 375: 12, 354: 12, 686: 11, 556: 11, 551: 11, 511: 11, 481: 11, 479: 11, 474: 11, 471: 11, 459: 11, 455: 11, 453: 11, 430: 11, 428: 11, 413: 11, 412: 11, 406: 11, 394: 11, 389: 11, 346: 11, 719: 10, 650: 10, 624: 10, 614: 10, 563: 10, 560: 10, 550: 10, 548: 10, 537: 10, 515: 10, 512: 10, 502: 10, 495: 10, 490: 10, 485: 10, 473: 10, 470: 10, 469: 10, 465: 10, 445: 10, 439: 10, 433: 10, 432: 10, 427: 10, 411: 10, 408: 10, 407: 10, 369: 10, 360: 10, 358: 10, 1246: 9, 695: 9, 630: 9, 617: 9, 593: 9, 587: 9, 547: 9, 540: 9, 530: 9, 520: 9, 518: 9, 513: 9, 487: 9, 486: 9, 482: 9, 478: 9, 472: 9, 460: 9, 458: 9, 450: 9, 447: 9, 431: 9, 426: 9, 425: 9, 415: 9, 414: 9, 393: 9, 381: 9, 715: 8, 714: 8, 684: 8, 665: 8, 655: 8, 627: 8, 613: 8, 594: 8, 590: 8, 582: 8, 580: 8, 573: 8, 565: 8, 541: 8, 538: 8, 533: 8, 527: 8, 524: 8, 510: 8, 501: 8, 468: 8, 456: 8, 449: 8, 448: 8, 438: 8, 436: 8, 786: 7, 780: 7, 778: 7, 767: 7, 765: 7, 762: 7, 756: 7, 747: 7, 728: 7, 723: 7, 711: 7, 705: 7, 685: 7, 675: 7, 656: 7, 653: 7, 649: 7, 633: 7, 632: 7, 629: 7, 615: 7, 598: 7, 567: 7, 561: 7, 535: 7, 529: 7, 514: 7, 509: 7, 494: 7, 444: 7, 434: 7, 1030: 6, 1007: 6, 997: 6, 970: 6, 965: 6, 940: 6, 936: 6, 930: 6, 927: 6, 921: 6, 883: 6, 881: 6, 865: 6, 852: 6, 835: 6, 832: 6, 823: 6, 812: 6, 787: 6, 784: 6, 770: 6, 764: 6, 763: 6, 752: 6, 734: 6, 722: 6, 716: 6, 713: 6, 709: 6, 694: 6, 692: 6, 691: 6, 682: 6, 681: 6, 678: 6, 664: 6, 646: 6, 642: 6, 639: 6, 638: 6, 636: 6, 626: 6, 625: 6, 623: 6, 620: 6, 605: 6, 603: 6, 601: 6, 584: 6, 581: 6, 578: 6, 575: 6, 574: 6, 572: 6, 570: 6, 569: 6, 566: 6, 564: 6, 559: 6, 558: 6, 553: 6, 552: 6, 543: 6, 536: 6, 526: 6, 517: 6, 505: 6, 504: 6, 500: 6, 499: 6, 493: 6, 484: 6, 461: 6, 405: 6, 2032: 5, 1417: 5, 1274: 5, 1250: 5, 1219: 5, 1200: 5, 1182: 5, 1160: 5, 1159: 5, 1140: 5, 1120: 5, 1105: 5, 1087: 5, 1074: 5, 1045: 5, 1040: 5, 1039: 5, 966: 5, 964: 5, 961: 5, 958: 5, 944: 5, 942: 5, 938: 5, 932: 5, 911: 5, 897: 5, 886: 5, 869: 5, 843: 5, 841: 5, 826: 5, 822: 5, 817: 5, 810: 5, 806: 5, 798: 5, 789: 5, 761: 5, 759: 5, 757: 5, 753: 5, 748: 5, 745: 5, 742: 5, 725: 5, 720: 5, 717: 5, 707: 5, 704: 5, 697: 5, 688: 5, 683: 5, 679: 5, 676: 5, 669: 5, 667: 5, 666: 5, 662: 5, 647: 5, 641: 5, 631: 5, 628: 5, 619: 5, 618: 5, 612: 5, 611: 5, 607: 5, 597: 5, 591: 5, 589: 5, 586: 5, 585: 5, 583: 5, 577: 5, 544: 5, 539: 5, 525: 5, 522: 5, 516: 5, 507: 5, 497: 5, 491: 5, 483: 5, 467: 5, 464: 5, 440: 5, 3589: 4, 2094: 4, 1919: 4, 1885: 4, 1871: 4, 1842: 4, 1620: 4, 1619: 4, 1604: 4, 1574: 4, 1568: 4, 1517: 4, 1496: 4, 1388: 4, 1385: 4, 1354: 4, 1314: 4, 1269: 4, 1251: 4, 1232: 4, 1224: 4, 1218: 4, 1212: 4, 1193: 4, 1173: 4, 1153: 4, 1141: 4, 1125: 4, 1119: 4, 1073: 4, 1067: 4, 1047: 4, 1020: 4, 1018: 4, 1015: 4, 1013: 4, 1000: 4, 995: 4, 985: 4, 984: 4, 979: 4, 971: 4, 962: 4, 959: 4, 950: 4, 946: 4, 933: 4, 929: 4, 925: 4, 905: 4, 904: 4, 892: 4, 891: 4, 880: 4, 875: 4, 867: 4, 862: 4, 859: 4, 857: 4, 851: 4, 850: 4, 848: 4, 831: 4, 814: 4, 811: 4, 808: 4, 807: 4, 803: 4, 800: 4, 797: 4, 785: 4, 783: 4, 782: 4, 771: 4, 768: 4, 755: 4, 754: 4, 750: 4, 749: 4, 744: 4, 740: 4, 733: 4, 731: 4, 730: 4, 727: 4, 718: 4, 708: 4, 706: 4, 701: 4, 700: 4, 693: 4, 689: 4, 673: 4, 672: 4, 671: 4, 670: 4, 661: 4, 660: 4, 659: 4, 658: 4, 657: 4, 651: 4, 635: 4, 610: 4, 609: 4, 608: 4, 602: 4, 600: 4, 599: 4, 596: 4, 595: 4, 592: 4, 562: 4, 531: 4, 521: 4, 454: 4, 5577: 3, 4515: 3, 4202: 3, 3723: 3, 3603: 3, 3416: 3, 3181: 3, 3164: 3, 2920: 3, 2748: 3, 2633: 3, 2603: 3, 2421: 3, 2332: 3, 2328: 3, 2301: 3, 2212: 3, 2179: 3, 2106: 3, 2090: 3, 2047: 3, 2034: 3, 2012: 3, 1945: 3, 1937: 3, 1931: 3, 1894: 3, 1858: 3, 1819: 3, 1817: 3, 1801: 3, 1796: 3, 1756: 3, 1755: 3, 1736: 3, 1717: 3, 1693: 3, 1670: 3, 1669: 3, 1663: 3, 1646: 3, 1645: 3, 1587: 3, 1558: 3, 1555: 3, 1538: 3, 1530: 3, 1525: 3, 1516: 3, 1514: 3, 1511: 3, 1504: 3, 1479: 3, 1472: 3, 1470: 3, 1452: 3, 1409: 3, 1386: 3, 1380: 3, 1368: 3, 1356: 3, 1349: 3, 1345: 3, 1336: 3, 1329: 3, 1320: 3, 1318: 3, 1313: 3, 1312: 3, 1308: 3, 1306: 3, 1305: 3, 1303: 3, 1300: 3, 1295: 3, 1290: 3, 1286: 3, 1283: 3, 1281: 3, 1272: 3, 1268: 3, 1267: 3, 1264: 3, 1263: 3, 1245: 3, 1234: 3, 1225: 3, 1221: 3, 1216: 3, 1213: 3, 1203: 3, 1198:

