Personalized Cancer Diagonasis

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 [2]:

```python
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?
client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleuser
content.com&redirect_uri=urn%3aietf%3awg%3aoauth%3a2.0%3aoob&response_t
ype=code&scope=email%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdocs.t
est%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdrive%20https%3a%2f%2fw
ww.googleapis.com%2fauth%2fdrive.photos.readonly%20https%3a%2f%2fwww.go
ogleapis.com%2fauth%2fpeopleapi.readonly

Enter your authorization code:
··········
Mounted at /content/drive
```

In [0]:

```python
import pandas as pd
import matplotlib.pyplot as plt
import re
import time
import warnings
import numpy as np
import seaborn as sns
from collections import Counter, defaultdict
from nltk.corpus import stopwords
from sklearn.decomposition import TruncatedSVD
from sklearn.preprocessing import normalize
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.manifold import TSNE
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, normalized_mutual_info_score
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.calibration import CalibratedClassifierCV
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn.model_selection import train_test_split, GridSearchCV, StratifiedKFold

import math
from sklearn.ensemble import RandomForestClassifier
warnings.filterwarnings("ignore")

from mlxtend.classifier import StackingClassifier

from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
```

3.1. Reading Data

3.1.1. Reading Gene and Variation Data

In [7]:

```
data_variants = pd.read_csv('/content/drive/My Drive/cancer study/training_variants')
print('Number of data points : ', data_variants.shape[0])
print('Number of features : ', data_variants.shape[1])
print('Features : ', data_variants.columns.values)
data_variants.head()
```

```
Number of data points :  3321
Number of features :  4
Features :  ['ID' 'Gene' 'Variation' 'Class']
```

Out[7]:

	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 [8]:

```
# note the seprator in this file
data_text =pd.read_csv("/content/drive/My Drive/cancer study/training_t
ext",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[8]:

	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 [10]:

```
import nltk
nltk.download('stopwords')
# 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 t
he data
            if not word in stop_words:
                string += word + " "

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

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

In [11]:

```
# Text processing stage.
start_time = time.clock()
for index, row in 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_tim
e, "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 : 28.178103 seconds
```

In [12]:

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

Out[12]:

	ID	Gene	Variation	Class	TEXT

	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 [14]:

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

Out[14]:

	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 [0]:

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

In [16]:

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

Out[16]:

	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 [21]:

```
result.Gene = result.Gene.str.replace('\s+', '_')
result.Variation = result.Variation.str.replace('\s+', '_')
y_true = result[['Class']]
x_true = result.drop(['Class'], axis=1)


print("Feature columns in dataset: ")
print(x_true.head())
print()
print("Target columns in dataset: ")
print(y_true.head())
```

```
Feature columns in dataset: 
   ID  ...                                               TEXT
0   0  ...  cyclin dependent kinases cdks regulate variety...
1   1  ...  abstract background non small cell lung cancer...
2   2  ...  abstract background non small cell lung cancer...
3   3  ...  recent evidence demonstrated acquired uniparen...
4   4  ...  oncogenic mutations monomeric casitas b lineag...

[5 rows x 4 columns]

Target columns in dataset: 
   Class
0      1
1      2
2      2
3      3
4      4
```

In [0]:

```
# Split the data into test and train by maintaining same distribution o
f output varaible 'y_true' [stratify=y_true]
```

```
x_train, x_test, y_train, y_test = train_test_split(x_true, y_true, str
atify=y_true, test_size=0.2)

# Split the train data into train and cross validation by maintaining s
ame distribution of output varaible 'y_train' [stratify=y_train]
x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, strat
ify=y_train, test_size=0.2)
```

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

In [23]:

```
print('Number of data points in train data:', x_train.shape[0])
print('Number of data points in test data:', x_test.shape[0])
print('Number of data points in cross validation data:', x_cv.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 [33]:

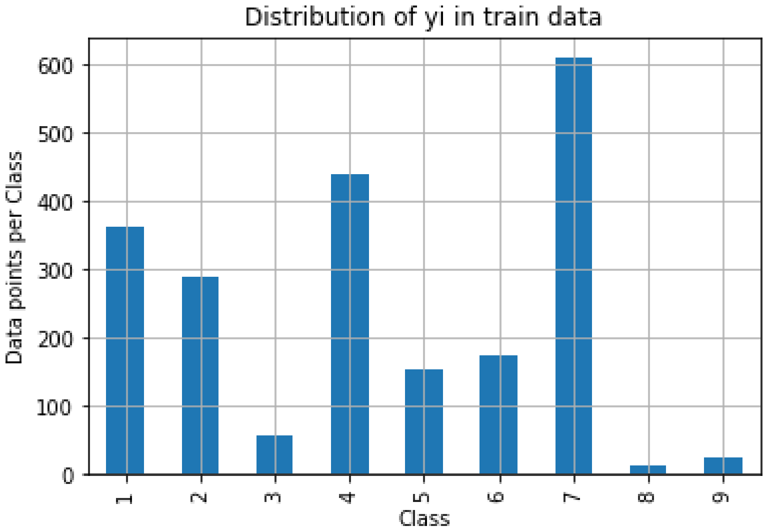
```
pip install pandas -U
```