3, 1168: 3, 1166: 3, 1154: 3, 1149: 3, 1135: 3, 1114: 3, 1101: 3, 1098: 3, 1086: 3, 1082: 3, 1072: 3, 1
070: 3, 1069: 3, 1065: 3, 1063: 3, 1050: 3, 1049: 3, 1032: 3, 1031: 3, 1024: 3, 1017: 3, 1010: 3, 1008:
3, 1003: 3, 989: 3, 988: 3, 983: 3, 976: 3, 960: 3, 957: 3, 953: 3, 949: 3, 947: 3, 945: 3, 935: 3, 922
: 3, 910: 3, 906: 3, 903: 3, 899: 3, 894: 3, 885: 3, 879: 3, 878: 3, 871: 3, 866: 3, 863: 3, 861: 3, 86
0: 3, 856: 3, 855: 3, 845: 3, 837: 3, 836: 3, 828: 3, 821: 3, 818: 3, 815: 3, 801: 3, 794: 3, 793: 3, 7
92: 3, 781: 3, 779: 3, 775: 3, 774: 3, 769: 3, 766: 3, 743: 3, 741: 3, 739: 3, 738: 3, 735: 3, 732: 3,
729: 3, 726: 3, 712: 3, 710: 3, 702: 3, 699: 3, 696: 3, 668: 3, 663: 3, 652: 3, 643: 3, 621: 3, 616: 3,
606: 3, 604: 3, 588: 3, 579: 3, 571: 3, 554: 3, 545: 3, 532: 3, 506: 3, 503: 3, 489: 3, 477: 3, 7308: 2
, 7285: 2, 7211: 2, 6944: 2, 6409: 2, 6001: 2, 5984: 2, 5855: 2, 5426: 2, 5252: 2, 4622: 2, 4530: 2, 44
60: 2, 4322: 2, 4241: 2, 4193: 2, 4162: 2, 4079: 2, 4066: 2, 3999: 2, 3912: 2, 3866: 2, 3776: 2, 3765:
2, 3733: 2, 3727: 2, 3623: 2, 3486: 2, 3465: 2, 3461: 2, 3455: 2, 3413: 2, 3405: 2, 3328: 2, 3315: 2, 3
314: 2, 3311: 2, 3299: 2, 3249: 2, 3242: 2, 3238: 2, 3211: 2, 3209: 2, 3177: 2, 3152: 2, 3133: 2, 3117:
2, 3113: 2, 3071: 2, 3057: 2, 3041: 2, 3033: 2, 2995: 2, 2940: 2, 2910: 2, 2858: 2, 2749: 2, 2743: 2, 2
732: 2, 2707: 2, 2701: 2, 2690: 2, 2680: 2, 2671: 2, 2622: 2, 2598: 2, 2586: 2, 2575: 2, 2558: 2, 2550:
2, 2545: 2, 2526: 2, 2518: 2, 2510: 2, 2497: 2, 2478: 2, 2476: 2, 2475: 2, 2446: 2, 2430: 2, 2429: 2, 2
420: 2, 2418: 2, 2414: 2, 2410: 2, 2404: 2, 2395: 2, 2382: 2, 2369: 2, 2367: 2, 2361: 2, 2330: 2, 2321:
2, 2306: 2, 2295: 2, 2263: 2, 2250: 2, 2245: 2, 2234: 2, 2223: 2, 2222: 2, 2199: 2, 2168: 2, 2164: 2, 2
163: 2, 2113: 2, 2108: 2, 2100: 2, 2098: 2, 2096: 2, 2083: 2, 2081: 2, 2074: 2, 2068: 2, 2058: 2, 2053:
2, 2028: 2, 2019: 2, 2018: 2, 2013: 2, 2003: 2, 1999: 2, 1998: 2, 1994: 2, 1985: 2, 1978: 2, 1977: 2, 1
972: 2, 1971: 2, 1968: 2, 1948: 2, 1946: 2, 1938: 2, 1936: 2, 1930: 2, 1903: 2, 1897: 2, 1895: 2, 1891:
2, 1886: 2, 1883: 2, 1878: 2, 1874: 2, 1873: 2, 1872: 2, 1863: 2, 1859: 2, 1848: 2, 1847: 2, 1839: 2, 1
832: 2, 1830: 2, 1828: 2, 1818: 2, 1815: 2, 1814: 2, 1813: 2, 1811: 2, 1787: 2, 1786: 2, 1784: 2, 1777:
2, 1776: 2, 1771: 2, 1760: 2, 1751: 2, 1746: 2, 1737: 2, 1729: 2, 1712: 2, 1699: 2, 1690: 2, 1685: 2, 1
684: 2, 1678: 2, 1673: 2, 1665: 2, 1661: 2, 1658: 2, 1654: 2, 1652: 2, 1648: 2, 1634: 2, 1630: 2, 1628:
2, 1625: 2, 1617: 2, 1616: 2, 1606: 2, 1598: 2, 1593: 2, 1589: 2, 1582: 2, 1571: 2, 1563: 2, 1561: 2, 1
560: 2, 1556: 2, 1553: 2, 1552: 2, 1549: 2, 1543: 2, 1540: 2, 1535: 2, 1534: 2, 1526: 2, 1519: 2, 1515:
2, 1508: 2, 1505: 2, 1503: 2, 1501: 2, 1495: 2, 1493: 2, 1492: 2, 1486: 2, 1485: 2, 1481: 2, 1477: 2, 1
471: 2, 1463: 2, 1460: 2, 1458: 2, 1451: 2, 1444: 2, 1438: 2, 1434: 2, 1422: 2, 1420: 2, 1418: 2, 1415:
2, 1414: 2, 1407: 2, 1404: 2, 1402: 2, 1401: 2, 1397: 2, 1384: 2, 1381: 2, 1378: 2, 1374: 2, 1372: 2, 1
360: 2, 1358: 2, 1351: 2, 1350: 2, 1347: 2, 1342: 2, 1338: 2, 1337: 2, 1332: 2, 1330: 2, 1328: 2, 1324:
2, 1323: 2, 1322: 2, 1317: 2, 1315: 2, 1311: 2, 1297: 2, 1296: 2, 1293: 2, 1292: 2, 1291: 2, 1288: 2, 1
279: 2, 1278: 2, 1265: 2, 1256: 2, 1255: 2, 1253: 2, 1244: 2, 1243: 2, 1242: 2, 1241: 2, 1233: 2, 1231:
2, 1222: 2, 1220: 2, 1217: 2, 1215: 2, 1211: 2, 1207: 2, 1205: 2, 1199: 2, 1196: 2, 1192: 2, 1190: 2, 1
188: 2, 1186: 2, 1184: 2, 1183: 2, 1180: 2, 1176: 2, 1174: 2, 1164: 2, 1163: 2, 1158: 2, 1155: 2, 1152:
2, 1150: 2, 1148: 2, 1147: 2, 1143: 2, 1138: 2, 1131: 2, 1130: 2, 1128: 2, 1124: 2, 1122: 2, 1117: 2, 1
116: 2, 1111: 2, 1110: 2, 1107: 2, 1103: 2, 1102: 2, 1096: 2, 1095: 2, 1090: 2, 1089: 2, 1084: 2, 1083:
2, 1081: 2, 1079: 2, 1078: 2, 1075: 2, 1071: 2, 1060: 2, 1058: 2, 1056: 2, 1054: 2, 1044: 2, 1042: 2, 1
037: 2, 1035: 2, 1034: 2, 1033: 2, 1029: 2, 1028: 2, 1023: 2, 1022: 2, 1019: 2, 1014: 2, 1011: 2, 1006:
2, 1005: 2, 999: 2, 994: 2, 990: 2, 987: 2, 986: 2, 982: 2, 981: 2, 978: 2, 975: 2, 972: 2, 968: 2, 963
: 2, 956: 2, 954: 2, 952: 2, 948: 2, 943: 2, 934: 2, 931: 2, 928: 2, 920: 2, 919: 2, 918: 2, 917: 2, 91
6: 2, 915: 2, 912: 2, 909: 2, 907: 2, 900: 2, 898: 2, 893: 2, 890: 2, 888: 2, 877: 2, 876: 2, 873: 2, 8
70: 2, 858: 2, 854: 2, 849: 2, 847: 2, 846: 2, 844: 2, 840: 2, 839: 2, 838: 2, 834: 2, 830: 2, 829: 2,
827: 2, 824: 2, 820: 2, 819: 2, 816: 2, 813: 2, 809: 2, 804: 2, 802: 2, 799: 2, 790: 2, 788: 2, 777: 2,
776: 2, 773: 2, 772: 2, 760: 2, 758: 2, 751: 2, 746: 2, 737: 2, 736: 2, 724: 2, 703: 2, 690: 2, 680: 2,
654: 2, 648: 2, 645: 2, 640: 2, 634: 2, 153450: 1, 121554: 1, 82590: 1, 68228: 1, 68181: 1, 67656: 1, 6
7557: 1, 66612: 1, 63296: 1, 63147: 1, 56269: 1, 54175: 1, 50204: 1, 49785: 1, 47663: 1, 46527: 1, 4467
0: 1, 42573: 1, 41890: 1, 41701: 1, 41275: 1, 40518: 1, 39849: 1, 39039: 1, 38647: 1, 37928: 1, 37840:
1, 36348: 1, 35977: 1, 35758: 1, 35193: 1, 34443: 1, 34290: 1, 33343: 1, 32529: 1, 32408: 1, 32350: 1,
32026: 1, 29285: 1, 28646: 1, 28030: 1, 27331: 1, 26061: 1, 26042: 1, 26015: 1, 25113: 1, 25094: 1, 248
91: 1, 24773: 1, 24645: 1, 24564: 1, 24529: 1, 24508: 1, 24168: 1, 23982: 1, 22711: 1, 22373: 1, 22363:
1, 22059: 1, 21684: 1, 21293: 1, 21273: 1, 20775: 1, 20710: 1, 20493: 1, 20385: 1, 19683: 1, 19626: 1,
19620: 1, 19332: 1, 19014: 1, 18932: 1, 18905: 1, 18666: 1, 18572: 1, 18568: 1, 18448: 1, 18334: 1, 183
27: 1, 18254: 1, 18218: 1, 18148: 1, 18139: 1, 18067: 1, 17859: 1, 17843: 1, 17752: 1, 17718: 1, 17633:
1, 17445: 1, 17367: 1, 17001: 1, 16986: 1, 16972: 1, 16870: 1, 16793: 1, 16754: 1, 16473: 1, 16366: 1,
16287: 1, 16152: 1, 16066: 1, 16028: 1, 15872: 1, 15835: 1, 15827: 1, 15826: 1, 15730: 1, 15696: 1, 156
48: 1, 15286: 1, 15162: 1, 15080: 1, 15002: 1, 14814: 1, 14739: 1, 14685: 1, 14668: 1, 14631: 1, 14493:
1, 14324: 1, 14217: 1, 14142: 1, 13944: 1, 13813: 1, 13783: 1, 13771: 1, 13728: 1, 13689: 1, 13674: 1,
13528: 1, 13519: 1, 13481: 1, 13445: 1, 13172: 1, 13131: 1, 13122: 1, 13089: 1, 13022: 1, 12805: 1, 127
84: 1, 12707: 1, 12688: 1, 12665: 1, 12597: 1, 12589: 1, 12452: 1, 12451: 1, 12450: 1, 12436: 1, 12407:
1, 12404: 1, 12392: 1, 12385: 1, 12351: 1, 12335: 1, 12334: 1, 12318: 1, 12260: 1, 12207: 1, 12191: 1,
12140: 1, 12135: 1, 12131: 1, 11997: 1, 11996: 1, 11994: 1, 11933: 1, 11874: 1, 11834: 1, 11788: 1, 117
60: 1, 11751: 1, 11739: 1, 11692: 1, 11611: 1, 11417: 1, 11416: 1, 11401: 1, 11393: 1, 11344: 1, 11200:
1, 11178: 1, 11140: 1, 11035: 1, 10936: 1, 10771: 1, 10770: 1, 10766: 1, 10724: 1, 10722: 1, 10688: 1,
10682: 1, 10626: 1, 10522: 1, 10502: 1, 10341: 1, 10340: 1, 10315: 1, 10303: 1, 10227: 1, 10216: 1, 101
89: 1, 10182: 1, 10176: 1, 10131: 1, 10083: 1, 10048: 1, 10010: 1, 9937: 1, 9924: 1, 9902: 1, 9894: 1,
9764: 1, 9751: 1, 9736: 1, 9718: 1, 9706: 1, 9672: 1, 9671: 1, 9547: 1, 9511: 1, 9479: 1, 9471: 1, 9456
: 1, 9303: 1, 9270: 1, 9253: 1, 9240: 1, 9219: 1, 9200: 1, 9170: 1, 9150: 1, 9119: 1, 9093: 1, 9091: 1,
9073: 1, 9053: 1, 9032: 1, 9008: 1, 8996: 1, 8967: 1, 8945: 1, 8930: 1, 8918: 1, 8893: 1, 8869: 1, 8862
: 1, 8845: 1, 8777: 1, 8746: 1, 8702: 1, 8675: 1, 8604: 1, 8517: 1, 8511: 1, 8450: 1, 8439: 1, 8419: 1,
8415: 1, 8364: 1, 8334: 1, 8327: 1, 8284: 1, 8258: 1, 8215: 1, 8188: 1, 8169: 1, 8139: 1, 8118: 1, 8117
: 1, 8113: 1, 8109: 1, 8074: 1, 8029: 1, 8003: 1, 8001: 1, 7995: 1, 7981: 1, 7956: 1, 7934: 1, 7920: 1,
7904: 1, 7903: 1, 7900: 1, 7888: 1, 7875: 1, 7864: 1, 7857: 1, 7852: 1, 7844: 1, 7834: 1, 7808: 1, 7806
: 1, 7788: 1, 7703: 1, 7664: 1, 7657: 1, 7654: 1, 7652: 1, 7586: 1, 7553: 1, 7546: 1, 7543: 1, 7528: 1,
7501: 1, 7469: 1, 7446: 1, 7433: 1, 7399: 1, 7371: 1, 7361: 1, 7352: 1, 7341: 1, 7336: 1, 7312: 1, 7304
: 1, 7298: 1, 7281: 1, 7260: 1, 7234: 1, 7207: 1, 7165: 1, 7143: 1, 7108: 1, 7086: 1, 7077: 1, 7041: 1,
7037: 1, 7024: 1, 7022: 1, 6995: 1, 6989: 1, 6986: 1, 6966: 1, 6962: 1, 6957: 1, 6955: 1, 6948: 1, 6920
: 1, 6908: 1, 6887: 1, 6885: 1, 6883: 1, 6829: 1, 6800: 1, 6799: 1, 6790: 1, 6787: 1, 6707: 1, 6706: 1,