```
Requirement already up-to-date: pandas in /usr/local/lib/python3.6/dist
-packages (0.25.3)
Requirement already satisfied, skipping upgrade: numpy>=1.13.3 in /usr/
local/lib/python3.6/dist-packages (from pandas) (1.17.4)
Requirement already satisfied, skipping upgrade: pytz>=2017.2 in /usr/l
ocal/lib/python3.6/dist-packages (from pandas) (2018.9)
Requirement already satisfied, skipping upgrade: python-dateutil>=2.6.1
in /usr/local/lib/python3.6/dist-packages (from pandas) (2.6.1)
Requirement already satisfied, skipping upgrade: six>=1.5 in /usr/loca
l/lib/python3.6/dist-packages (from python-dateutil>=2.6.1->pandas) (1.
12.0)
```

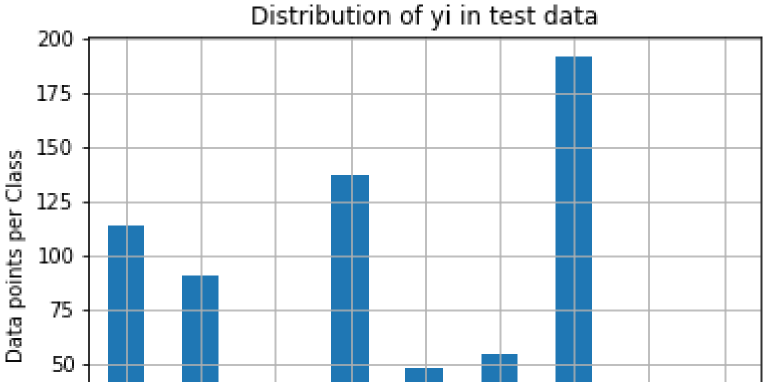
In [35]:

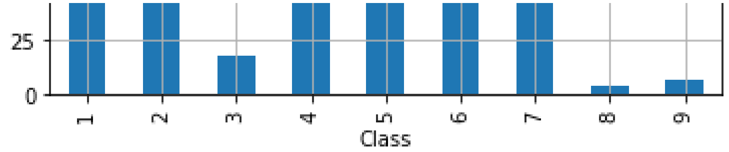
```
def plot_distribution(class_distribution,title,xlabel,ylabel):
    class_distribution.plot(kind='bar')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.title(title)
    plt.grid()
    plt.show()


# it returns a dict, keys as class labels and values as the number of d
ata points in that class
train_class_distribution = y_train['Class'].value_counts().sort_index()
test_class_distribution = y_test['Class'].value_counts().sort_index()
cv_class_distribution = y_cv['Class'].value_counts().sort_index()

plot_distribution(train_class_distribution,
                 'Distribution of yi in train data',
                 'Class',
                 'Data points per Class')

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/num
py.argsort.html
# -(train_class_distribution.values): the minus sign will give us in de
creasing 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_distri
bution.values[i],
          '(', np.round((train_class_distribution.values[i]/x_train.sha
pe[0]*100), 3), '%)')

print('-'*80)



plot_distribution(test_class_distribution,
                 'Distribution of yi in test data',
                 'Class',
                 'Data points per Class')

# ref: argsort https://docs.scipy.org/doc/numpy/reference/generated/num
```

```
py.argsort.html
# -(test_class_distribution.values): the minus sign will give us in dec
reasing 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_distrib
ution.values[i],
          '(', np.round((test_class_distribution.values[i]/x_test.shape
[0]*100), 3), '%)')

print('-'*80)



plot_distribution(cv_class_distribution,
                 'Distribution of yi in cross validation data',
                 'Class',
                 'Data points per Class')

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

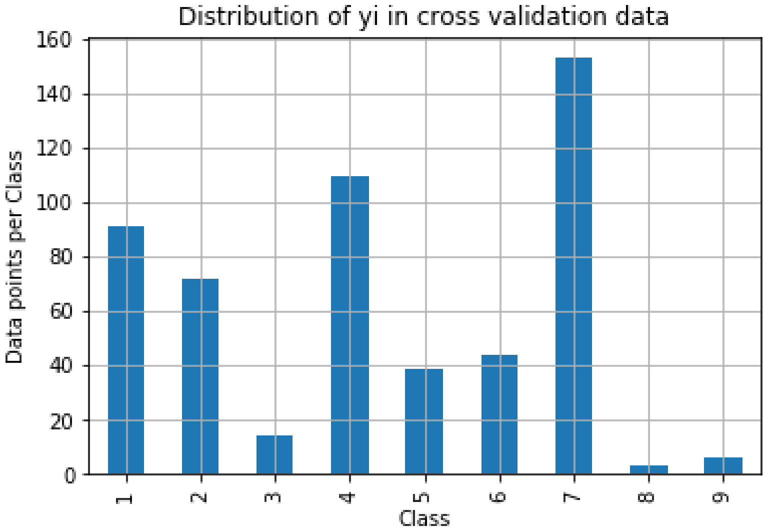


```
Number of data points in class 7 : 609 ( 28.672 %)
Number of data points in class 4 : 439 ( 20.669 %)
Number of data points in class 1 : 363 ( 17.09 %)
Number of data points in class 2 : 289 ( 13.606 %)
Number of data points in class 6 : 176 ( 8.286 %)
Number of data points in class 5 : 155 ( 7.298 %)
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 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 '9' class probabilites randomly such that they sum to 1.

In [0]:

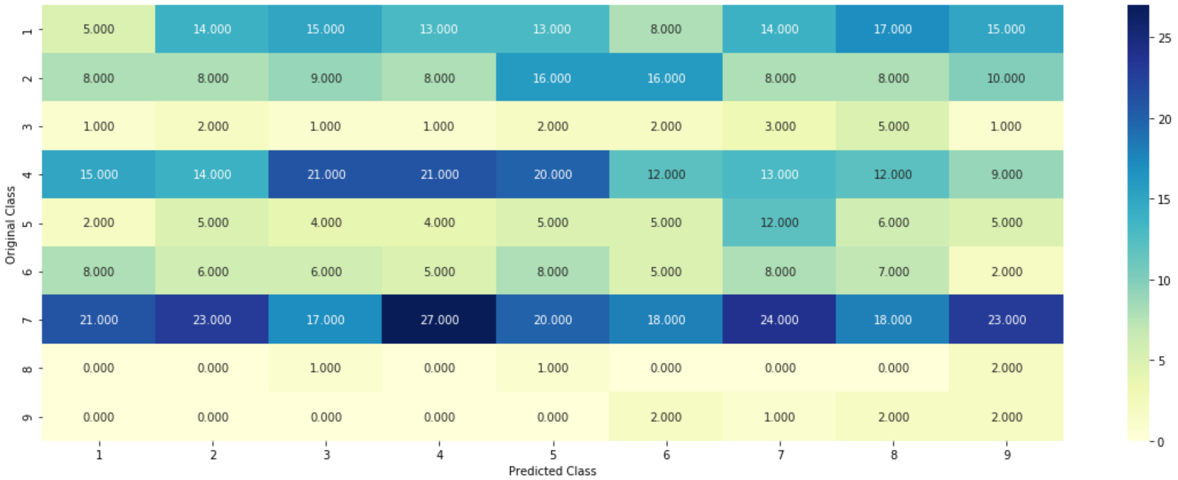
```
def plot_matrix(matrix,labels):
    plt.figure(figsize=(20,7))
    sns.heatmap(matrix, annot=True, cmap="YlGnBu", fmt=".3f", xticklabe
ls=labels, yticklabels=labels)
    plt.xlabel('Predicted Class')
    plt.ylabel('Original Class')
    plt.show()

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

    recall_table =(((cm.T)/(cm.sum(axis=1))).T)
    # How did we calculateed recall_table :
    # divide each element of the confusion matrix with the sum of eleme
nts 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 correspo
nds 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

    precision_table =(cm/cm.sum(axis=0))
    # How did we calculateed precision_table :
    # divide each element of the confusion matrix with the sum of eleme
nts in that row
    # C = [[1, 2],
    #      [3, 4]]
    # C.sum(axis = 0)  axis=0 corresonds to columns and axis=1 correspo
nds 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]
    print()
    print("-"*20, "Confusion matrix", "-"*20)
    plot_matrix(cm,labels)

    print("-"*20, "Precision matrix (Columm Sum=1)", "-"*20)
    plot_matrix(precision_table,labels)

    print("-"*20, "Recall matrix (Row sum=1)", "-"*20)
    plot_matrix(recall_table,labels)
```

In [37]:

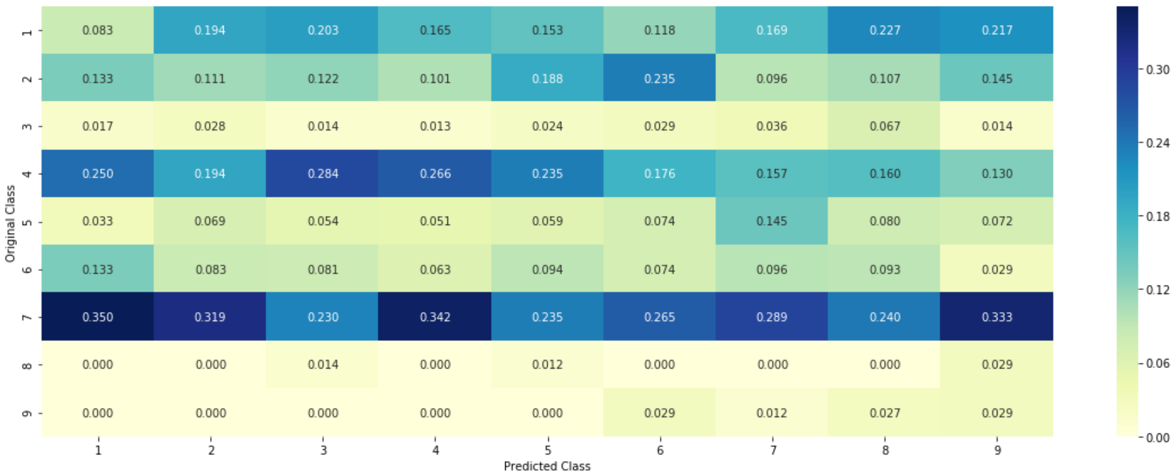
```
# We need to generate 9 numbers and the sum of numbers should be 1
# one solution is to genarate 9 numbers and divide each of the numbers
 by their sum
# ref: https://stackoverflow.com/a/18662466/4084039
test_data_len = x_test.shape[0]
cv_data_len = x_cv.shape[0]

# 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_p
redicted_y, eps=1e-15))

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

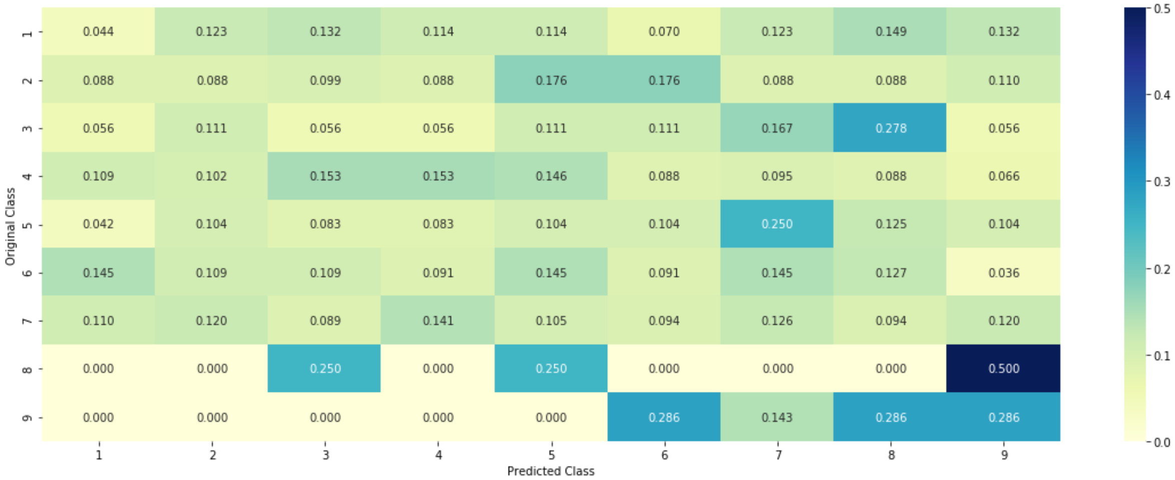
```
Log loss on Cross Validation Data using Random Model 2.47148139847889
3
Log loss on Test Data using Random Model 2.531700546464972

-------------------- Confusion matrix --------------------
```