: 1, 6900: 1, 6887: 1, 6865: 1, 6865: 1, 6829: 1, 6800: 1, 6799: 1, 6790: 1, 6767: 1, 6767: 1, 6766: 1,
6698: 1, 6689: 1, 6675: 1, 6663: 1, 6657: 1, 6648: 1, 6629: 1, 6603: 1, 6588: 1, 6583: 1, 6574: 1, 6572
: 1, 6569: 1, 6544: 1, 6526: 1, 6489: 1, 6488: 1, 6480: 1, 6459: 1, 6447: 1, 6445: 1, 6426: 1, 6415: 1,
6405: 1, 6375: 1, 6368: 1, 6351: 1, 6344: 1, 6337: 1, 6325: 1, 6312: 1, 6307: 1, 6304: 1, 6296: 1, 6295
: 1, 6257: 1, 6232: 1, 6226: 1, 6221: 1, 6186: 1, 6154: 1, 6149: 1, 6143: 1, 6138: 1, 6116: 1, 6108: 1,
6097: 1, 6074: 1, 6068: 1, 6046: 1, 6031: 1, 5996: 1, 5983: 1, 5973: 1, 5955: 1, 5946: 1, 5906: 1, 5824
: 1, 5823: 1, 5811: 1, 5810: 1, 5809: 1, 5799: 1, 5769: 1, 5763: 1, 5761: 1, 5742: 1, 5731: 1, 5730: 1,
5727: 1, 5717: 1, 5712: 1, 5711: 1, 5685: 1, 5679: 1, 5676: 1, 5663: 1, 5661: 1, 5659: 1, 5658: 1, 5628
: 1, 5615: 1, 5595: 1, 5591: 1, 5582: 1, 5558: 1, 5547: 1, 5518: 1, 5516: 1, 5491: 1, 5481: 1, 5455: 1,
5454: 1, 5437: 1, 5421: 1, 5417: 1, 5391: 1, 5384: 1, 5375: 1, 5359: 1, 5351: 1, 5346: 1, 5331: 1, 5324
: 1, 5310: 1, 5292: 1, 5287: 1, 5283: 1, 5256: 1, 5255: 1, 5242: 1, 5212: 1, 5209: 1, 5203: 1, 5198: 1,
5192: 1, 5171: 1, 5163: 1, 5150: 1, 5140: 1, 5130: 1, 5119: 1, 5107: 1, 5102: 1, 5101: 1, 5084: 1, 5073
: 1, 5071: 1, 5070: 1, 5061: 1, 5051: 1, 5047: 1, 5041: 1, 5039: 1, 5033: 1, 5020: 1, 5013: 1, 5002: 1,
5000: 1, 4999: 1, 4983: 1, 4980: 1, 4953: 1, 4946: 1, 4940: 1, 4926: 1, 4920: 1, 4915: 1, 4903: 1, 4897
: 1, 4889: 1, 4888: 1, 4885: 1, 4881: 1, 4874: 1, 4872: 1, 4860: 1, 4857: 1, 4855: 1, 4852: 1, 4842: 1,
4834: 1, 4833: 1, 4825: 1, 4812: 1, 4807: 1, 4805: 1, 4795: 1, 4792: 1, 4791: 1, 4789: 1, 4760: 1, 4746
: 1, 4736: 1, 4732: 1, 4724: 1, 4722: 1, 4676: 1, 4661: 1, 4642: 1, 4634: 1, 4629: 1, 4624: 1, 4618: 1,
4614: 1, 4583: 1, 4574: 1, 4548: 1, 4546: 1, 4541: 1, 4538: 1, 4521: 1, 4512: 1, 4506: 1, 4504: 1, 4503
: 1, 4501: 1, 4500: 1, 4468: 1, 4463: 1, 4456: 1, 4454: 1, 4444: 1, 4443: 1, 4439: 1, 4422: 1, 4413: 1,
4404: 1, 4403: 1, 4394: 1, 4364: 1, 4359: 1, 4350: 1, 4342: 1, 4340: 1, 4330: 1, 4329: 1, 4318: 1, 4311
: 1, 4304: 1, 4302: 1, 4297: 1, 4285: 1, 4280: 1, 4272: 1, 4253: 1, 4243: 1, 4242: 1, 4239: 1, 4232: 1,
4230: 1, 4220: 1, 4215: 1, 4211: 1, 4209: 1, 4198: 1, 4194: 1, 4190: 1, 4165: 1, 4161: 1, 4158: 1, 4135
: 1, 4132: 1, 4130: 1, 4126: 1, 4116: 1, 4114: 1, 4108: 1, 4107: 1, 4089: 1, 4083: 1, 4073: 1, 4069: 1,
4043: 1, 4040: 1, 4037: 1, 4035: 1, 4012: 1, 4005: 1, 3995: 1, 3993: 1, 3990: 1, 3987: 1, 3983: 1, 3978
: 1, 3971: 1, 3964: 1, 3961: 1, 3957: 1, 3954: 1, 3951: 1, 3944: 1, 3939: 1, 3937: 1, 3931: 1, 3927: 1,
3922: 1, 3916: 1, 3905: 1, 3904: 1, 3892: 1, 3889: 1, 3876: 1, 3844: 1, 3841: 1, 3839: 1, 3836: 1, 3834
: 1, 3832: 1, 3830: 1, 3823: 1, 3805: 1, 3798: 1, 3796: 1, 3795: 1, 3793: 1, 3783: 1, 3782: 1, 3770: 1,
3758: 1, 3752: 1, 3748: 1, 3743: 1, 3742: 1, 3738: 1, 3730: 1, 3728: 1, 3725: 1, 3711: 1, 3708: 1, 3706
: 1, 3705: 1, 3699: 1, 3697: 1, 3695: 1, 3688: 1, 3687: 1, 3684: 1, 3678: 1, 3672: 1, 3670: 1, 3665: 1,
3661: 1, 3653: 1, 3652: 1, 3650: 1, 3644: 1, 3640: 1, 3636: 1, 3631: 1, 3624: 1, 3614: 1, 3608: 1, 3606
: 1, 3599: 1, 3590: 1, 3576: 1, 3569: 1, 3564: 1, 3562: 1, 3557: 1, 3548: 1, 3545: 1, 3541: 1, 3540: 1,
3530: 1, 3529: 1, 3523: 1, 3519: 1, 3512: 1, 3511: 1, 3510: 1, 3507: 1, 3504: 1, 3500: 1, 3496: 1, 3488
: 1, 3487: 1, 3481: 1, 3479: 1, 3471: 1, 3466: 1, 3458: 1, 3457: 1, 3448: 1, 3444: 1, 3433: 1, 3432: 1,
3431: 1, 3424: 1, 3423: 1, 3417: 1, 3414: 1, 3399: 1, 3397: 1, 3396: 1, 3391: 1, 3386: 1, 3375: 1, 3370
: 1, 3369: 1, 3364: 1, 3362: 1, 3353: 1, 3352: 1, 3350: 1, 3349: 1, 3335: 1, 3334: 1, 3333: 1, 3332: 1,
3326: 1, 3323: 1, 3306: 1, 3305: 1, 3304: 1, 3297: 1, 3291: 1, 3289: 1, 3287: 1, 3280: 1, 3278: 1, 3276
: 1, 3275: 1, 3272: 1, 3270: 1, 3262: 1, 3259: 1, 3252: 1, 3240: 1, 3227: 1, 3223: 1, 3222: 1, 3220: 1,
3219: 1, 3218: 1, 3213: 1, 3210: 1, 3197: 1, 3192: 1, 3189: 1, 3184: 1, 3182: 1, 3168: 1, 3149: 1, 3147
: 1, 3137: 1, 3134: 1, 3128: 1, 3125: 1, 3123: 1, 3120: 1, 3115: 1, 3112: 1, 3109: 1, 3105: 1, 3093: 1,
3089: 1, 3088: 1, 3068: 1, 3062: 1, 3055: 1, 3052: 1, 3045: 1, 3044: 1, 3039: 1, 3030: 1, 3026: 1, 3025
: 1, 3023: 1, 3022: 1, 3014: 1, 3010: 1, 3008: 1, 3007: 1, 3005: 1, 3000: 1, 2989: 1, 2982: 1, 2980: 1,
2972: 1, 2971: 1, 2967: 1, 2963: 1, 2958: 1, 2957: 1, 2944: 1, 2941: 1, 2939: 1, 2936: 1, 2932: 1, 2930
: 1, 2929: 1, 2921: 1, 2913: 1, 2909: 1, 2902: 1, 2900: 1, 2899: 1, 2890: 1, 2888: 1, 2887: 1, 2885: 1,
2884: 1, 2883: 1, 2882: 1, 2881: 1, 2878: 1, 2877: 1, 2867: 1, 2864: 1, 2859: 1, 2854: 1, 2843: 1, 2841
: 1, 2830: 1, 2827: 1, 2816: 1, 2812: 1, 2811: 1, 2810: 1, 2806: 1, 2797: 1, 2795: 1, 2792: 1, 2779: 1,
2775: 1, 2772: 1, 2765: 1, 2764: 1, 2752: 1, 2751: 1, 2747: 1, 2742: 1, 2737: 1, 2734: 1, 2733: 1, 2727
: 1, 2723: 1, 2720: 1, 2718: 1, 2717: 1, 2712: 1, 2711: 1, 2702: 1, 2698: 1, 2689: 1, 2679: 1, 2677: 1,
2675: 1, 2672: 1, 2670: 1, 2666: 1, 2665: 1, 2662: 1, 2661: 1, 2659: 1, 2658: 1, 2641: 1, 2640: 1, 2637
: 1, 2630: 1, 2627: 1, 2625: 1, 2624: 1, 2621: 1, 2616: 1, 2613: 1, 2606: 1, 2605: 1, 2602: 1, 2597: 1,
2596: 1, 2594: 1, 2591: 1, 2590: 1, 2585: 1, 2584: 1, 2581: 1, 2579: 1, 2577: 1, 2573: 1, 2566: 1, 2564
: 1, 2562: 1, 2560: 1, 2559: 1, 2556: 1, 2555: 1, 2554: 1, 2544: 1, 2541: 1, 2537: 1, 2528: 1, 2523: 1,
2515: 1, 2514: 1, 2512: 1, 2508: 1, 2507: 1, 2506: 1, 2504: 1, 2502: 1, 2498: 1, 2494: 1, 2491: 1, 2490
: 1, 2486: 1, 2485: 1, 2483: 1, 2482: 1, 2477: 1, 2474: 1, 2472: 1, 2471: 1, 2467: 1, 2460: 1, 2454: 1,
2447: 1, 2439: 1, 2434: 1, 2433: 1, 2432: 1, 2431: 1, 2428: 1, 2427: 1, 2426: 1, 2419: 1, 2416: 1, 2413
: 1, 2408: 1, 2407: 1, 2402: 1, 2396: 1, 2391: 1, 2385: 1, 2379: 1, 2376: 1, 2372: 1, 2370: 1, 2368: 1,
2366: 1, 2363: 1, 2358: 1, 2357: 1, 2356: 1, 2348: 1, 2347: 1, 2345: 1, 2339: 1, 2337: 1, 2326: 1, 2325
: 1, 2324: 1, 2320: 1, 2319: 1, 2318: 1, 2316: 1, 2313: 1, 2310: 1, 2307: 1, 2304: 1, 2300: 1, 2293: 1,
2285: 1, 2282: 1, 2275: 1, 2268: 1, 2267: 1, 2265: 1, 2262: 1, 2261: 1, 2260: 1, 2259: 1, 2255: 1, 2254
: 1, 2253: 1, 2246: 1, 2242: 1, 2240: 1, 2238: 1, 2235: 1, 2230: 1, 2229: 1, 2228: 1, 2227: 1, 2224: 1,
2217: 1, 2216: 1, 2207: 1, 2203: 1, 2202: 1, 2201: 1, 2198: 1, 2192: 1, 2190: 1, 2189: 1, 2183: 1, 2180
: 1, 2178: 1, 2176: 1, 2174: 1, 2173: 1, 2172: 1, 2170: 1, 2167: 1, 2161: 1, 2159: 1, 2155: 1, 2153: 1,
2151: 1, 2144: 1, 2143: 1, 2141: 1, 2140: 1, 2136: 1, 2135: 1, 2130: 1, 2127: 1, 2125: 1, 2122: 1, 2121
: 1, 2117: 1, 2115: 1, 2112: 1, 2109: 1, 2107: 1, 2105: 1, 2104: 1, 2099: 1, 2095: 1, 2093: 1, 2091: 1,
2086: 1, 2085: 1, 2084: 1, 2082: 1, 2080: 1, 2079: 1, 2077: 1, 2071: 1, 2069: 1, 2067: 1, 2065: 1, 2063
: 1, 2062: 1, 2061: 1, 2052: 1, 2049: 1, 2046: 1, 2044: 1, 2043: 1, 2042: 1, 2040: 1, 2038: 1, 2033: 1,
2030: 1, 2029: 1, 2025: 1, 2021: 1, 2014: 1, 2011: 1, 2009: 1, 2005: 1, 2001: 1, 1992: 1, 1987: 1, 1982
: 1, 1981: 1, 1979: 1, 1973: 1, 1967: 1, 1966: 1, 1961: 1, 1959: 1, 1957: 1, 1956: 1, 1953: 1, 1952: 1,
1951: 1, 1941: 1, 1932: 1, 1928: 1, 1924: 1, 1923: 1, 1922: 1, 1916: 1, 1915: 1, 1914: 1, 1913: 1, 1912
: 1, 1911: 1, 1910: 1, 1906: 1, 1905: 1, 1904: 1, 1902: 1, 1892: 1, 1887: 1, 1884: 1, 1877: 1, 1875: 1,
1869: 1, 1866: 1, 1865: 1, 1861: 1, 1860: 1, 1857: 1, 1856: 1, 1855: 1, 1850: 1, 1849: 1, 1844: 1, 1841
: 1, 1840: 1, 1838: 1, 1836: 1, 1835: 1, 1834: 1, 1829: 1, 1822: 1, 1821: 1, 1816: 1, 1812: 1, 1808: 1,
1804: 1, 1799: 1, 1789: 1, 1788: 1, 1781: 1, 1780: 1, 1779: 1, 1775: 1, 1768: 1, 1764: 1, 1762: 1, 1759
: 1, 1758: 1, 1754: 1, 1752: 1, 1749: 1, 1748: 1, 1744: 1, 1743: 1, 1742: 1, 1741: 1, 1739: 1, 1734: 1,
1727: 1, 1725: 1, 1724: 1, 1723: 1, 1716: 1, 1715: 1, 1713: 1, 1711: 1, 1710: 1, 1708: 1, 1707: 1, 1705
: 1, 1704: 1, 1703: 1, 1702: 1, 1701: 1, 1696: 1, 1695: 1, 1694: 1, 1692: 1, 1691: 1, 1689: 1, 1688: 1,
1682: 1, 1680: 1, 1674: 1, 1672: 1, 1668: 1, 1667: 1, 1664: 1, 1662: 1, 1660: 1, 1659: 1, 1656: 1, 1655
: 1, 1649: 1, 1647: 1, 1644: 1, 1642: 1, 1640: 1, 1638: 1, 1637: 1, 1636: 1, 1633: 1, 1627: 1, 1626: 1,
1623: 1, 1621: 1, 1615: 1, 1613: 1, 1612: 1, 1611: 1, 1608: 1, 1602: 1, 1600: 1, 1592: 1, 1591: 1, 1590