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



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



3.3 Univariate Analysis

In [0]:

```
# code for response coding with Laplace smoothing.
# alpha : used for laplace smoothing
# feature: ['gene', 'variation']
# df: ['x_train', 'x_test', 'x_cv']
# algorithm
# ----------
# Consider all unique values and the number of occurances of given feat
ure in train data dataframe
# build a vector (1*9) , the first element = (number of times it occure
d in class1 + 10*alpha / number of time it occurred in total data+90*al
pha)
# gv_dict is like a look up table, for every gene it store a (1*9) repr
esentation 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_f
ea'
# 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 = x_train[feature].value_counts()

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

    # denominator will contain the number of time that particular featu
re occured in whole data
    for i, denominator in value_count.items():
        # vec will contain (p(yi==1/Gi) probability of gene/variation b
elongs 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['Ge
ne']=='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 = x_train.loc[(y_train['Class']==k) & (x_train[feat
ure]==i)]
```

```
            # cls_cnt.shape[0](numerator) will contain the number of time that particular feature occured 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)
    #     {'BRCA1': [0.20075757575757575, 0.03787878787878788, 0.068181818181818177, 0.13636363636363635, 0.25, 0.19318181818181818, 0.03787878787878788, 0.03787878787878788, 0.03787878787878788],
    #      'TP53': [0.32142857142857145, 0.061224489795918366, 0.061224489795918366, 0.27040816326530615, 0.061224489795918366, 0.066326530612244902, 0.051020408163265307, 0.051020408163265307, 0.056122448979591837],
    #      'EGFR': [0.056818181818181816, 0.21590909090909091, 0.0625, 0.068181818181818177, 0.068181818181818177, 0.0625, 0.34659090909090912, 0.0625, 0.056818181818181816],
    #      'BRCA2': [0.13333333333333333, 0.060606060606060608, 0.060606060606060608, 0.078787878787878782, 0.13939393939393939, 0.34545454545454546, 0.060606060606060608, 0.060606060606060608, 0.060606060606060608],
    #      'PTEN': [0.069182389937106917, 0.062893081761006289, 0.069182389937106917, 0.46540880503144655, 0.075471698113207544, 0.062893081761006289, 0.069182389937106917, 0.062893081761006289, 0.062893081761006289],
    #      'KIT': [0.066225165562913912, 0.25165562913907286, 0.072847682119205295, 0.072847682119205295, 0.066225165562913912, 0.066225165562913912, 0.27152317880794702, 0.066225165562913912, 0.066225165562913912],
    #      'BRAF': [0.066666666666666666, 0.17999999999999999, 0.073333333333333334, 0.073333333333333334, 0.093333333333333338, 0.080000000000000002, 0.29999999999999999, 0.066666666666666666, 0.066666666666666666],
    #       ...
    #     }
    gv_dict = get_gv_fea_dict(alpha, feature, df)
    # value_count is similar in get_gv_fea_dict
    value_count = x_train[feature].value_counts()

    # gv_fea: Gene_variation feature, it will contain the feature for e
ach 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 then 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_fe
a
    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 [39]:
```
unique_genes = x_train['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 : 233
BRCA1     176
TP53      102
BRCA2      85
PTEN       77
EGFR       77
BRAF       61
KIT        58
ALK        42
ERBB2      39
PIK3CA     37
Name: Gene, dtype: int64
```

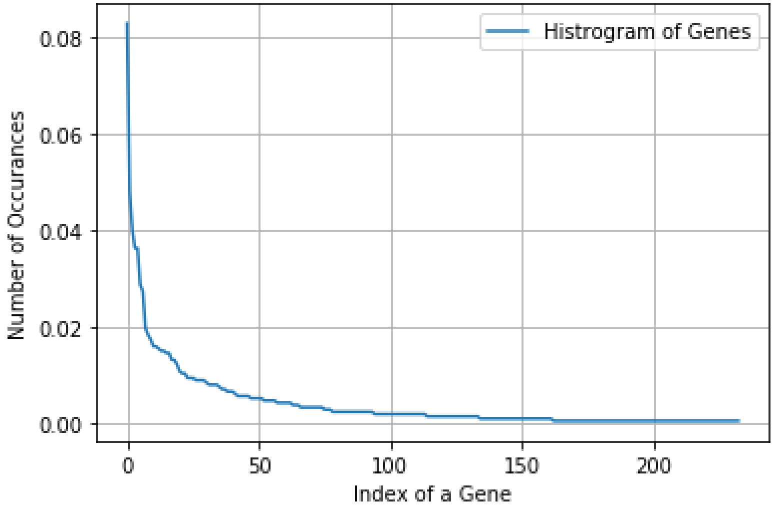
In [40]:

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

```
Ans: There are 233 different categories of genes in the train data, and
they are distibuted as follows
```

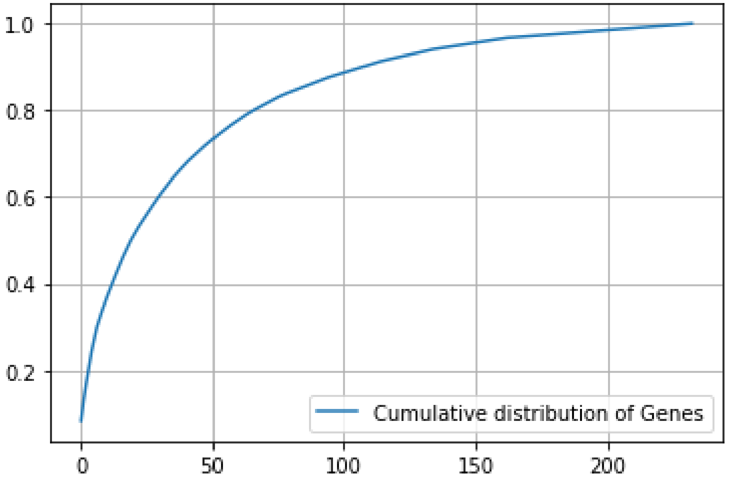
In [41]:

```
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 [42]:

```
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 [0]:

```
#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", x_train))
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))
# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [44]:

```
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 feature: (2124, 9)
```

In [0]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = TfidfVectorizer()
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(x_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(x_test['Gen
e'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(x_cv['Gene'])
```

In [46]:
```
train_gene_feature_onehotCoding
```

Out[46]:
```
<2124x232 sparse matrix of type '<class 'numpy.float64'>'
	with 2124 stored elements in Compressed Sparse Row format>
```

In [47]:
```
print("train_gene_feature_onehotCoding is converted feature using one-h
ot encoding method. The shape of gene feature:",
      train_gene_feature_onehotCoding.shape)
```

```
train_gene_feature_onehotCoding is converted feature using one-hot enco
ding method. The shape of gene feature: (2124, 232)
```

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

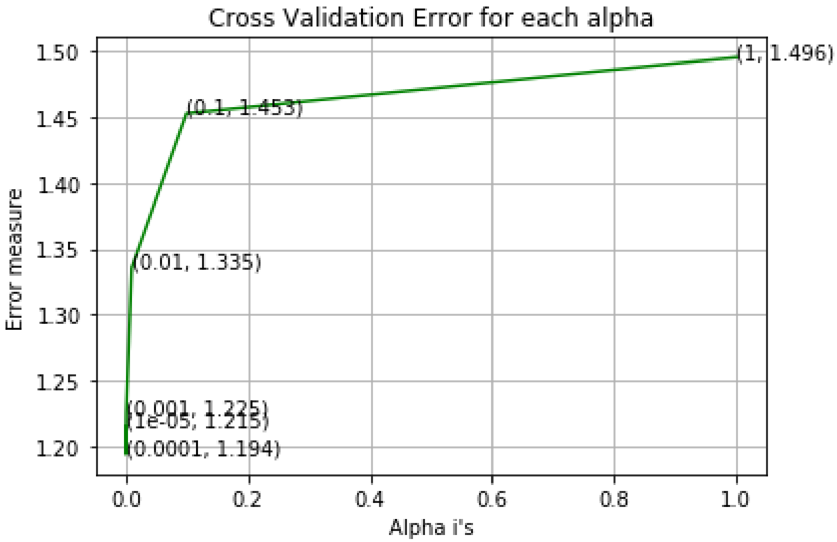
There are many ways to estimate how good a feature is, in predicting y_i. One of the good methods is to build a proper ML model using just this feature. In this case, we will build a logistic regression model using only Gene feature (one hot encoded) to predict y_i.

In [48]:
```
alpha = [10 ** x for x in range(-5, 1)] # hyperparam for SGD classifie
r.

# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_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 S
tochastic 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.clas
ses_, 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_arra
y[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, predict_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, predict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.21509265096772
For values of alpha =  0.0001 The log loss is: 1.1935588304691498
For values of alpha =  0.001 The log loss is: 1.225131040452187
For values of alpha =  0.01 The log loss is: 1.3354750081327889
For values of alpha =  0.1 The log loss is: 1.4529597747973626
For values of alpha =  1 The log loss is: 1.495568541586665
```



```
For values of best alpha =  0.0001 The train log loss is: 1.013773862
2711105
For values of best alpha =  0.0001 The cross validation log loss is:
1.1935588304691498
For values of best alpha =  0.0001 The test log loss is: 1.1674872485
05434
```

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 [49]:

```
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=x_test[x_test['Gene'].isin(list(set(x_train['Gene'])))].s
hape[0]
cv_coverage=x_cv[x_cv['Gene'].isin(list(set(x_train['Gene'])))].shape[0
]
```

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

```
Q6. How many data points in Test and CV datasets are covered by the  23
3  genes in train dataset?
Ans
1. In test data 645 out of 665 : 96.99248120300751
2. In cross validation data 512 out of  532 : 96.2406015037594
```

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 [50]:

```
unique_variations = x_train['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 : 1926
Truncating_Mutations    61
Deletion                50
Amplification           47
Fusions                 24
G12V                     3
E17K                     2
T73I                     2
P34R                     2
Overexpression           2
EWSR1-ETV1_Fusion        2
Name: Variation, dtype: int64
```

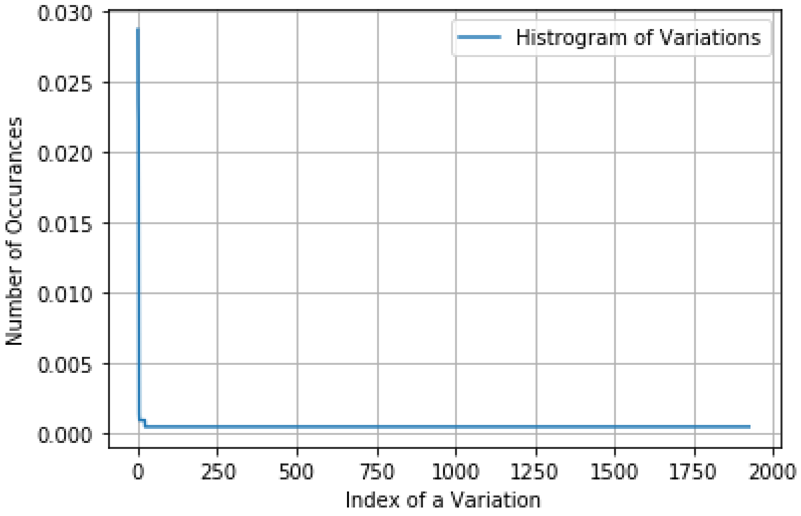
In [51]:

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