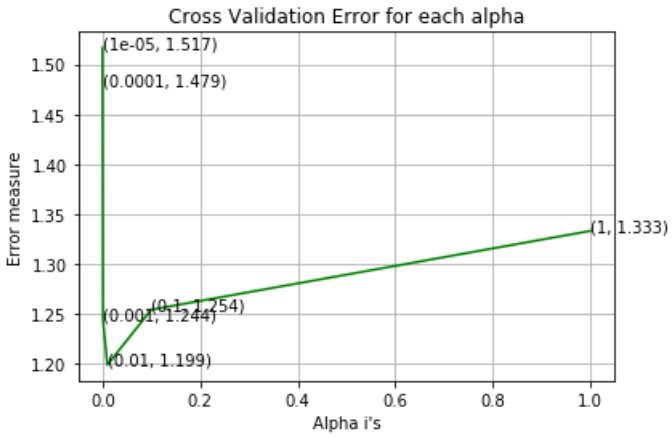
```
1623: 1, 1621: 1, 1619: 1, 1615: 1, 1612: 1, 1611: 1, 1608: 1, 1602: 1, 1600: 1, 1592: 1, 1591: 1, 1590
: 1, 1585: 1, 1584: 1, 1583: 1, 1579: 1, 1576: 1, 1575: 1, 1573: 1, 1572: 1, 1567: 1, 1564: 1, 1559: 1,
1554: 1, 1545: 1, 1533: 1, 1532: 1, 1531: 1, 1529: 1, 1527: 1, 1522: 1, 1521: 1, 1520: 1, 1518: 1, 1509
: 1, 1507: 1, 1502: 1, 1500: 1, 1499: 1, 1498: 1, 1497: 1, 1494: 1, 1488: 1, 1484: 1, 1483: 1, 1482: 1,
1480: 1, 1478: 1, 1476: 1, 1473: 1, 1469: 1, 1467: 1, 1465: 1, 1464: 1, 1462: 1, 1461: 1, 1459: 1, 1454
: 1, 1450: 1, 1449: 1, 1448: 1, 1445: 1, 1442: 1, 1441: 1, 1439: 1, 1437: 1, 1436: 1, 1435: 1, 1432: 1,
1430: 1, 1429: 1, 1427: 1, 1426: 1, 1425: 1, 1424: 1, 1423: 1, 1421: 1, 1419: 1, 1411: 1, 1410: 1, 1408
: 1, 1406: 1, 1403: 1, 1396: 1, 1394: 1, 1390: 1, 1389: 1, 1387: 1, 1382: 1, 1377: 1, 1376: 1, 1375: 1,
1370: 1, 1369: 1, 1367: 1, 1365: 1, 1363: 1, 1361: 1, 1353: 1, 1343: 1, 1340: 1, 1339: 1, 1335: 1, 1333
: 1, 1325: 1, 1321: 1, 1316: 1, 1310: 1, 1309: 1, 1307: 1, 1304: 1, 1302: 1, 1299: 1, 1294: 1, 1289: 1,
1287: 1, 1285: 1, 1284: 1, 1282: 1, 1280: 1, 1277: 1, 1275: 1, 1273: 1, 1270: 1, 1266: 1, 1262: 1, 1261
: 1, 1260: 1, 1254: 1, 1249: 1, 1247: 1, 1240: 1, 1235: 1, 1230: 1, 1228: 1, 1227: 1, 1223: 1, 1214: 1,
1209: 1, 1206: 1, 1202: 1, 1201: 1, 1197: 1, 1195: 1, 1194: 1, 1191: 1, 1189: 1, 1187: 1, 1181: 1, 1178
: 1, 1177: 1, 1175: 1, 1171: 1, 1170: 1, 1169: 1, 1161: 1, 1157: 1, 1156: 1, 1151: 1, 1145: 1, 1144: 1,
1139: 1, 1137: 1, 1136: 1, 1134: 1, 1133: 1, 1132: 1, 1129: 1, 1127: 1, 1118: 1, 1115: 1, 1113: 1, 1112
: 1, 1109: 1, 1108: 1, 1097: 1, 1093: 1, 1092: 1, 1091: 1, 1080: 1, 1077: 1, 1076: 1, 1062: 1, 1061: 1,
1059: 1, 1057: 1, 1055: 1, 1053: 1, 1046: 1, 1043: 1, 1041: 1, 1038: 1, 1036: 1, 1027: 1, 1021: 1, 1016
: 1, 1012: 1, 1009: 1, 1004: 1, 1002: 1, 1001: 1, 993: 1, 992: 1, 991: 1, 977: 1, 974: 1, 967: 1, 955:
1, 951: 1, 941: 1, 939: 1, 937: 1, 926: 1, 924: 1, 923: 1, 914: 1, 913: 1, 902: 1, 896: 1, 889: 1, 887:
1, 884: 1, 874: 1, 872: 1, 868: 1, 842: 1, 833: 1, 825: 1, 805: 1, 796: 1, 795: 1, 791: 1, 721: 1, 687:
1, 677: 1, 674: 1, 644: 1, 622: 1, 568: 1})
```