```
Ans: There are 1926 different categories of variations in the train data, and they are distibuted as follows
```

In [52]:

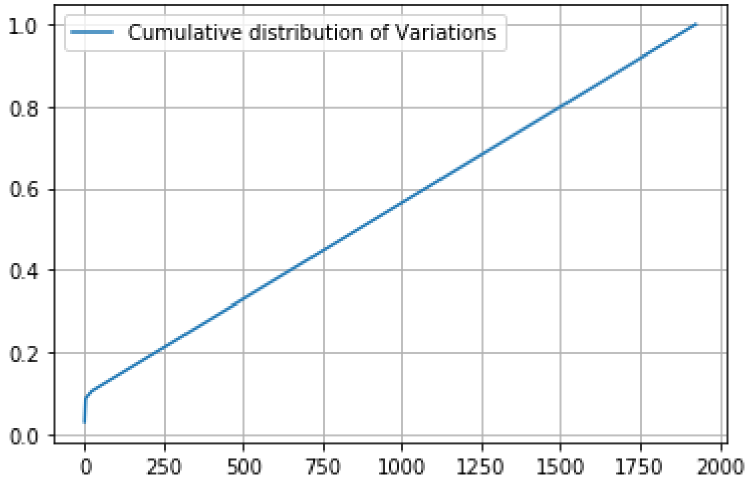
```
s = sum(unique_variations.values);
h = unique_variations.values/s;
plt.plot(h, label="Histrogram of Variations")
plt.xlabel('Index of a Variation')
plt.ylabel('Number of Occurances')
plt.legend()
plt.grid()
plt.show()
```



In [53]:

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

```
[0.0287194  0.05225989 0.07438795 ... 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 [0]:
```
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha,
 "Variation", x_train))

# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha,
"Variation", x_test))

# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "V
ariation", x_cv))
```

In [55]:

```
print("train_variation_feature_responseCoding is a converted feature us
ing 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 [0]:

```
# one-hot encoding of variation feature.
variation_vectorizer = TfidfVectorizer()
train_variation_feature_onehotCoding = variation_vectorizer.fit_transfo
rm(x_train['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(x_
test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv
['Variation'])
```

In [57]:

```
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 on
ne-hot encoding method. The shape of Variation feature: (2124, 1956)
```

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

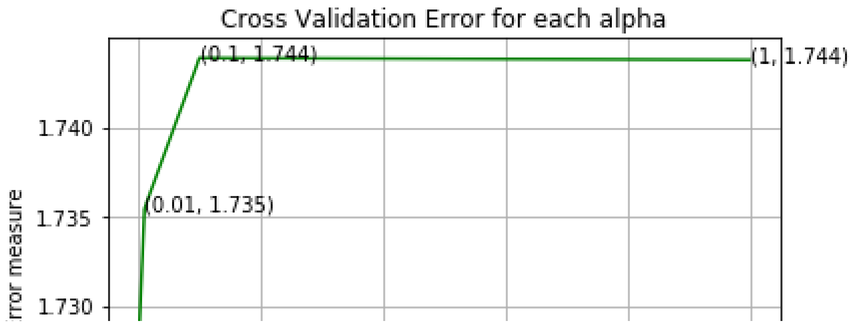
Let's build a model just like the earlier!

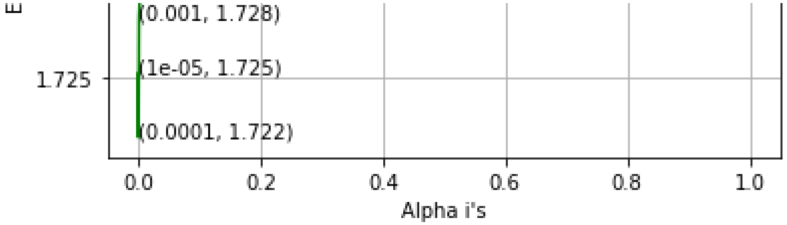
In [58]:

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

# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/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, predict_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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.725232987650247
For values of alpha =  0.0001 The log loss is: 1.7216475393608937
For values of alpha =  0.001 The log loss is: 1.7283142105934945
For values of alpha =  0.01 The log loss is: 1.7354313714766585
For values of alpha =  0.1 The log loss is: 1.7438493968177016
For values of alpha =  1 The log loss is: 1.7437454752985284
```





```
For values of best alpha =  0.0001 The train log loss is: 0.76440597003
83514
For values of best alpha =  0.0001 The cross validation log loss is: 1.
7216475393608937
For values of best alpha =  0.0001 The test log loss is: 1.709072029723
3053
```

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 [59]:

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

```
Q12. How many data points are covered by total  1926  genes in test and
cross validation data sets?
Ans
1. In test data 68 out of 665 : 10.225563909774436
2. In cross validation data 54 out of  532 : 10.150375939849624
```

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 [0]:

```
# 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 [0]:

```
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 )/(t
otal_dict.get(word,0)+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_pr
ob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [62]:
```
x_train['TEXT'].head()
```

Out[62]:
```
2015    background mek1 mutations melanoma confer resi...
2002    mek1 mek2 related protein kinases participate ...
727     mutational hotspots indicate selective pressur...
1504    abstract somatically acquired epigenetic chang...
2350    philadelphia chromosome like acute lymphoblast...
Name: TEXT, dtype: object
```

In [0]:
```
def top_tfidf_feats(row, features, top_n=25):
    ''' Get top n tfidf values in row and return them with their corres
ponding feature names.'''
    topn_ids = np.argsort(row)[::-1][:top_n]
    top_feats = [(features[i], row[i]) for i in topn_ids]
    df = pd.DataFrame(top_feats)
    df.columns = ['feature', 'tfidf']
    return df

def top_mean_feats(Xtr, features, min_tfidf=0.1, grp_ids=None, top_n=25
):
    ''' Return the top n features that on average are most important am
ongst documents in rows
        indentified by indices in grp_ids. '''
    if grp_ids:
        D = Xtr[grp_ids].toarray()
    else:
        D = Xtr.toarray()

    D[D < min_tfidf] = 0
    tfidf_means = np.mean(D, axis=0)
    return top_tfidf_feats(tfidf_means, features, top_n)
```

In [0]:
```
# building a CountVectorizer with all the words that occured minimum 3
 times in train data
text_vectorizer = TfidfVectorizer(min_df=3)
train_text_feature_onehotCoding = text_vectorizer.fit_transform(x_train
['TEXT'])

# getting top 1000 feature names (words)
```

```
train_text_features = top_mean_feats(train_text_feature_onehotCoding,
                                                        text_vectorizer.get_fea
ture_names(),
                                                        top_n=1000)['feature'].
tolist()
```

In [65]:
```
# 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
train_text_fea_counts
```

Out[65]:
```
array([9.78120387, 9.34827169, 0.0353608 , ..., 0.01130901, 0.03074788,
       0.07961424])
```

In [66]:
```
# zip(list(text_features),text_fea_counts) will zip a word with its num
ber of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_count
s))


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

```
Total number of unique words in train data : 1000
```

In [0]:
```
dict_list = []
# dict_list =[] contains 9 dictoinaries each corresponds to a class
for i in range(1,10):
    cls_text = x_train[y_train['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 = extract_dictionary_paddle(x_train)


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[i]+90))
    confuse_array.append(ratios)
confuse_array = np.array(confuse_array)
```

In [0]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)
```

In [0]:

```
# 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_responseCoding.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(axis=1)).T
```

In [0]:

```
# 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(x_test['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(x_cv['TEXT'])
# don't forget to normalize every feature
cv_text_feature_onehotCoding = normalize(cv_text_feature_onehotCoding, axis=0)
```

In [0]:

```
#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 [72]:

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

```
Counter({0.027348412448247397: 30, 0.027313539435983l2: 22, 0.070753151
53685217: 16, 0.034531123150474366: 9, 0.028323017655552007: 8, 0.36616
02984765746: 7, 0.037525226605373614: 7, 0.245072211469939l: 6, 0.02923
6315642720896: 5, 0.06906224630094873: 4, 0.04475811086358627: 4, 0.019
365134684061074: 4, 0.07123122491754474: 3, 0.05462707887196624: 3, 0.0
41969078763312885: 3, 0.02827931412099005: 3, 0.026566288473777108: 3,
0.020634680220092967: 3, 0.01658069940315386: 3, 0.015259417207678351:
3, 0.01201711235883088: 3, 0.01130901408163787З: 3, 0.00691053246424004
45: 3, 0.16880341786467004: 2, 0.10359336945142311: 2, 0.10178112673474
085: 2, 0.08992515476613207: 2, 0.06632913471494398: 2, 0.0649033141196
7093: 2, 0.056646035311104015: 2, 0.05160684630175373: 2, 0.04974209820
9461585: 2, 0.030915557196133522: 2, 0.02874005387736772: 2, 0.02746947
7846428614: 2, 0.026485700616552078: 2, 0.024992317222838865: 2, 0.0249
3969729979688: 2, 0.022618028163275745: 2, 0.02026918444589862: 2, 0.01
983401822758016: 2, 0.018375749826926642: 2, 0.013396675363969407: 2,
0.012496158611419433: 2, 0.010283305142038878: 2, 33.15213356305331: 1,
18.800329020923026: 1, 18.35210009570497: 1, 14.777069870430353: 1, 13.
10556349353904: 1, 9.781203873633551: 1, 9.348271689735814: 1, 6.714770
828166559: 1, 5.707055126454826: 1, 5.0804101341224674: 1, 4.7015438238
898986: 1, 4.460691275234803: 1, 4.357157533957519: 1, 4.28062991722362
6: 1, 4.071546428202246: 1, 3.9466564628244907: 1, 3.834097899372683:
1, 3.795210531373411: 1, 3.691555632740379: 1, 3.555997667073773: 1, 3.
2254615246433054: 1, 2.967446699927231: 1, 2.827332121929487: 1, 2.7666
33117070814: 1, 2.695794326169332: 1, 2.435549628106116: 1, 2.323695117
351719: 1, 2.1002198888613046: 1, 2.0365897745734705: 1, 2.005182945030
265: 1, 2.0015582465593167: 1, 1.964752214387699: 1, 1.914061956184876
6: 1, 1.8229812497110482: 1, 1.738631651824602: 1, 1.6792069361149524:
1, 1.615697494674298: 1, 1.5615691201547788: 1, 1.532894581529564: 1,
1.4699777765636042: 1, 1.4346455539018335: 1, 1.408549771610778: 1, 1.4
002171648872521: 1, 1.2571263385983653: 1, 1.2320415923463675: 1, 1.200
```