In [46]:

```
# Train a Logistic regression+Calibration model using text features whicha re on-hot encoded
alpha = [10 ** x for x in range(-5, 1)]

# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_mo
del.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=N
one, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0,
power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------


cv_log_error_array=[]
for i in alpha:
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_text_feature_onehotCoding, y_train)

    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_text_feature_onehotCoding, y_train)
    predict_y = sig_clf.predict_proba(cv_text_feature_onehotCoding)
    cv_log_error_array.append(log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
    print('For values of alpha = ', i, "The log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_
, eps=1e-15))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],np.round(txt,3)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_text_feature_onehotCoding, y_train)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_text_feature_onehotCoding, y_train)

predict_y = sig_clf.predict_proba(train_text_feature_onehotCoding)
```

```
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, pred
ict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_
cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_text_feature_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predic
t_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.5167231920253488
For values of alpha =  0.0001 The log loss is: 1.4794519724011637
For values of alpha =  0.001 The log loss is: 1.2444734717983619
For values of alpha =  0.01 The log loss is: 1.1988087536395524
For values of alpha =  0.1 The log loss is: 1.2541335677728367
For values of alpha =  1 The log loss is: 1.3331017378409953
```




```
For values of best alpha =  0.01 The train log loss is: 0.7301697590831281
For values of best alpha =  0.01 The cross validation log loss is: 1.1988087536395524
For values of best alpha =  0.01 The test log loss is: 1.1950351201142333
```

Q. Is the Text feature stable across all the data sets (Test, Train, Cross validation)?

Ans. Yes, it seems like!

In [59]:

```
def get_intersec_text(df):
    df_text_vec = CountVectorizer(min_df=3, ngram_range=(1,2))
    df_text_fea = df_text_vec.fit_transform(df['TEXT'])
    df_text_features = df_text_vec.get_feature_names()

    df_text_fea_counts = df_text_fea.sum(axis=0).A1
    df_text_fea_dict = dict(zip(list(df_text_features),df_text_fea_counts))
    len1 = len(set(df_text_features))
    len2 = len(set(train_text_features) & set(df_text_features))
    return len1,len2
```

In [60]:

```
len1,len2 = get_intersec_text(test_df)
print(np.round((len2/len1)*100, 3), "% of word of test data appeared in train data")
len1,len2 = get_intersec_text(cv_df)
print(np.round((len2/len1)*100, 3), "% of word of Cross Validation appeared in train data")
```

```
92.847 % of word of test data appeared in train data
95.593 % of word of Cross Validation appeared in train data
```

4. Machine Learning Models

In [50]:

```
#Data preparation for ML models.

#Misc. functionns for ML models


def predict_and_plot_confusion_matrix(train_x, train_y,test_x, test_y, clf):
    clf.fit(train_x, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x, train_y)
    pred_y = sig_clf.predict(test_x)

    # for calculating log_loss we willl provide the array of probabilities belongs to each class
    print("Log loss :",log_loss(test_y, sig_clf.predict_proba(test_x)))
    # calculating the number of data points that are misclassified
    print("Number of mis-classified points :", np.count_nonzero((pred_y- test_y))/test_y.shape[0])
    plot_confusion_matrix(test_y, pred_y)
```

In [51]:

```
def report_log_loss(train_x, train_y, test_x, test_y,  clf):
    clf.fit(train_x, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x, train_y)
    sig_clf_probs = sig_clf.predict_proba(test_x)
    return log_loss(test_y, sig_clf_probs, eps=1e-15)
```

In [78]:

```
# generate ngram from sentence
# Ref: http://www.albertauyeung.com/post/generating-ngrams-python/
def generate_ngrams(s, n):
    # Convert to lowercases
    s = s.lower()

    # Replace all none alphanumeric characters with spaces
    s = re.sub(r'[^a-zA-Z0-9\s]', ' ', s)

    # Break sentence in the token, remove empty tokens
    tokens = [token for token in s.split(" ") if token != ""]

    # Use the zip function to help us generate n-grams
    # Concatentate the tokens into ngrams and return
    ngrams = zip(*[tokens[i:] for i in range(n)])
    return [" ".join(ngram) for ngram in ngrams]
```

In [95]:

```
# this function will be used just for naive bayes
# for the given indices, we will print the name of the features
# and we will check whether the feature present in the test point text or not
def get_impfeature_names(indices, text, gene, var, no_features):
    gene_count_vec = CountVectorizer()
    var_count_vec = CountVectorizer()
    text_count_vec = CountVectorizer(min_df=3, ngram_range=(1,2))

    gene_vec = gene_count_vec.fit(train_df['Gene'])
    var_vec  = var_count_vec.fit(train_df['Variation'])
    text_vec = text_count_vec.fit(train_df['TEXT'])

    all_features = gene_count_vec.get_feature_names() + var_count_vec.get_feature_names() + text_count_
vec.get_feature_names()

    fea1_len = len(gene_vec.get_feature_names())
    fea2_len = len(var_count_vec.get_feature_names())

    word_present = 0
    for i,v in enumerate(indices):
        if (v < fea1_len):
            word = all_features[v]
            yes_no = True if word == gene else False
            if yes_no:
```

```
            if yes_no:
                word_present += 1
                print(i, "Gene feature [{}] present in test data point [{}]".format(word,yes_no))
        elif (v < fea1_len+fea2_len):
            word = all_features[v]
            yes_no = True if word == var else False
            if yes_no:
                word_present += 1
                print(i, "variation feature [{}] present in test data point [{}]".format(word,yes_no))
        else:
            word = all_features[v]
            # list of text contain unigram
            text_1 = generate_ngrams(text,1)
            # list of text contain bigram
            text_2 = generate_ngrams(text,2)

            yes_no_1 = True if word in text_1 else False
            yes_no_2 = True if word in text_2 else False
            if yes_no_1:
                word_present += 1
                print(i, "Text feature [{}] present in test data point [{}]".format(word,yes_no_1))
            if yes_no_2:
                word_present += 1
                print(i, "Text feature [{}] present in test data point [{}]".format(word,yes_no_2))

    print("Out of the top ",no_features," features ", word_present, "are present in query point")
```

In [93]:

```
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

```
147 Text feature [greater ability] present in test data point [True]
247 Text feature [likely sufficient] present in test data point [True]
506 Text feature [thus appear] present in test data point [True]
971 Text feature [inhibitors determine] present in test data point [True]
Out of the top  1000  features  4 are present in query point
```

Stacking the three types of features

In [53]:

```
# merging gene, variance and text features

# building train, test and cross validation data sets
# a = [[1, 2],
#      [3, 4]]
# b = [[4, 5],
#      [6, 7]]
# hstack(a, b) = [[1, 2, 4, 5],
#                 [ 3, 4, 6, 7]]

train_gene_var_onehotCoding = hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCod
ing))
test_gene_var_onehotCoding = hstack((test_gene_feature_onehotCoding,test_variation_feature_onehotCoding
))
cv_gene_var_onehotCoding = hstack((cv_gene_feature_onehotCoding,cv_variation_feature_onehotCoding))
```

```
train_x_onehotCoding = hstack((train_gene_var_onehotCoding, train_text_feature_onehotCoding)).tocsr()
train_y = np.array(list(train_df['Class']))

test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(test_df['Class']))

cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
cv_y = np.array(list(cv_df['Class']))


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseCoding,train_variation_feature_re
sponseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCoding,test_variation_feature_respo
nseCoding))
cv_gene_var_responseCoding = np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCod
ing))

train_x_responseCoding = np.hstack((train_gene_var_responseCoding, train_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_text_feature_responseCoding))
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_feature_responseCoding))
```

In [54]:

```
print("One hot encoding features :")
print("(number of data points * number of features) in train data = ", train_x_onehotCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_onehotCoding.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_onehotCoding.shap
e)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 782561)
(number of data points * number of features) in test data =  (665, 782561)
(number of data points * number of features) in cross validation data = (532, 782561)
```

In [55]:

```
print(" Response encoding features :")
print("(number of data points * number of features) in train data = ", train_x_responseCoding.shape)
print("(number of data points * number of features) in test data = ", test_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation data =", cv_x_responseCoding.sh
ape)
```

```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

4. Logistic Regression

4.1 With Class balancing

4.1.1. Hyper paramter tuning

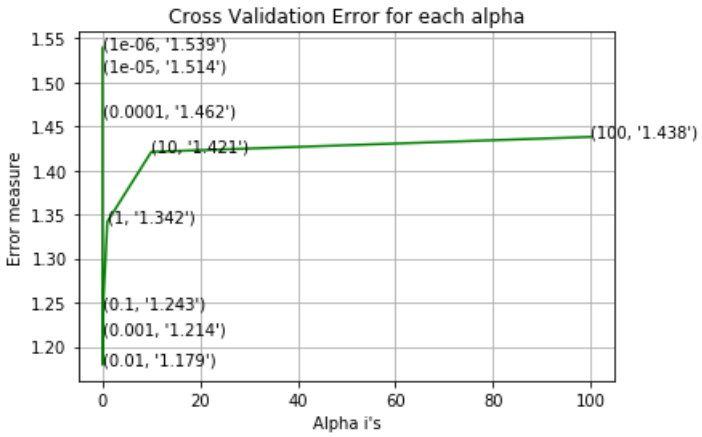
In [56]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_mo
del.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=N
one, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0,
power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
```

```
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-intuition-1/
#-------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-6, 3)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(class_weight='balanced', alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    # to avoid rounding error while multiplying probabilites we use log-probability estimates
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.5394824675008616
for alpha = 1e-05
Log Loss : 1.5139569947164189
for alpha = 0.0001
Log Loss : 1.4616013682227202
for alpha = 0.001
Log Loss : 1.2141890807075546
for alpha = 0.01
Log Loss : 1.1792090537315691
for alpha = 0.1
Log Loss : 1.243222167047315
```

```
for alpha = 1
Log Loss : 1.3419419370775192
for alpha = 10
Log Loss : 1.4211980261907622
for alpha = 100
Log Loss : 1.438172305924132
```



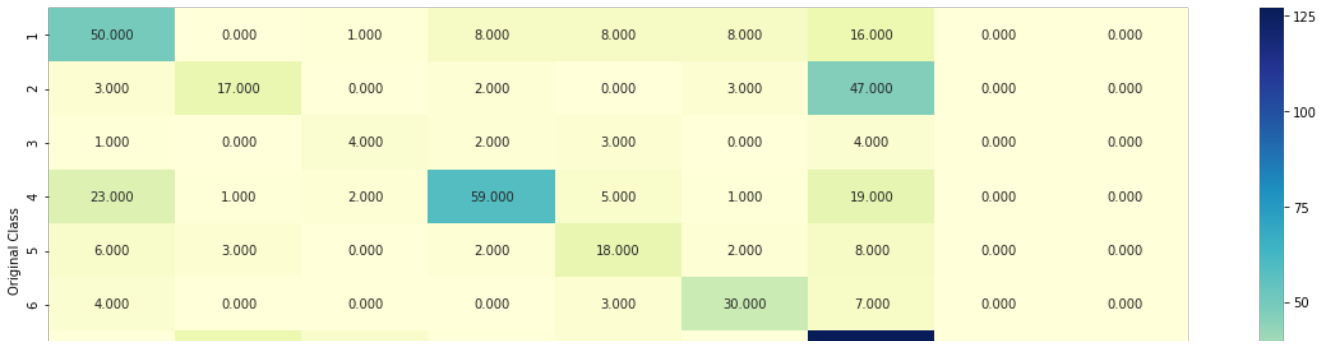

```
For values of best alpha =  0.01 The train log loss is: 0.7070545649046233
For values of best alpha =  0.01 The cross validation log loss is: 1.1792090537315691
For values of best alpha =  0.01 The test log loss is: 1.1776257465537177
```

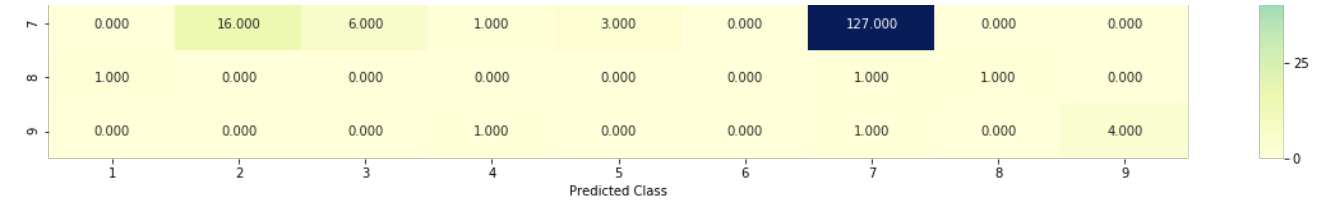
4.1.2. Testing the model with best hyper paramters

In [57]:

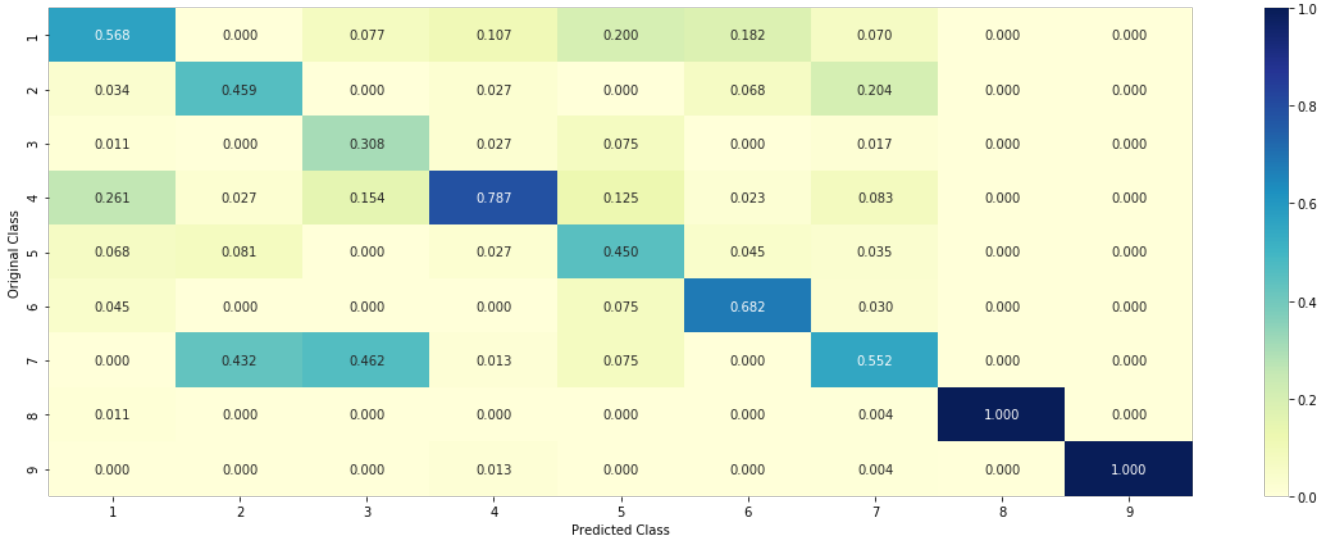
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-intuition-1/
#-------------------------------
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

```
Log loss : 1.1792090537315691
Number of mis-classified points : 0.41729323308270677
-------------------- Confusion matrix --------------------
```

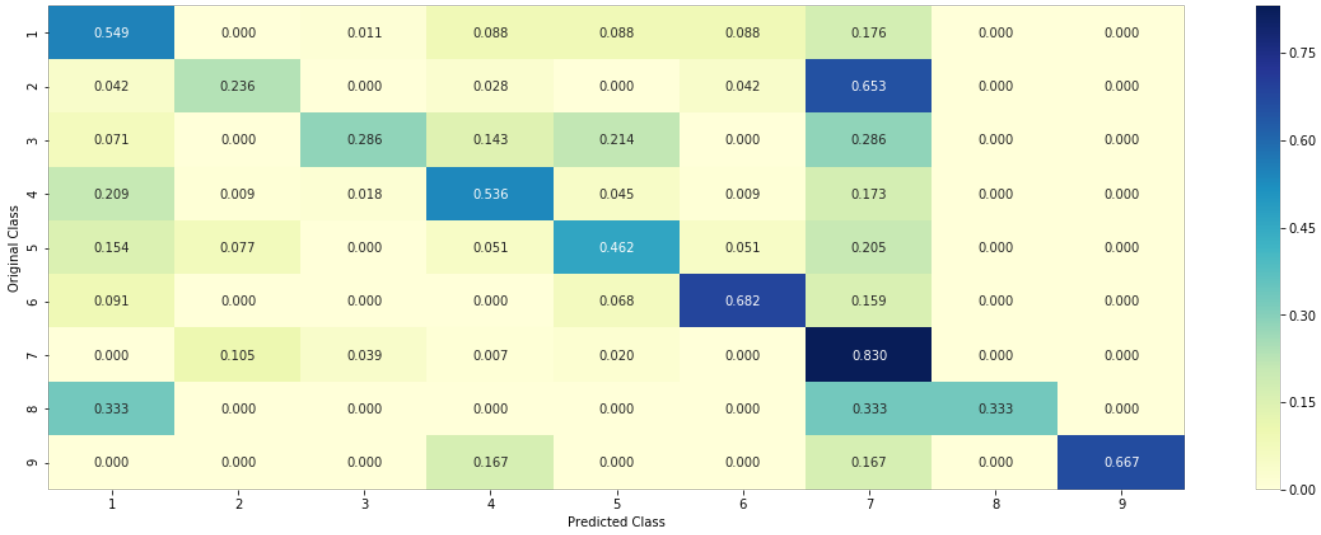





```
-------------------- Precision matrix (Column Sum=1) --------------------
```




```
-------------------- Recall matrix (Row sum=1) --------------------
```



4.1.3. Feature Importance

4.1.3.1. Correctly Classified point

In [98]:

```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='log', random_
state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 1500
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_i
```

```
ndex]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 6
Predicted Class Probabilities: [[0.0121 0.0106 0.001  0.0069 0.0656 0.892  0.0072 0.0028 0.0018]]
Actual Class : 6
--------------------------------------------------
900 Text feature [african] present in test data point [True]
1187 Text feature [given family] present in test data point [True]
Out of the top  1500  features  2 are present in query point
```

4.1.3.2. Incorrectly Classified point

In [97]:

```
test_point_index = 2
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_i
ndex]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.1203 0.1052 0.0144 0.1138 0.0458 0.0308 0.5566 0.0063 0.0068]]
Actual Class : 6
--------------------------------------------------
147 Text feature [greater ability] present in test data point [True]
247 Text feature [likely sufficient] present in test data point [True]
506 Text feature [thus appear] present in test data point [True]
971 Text feature [inhibitors determine] present in test data point [True]
Out of the top  1000  features  4 are present in query point
```

4.2. Without Class balancing

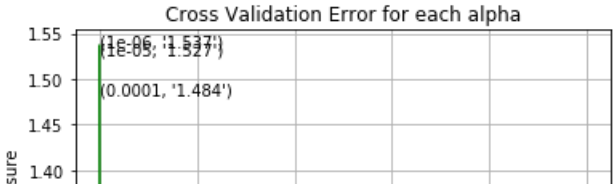
4.2.1. Hyper paramter tuning

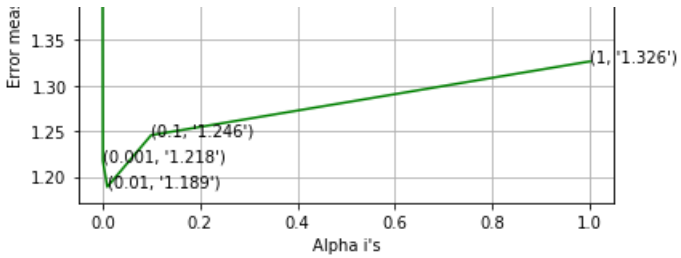
In [99]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_mo
del.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=N
one, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0,
power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/geometric-intuiti
on-1/
#-------------------------------



# find more about CalibratedClassifierCV here at http://scikit-learn.org/stable/modules/generated/sklea
```

```
rn.calibration.CalibratedClassifierCV.html
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, method='sigmoid', cv=3)
#
# some of the methods of CalibratedClassifierCV()
# fit(X, y[, sample_weight]) Fit the calibrated model
# get_params([deep]) Get parameters for this estimator.
# predict(X) Predict the target of new samples.
# predict_proba(X) Posterior probabilities of classification
#-------------------------------------
# video link:
#-------------------------------------

alpha = [10 ** x for x in range(-6, 1)]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = SGDClassifier(alpha=i, penalty='l2', loss='log', random_state=42)
    clf.fit(train_x_onehotCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_onehotCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_onehotCoding)
    cv_log_error_array.append(log_loss(cv_y, sig_clf_probs, labels=clf.classes_, eps=1e-15))
    print("Log Loss :",log_loss(cv_y, sig_clf_probs))

fig, ax = plt.subplots()
ax.plot(alpha, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[i],str(txt)), (alpha[i],cv_log_error_array[i]))
plt.grid()
plt.title("Cross Validation Error for each alpha")
plt.xlabel("Alpha i's")
plt.ylabel("Error measure")
plt.show()


best_alpha = np.argmin(cv_log_error_array)
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The train log loss is:",log_loss(y_train, pred
ict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(cv_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The cross validation log loss is:",log_loss(y_
cv, predict_y, labels=clf.classes_, eps=1e-15))
predict_y = sig_clf.predict_proba(test_x_onehotCoding)
print('For values of best alpha = ', alpha[best_alpha], "The test log loss is:",log_loss(y_test, predic
t_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-06
Log Loss : 1.536591020907485
for alpha = 1e-05
Log Loss : 1.5274646873956272
for alpha = 0.0001
Log Loss : 1.4838997135067387
for alpha = 0.001
Log Loss : 1.2175398616768762
for alpha = 0.01
Log Loss : 1.1888815510143445
for alpha = 0.1
Log Loss : 1.2455676860536395
for alpha = 1
Log Loss : 1.3264610661900667
```





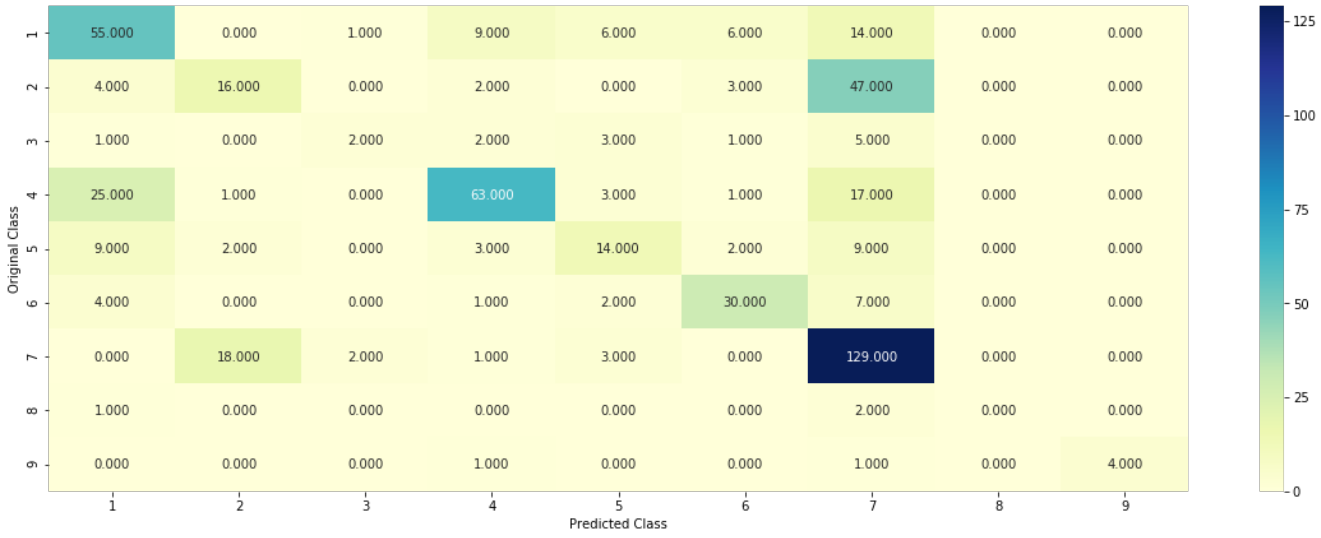

```
For values of best alpha =  0.01 The train log loss is: 0.7046715086493371
For values of best alpha =  0.01 The cross validation log loss is: 1.1888815510143445
For values of best alpha =  0.01 The test log loss is: 1.1815596595987967
```

4.3.2.2. Testing model with best hyper parameters

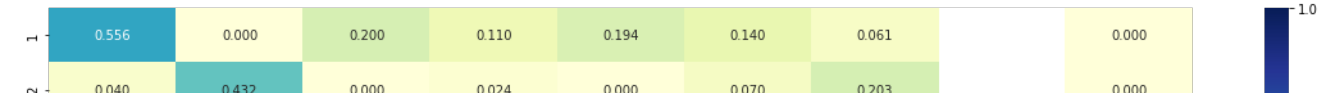
In [100]:

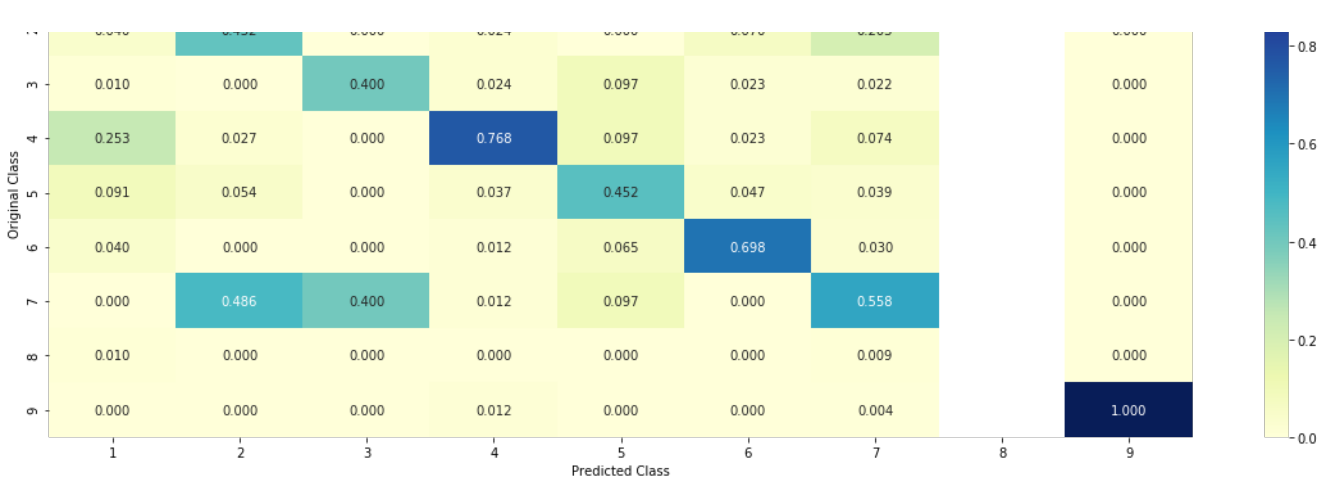
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, learning_rate='optimal', eta0=0.0, power_t=0.5,
# class_weight=None, warm_start=False, average=False, n_iter=None)

# some of methods
# fit(X, y[, coef_init, intercept_init, ...]) Fit linear model with Stochastic Gradient Descent.
# predict(X) Predict class labels for samples in X.

#-------------------------------
# video link:
#-------------------------------

clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_onehotCoding, cv_y, clf)
```

```
Log loss : 1.1888815510143445
Number of mis-classified points : 0.4116541353383459
-------------------- Confusion matrix --------------------
```




```
-------------------- Precision matrix (Column Sum=1) --------------------
```





---------- Recall matrix (Row sum=1) ----------

4.3.2.3. Feature Importance, Correctly Classified point

In [103]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
no_feature = 2000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_i
ndex]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 6
Predicted Class Probabilities: [[1.310e-02 1.130e-02 1.100e-03 1.020e-02 5.200e-02 9.008e-01 7.900e-03
  2.900e-03 6.000e-04]]
Actual Class : 6
--------------------------------------------------
1169 Text feature [african] present in test data point [True]
1367 Text feature [given family] present in test data point [True]
1790 Text feature [20c] present in test data point [True]
Out of the top  2000  features  3 are present in query point
```

4.3.2.4. Feature Importance, Incorrectly Classified point

In [104]:

```
test_point_index = 2
no_feature = 2000
predicted_cls = sig_clf.predict(test_x_onehotCoding[test_point_index])
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(test_x_onehotCoding[test_point_i
ndex]),4))
print("Actual Class :", test_y[test_point_index])
indices = np.argsort(-clf.coef_)[predicted_cls-1][:,:no_feature]
print("-"*50)
get_impfeature_names(indices[0], test_df['TEXT'].iloc[test_point_index],test_df['Gene'].iloc[test_point
_index],test_df['Variation'].iloc[test_point_index], no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.1149 0.103  0.0173 0.1032 0.047  0.0303 0.5727 0.0053 0.0064]]
Actual Class : 6
--------------------------------------------------
344 Text feature [likely sufficient] present in test data point [True]
449 Text feature [greater ability] present in test data point [True]
499 Text feature [transforming] present in test data point [True]
670 Text feature [downstream signaling] present in test data point [True]
771 Text feature [p85 heterodimer] present in test data point [True]
781 Text feature [within helical] present in test data point [True]
1011 Text feature [levels phosphorylated] present in test data point [True]
1012 Text feature [5a left] present in test data point [True]
1071 Text feature [effects downstream] present in test data point [True]
1464 Text feature [genetic aberrations] present in test data point [True]
1510 Text feature [thus appear] present in test data point [True]
1570 Text feature [activation event] present in test data point [True]
1632 Text feature [cancer lineage] present in test data point [True]
1645 Text feature [inhibitors determine] present in test data point [True]
1862 Text feature [present 30] present in test data point [True]
1879 Text feature [determine effects] present in test data point [True]
1977 Text feature [studies 26] present in test data point [True]
Out of the top  2000  features  17 are present in query point
```

Conclusion

1. EDA
   - Read training data 'Gene and Variation'
   - Read training data 'Text'
   - Preprocessing 'Text' data
   - Handling missing 'Text' data and merge all features
2. Split data into 3 parts in ratio 64:20:16
3. See the distribution of each class in train, cv and test data
4. Train the Random Model (worst/dump) to find the upperbound logloss
5. Univariant analysis on Gene and Variant
   - What type of feature is?
   - How many categories are present
   - How they are distributed
   - How to featurize the feature: responsecoding with Laplace smoothing and onehotencoding (CountVectorizer)
   - How good is this feature in prediction y?
6. Univariant Analysis on Text data
   - Extract the number of words from each class
   - Featurized the text with onhotencoding(CountVectorizer) witn unigram and bigram and responsecoding with Laplace smoothing
   - How good is this feature in predicting y?
7. Stacked all features
8. Finally , train model and plot and observe confusion matrix, precision and recall
   - Logistic Regr (LR) with class balance and get feature importance
   - Logistic Regr (LR) without class balance and get feature importance

In [3]:

```
from prettytable import PrettyTable
```

```
x = PrettyTable()
x.field_names = ["Feature", "Hyperparameter ","Model", "Train logloss", "CV logloss", "Test logloss", "
# of Misclassified pts"]

print('OHE -> OneHotEncoding\nRC-> ResponseCoding with Laplace smoothing')

feature = ['OHE','OHE']
param_ = [0.01,0.01]
model_ = ['LR (class balance)','LR (no class balance)']
train_loss = [0.70,0.70]
cv_loss = [1.18, 1.19]
test_loss = [1.18,1.18]
mis_class = [0.47,0.412]

for i in range(0,2):
    x.add_row([feature[i],param_[i],model_[i],train_loss[i],cv_loss[i],test_loss[i],mis_class[i]])

print(x)

print('\n\n1. From the observation above, LR(no class balance) are perform better than others.')
```

```
OHE -> OneHotEncoding
RC-> ResponseCoding with Laplace smoothing
+---------+-----------------+----------------------+--------------+------------+--------------+------
------------------+
| Feature | Hyperparameter  |         Model         | Train logloss | CV logloss | Test logloss | # of
Misclassified pts |
+---------+-----------------+----------------------+--------------+------------+--------------+------
------------------+
|   OHE   |       0.01      |   LR (class balance)  |      0.7      |    1.18    |     1.18     |
0.47          |
|   OHE   |       0.01      | LR (no class balance) |      0.7      |    1.19    |     1.18     |
0.412          |
+---------+-----------------+----------------------+--------------+------------+--------------+------
------------------+


1. From the observation above, LR(no class balance) are perform better than others.
```

In [ ]:

```
```