```
2795809646847: 1, 1.0888973509338777: 1, 1.0296010458464397: 1, 1.02749
90290362358: 1, 1.026949698767862: 1, 1.0017004957240376: 1, 0.99467727
19569529: 1, 0.9892919430208121: 1, 0.9789499350623859: 1, 0.9511883365
876501: 1, 0.8967362831219837: 1, 0.8775497493510961: 1, 0.875081585685
7693: 1, 0.871313090546159: 1, 0.8703812986334899: 1, 0.845199105856910
9: 1, 0.8351048761364598: 1, 0.8159662655943711: 1, 0.8007353643741917:
1, 0.7813759358595486: 1, 0.7649493279588543: 1, 0.7487917928425678: 1,
0.7463169999263384: 1, 0.7442916163103092: 1, 0.7324468297336642: 1, 0.
7313145880434785: 1, 0.7179835190631199: 1, 0.7129088272154156: 1, 0.68
45931445086815: 1, 0.6386196416905288: 1, 0.5982267729426565: 1, 0.5828
875821397915: 1, 0.5538755945416597: 1, 0.5457388720320702: 1, 0.530759
5308026725: 1, 0.5288613360437896: 1, 0.5062130416775504: 1, 0.50489361
14124587: 1, 0.48225266234873826: 1, 0.4821030332573601: 1, 0.471074567
7028782: 1, 0.46731394221346073: 1, 0.4648114115919422: 1, 0.4619037519
265363: 1, 0.46098741586532566: 1, 0.45725488879795606: 1, 0.4554872342
9731486: 1, 0.45426362861020997: 1, 0.4526317437443857: 1, 0.4452856375
786043: 1, 0.44401991619138304: 1, 0.43946657515912263: 1, 0.4384663003
1899236: 1, 0.43660262350277235: 1, 0.4336022902335166: 1, 0.4303090610
4779765: 1, 0.4275603794836571: 1, 0.4236719545716231: 1, 0.42185306821
974217: 1, 0.414476360052037 3: 1, 0.39957111783018134: 1, 0.39754185137
46027: 1, 0.39597211941328897: 1, 0.39493747099035503: 1, 0.39452128940
448505: 1, 0.39365206881057446: 1, 0.3924856361260656: 1, 0.39055277210
86557: 1, 0.38766095619461044: 1, 0.3808615308987661: 1, 0.378696495463
5649: 1, 0.37836656724900714: 1, 0.3663019668294163: 1, 0.3656044461339
9234: 1, 0.36516764037443283: 1, 0.35662625436289613: 1, 0.350406776263
59845: 1, 0.3496308481010539: 1, 0.3432296375491955: 1, 0.3419213391117
7947: 1, 0.3418059068945052: 1, 0.33682842542103913: 1, 0.3350818991900
452: 1, 0.33407776173461473: 1, 0.33360103779245787: 1, 0.3320604826704
431: 1, 0.33108707397427534: 1, 0.3274784343098888: 1, 0.32509555847439
53: 1, 0.3246235026196641: 1, 0.32353682184263804: 1, 0.322245777782994
96: 1, 0.32159947206992984: 1, 0.31759675620877753: 1, 0.31422782633875
035: 1, 0.3104531949577192: 1, 0.3094190023311168: 1, 0.302540471878026
5: 1, 0.3017518210668883: 1, 0.298126957071823: 1, 0.29745823646 4815:
1, 0.29343464944669984: 1, 0.29260627278409224: 1, 0.28706886747448745:
1, 0.28202763279337345: 1, 0.27900854577417045: 1, 0.278345623465707:
1, 0.2781173889169328: 1, 0.277340261260 1492: 1, 0.27126906321 27033: 1,
0.2704813640308251: 1, 0.2694565436056447: 1, 0.26570125595725247: 1,
0.264676312668 0763: 1, 0.26220965991064127: 1, 0.2601336883602317: 1,
0.25786974941013735: 1, 0.25668404811489803: 1, 0.25456567460695295: 1,
0.25143157389237564: 1, 0.251153613898335: 1, 0.2509381763986581: 1, 0.
2502803780938275: 1, 0.24935987623463773: 1, 0.24729331633300514: 1, 0.
24711805373913395: 1, 0.24701987321543983: 1, 0.2469331214596581: 1, 0.
24635764587900108: 1, 0.24601761656274398: 1, 0.24211370922871667: 1,
0.2404167004697405: 1, 0.23973707971912583: 1, 0.2351152064360039: 1,
0.23439913273240331: 1, 0.22986688362617322: 1, 0.22770916799590257: 1,
0.22475306348263327: 1, 0.22313233167861704: 1, 0.2228728880170569: 1,
0.22139100603723932: 1, 0.22032304225104102: 1, 0.2190467459346997: 1,
0.21501155894966908: 1, 0.21493804667240912: 1, 0.21444622334185445: 1,
0.21425941151851197: 1, 0.20866593876957323: 1, 0.20778047279715306: 1,
0.20745639041017963: 1, 0.20550649715355465: 1, 0.20496298157781573: 1,
0.20434491241258038: 1, 0.20383420217881357: 1, 0.20366708223055027: 1,
0.20302016628876107: 1, 0.2027458596927876: 1, 0.2016824355357336: 1,
0.20100871098884662: 1, 0.1987490455393472: 1, 0.1970899257014977: 1,
0.19632235598989997: 1, 0.1926857217377188: 1, 0.19242691658020245: 1,
0.1901367119475754: 1, 0.19004130153235482: 1, 0.1866566673689146: 1,
0.18496783491076693: 1, 0.18287623274421944: 1, 0.18243047180090077: 1,
0.18211442173833603: 1, 0.18020444855827492: 1, 0.1793177920658819: 1,
0.17873474319256766: 1, 0.17821744177737225: 1, 0.17548223770811489: 1,
0.17446040556833545: 1, 0.17405756193949853: 1, 0.17352605865459705: 1,
0.17324121759830724: 1, 0.17286612871997264: 1, 0.17043837590869723: 1,
0.17006999045841534: 1, 0.16993810593331204: 1, 0.16919756690341042: 1,
0.16736520969679092: 1, 0.16624600389983168: 1, 0.16621144658128809: 1,
0.16284039677825085: 1, 0.16236282450954376: 1, 0.16096472144218094: 1,
0.16009529331000244: 1, 0.16007252911651773: 1, 0.15896702793243825: 1,
0.15885000501840546: 1, 0.15795514037264777: 1, 0.15658660197344357: 1,
0.15512526731827278: 1, 0.15464965593560065: 1, 0.15441785733528865: 1,
0.15366597880984517: 1, 0.15351478932749288: 1, 0.15193065501620864: 1,
0.15163252316496614: 1, 0.15069059503530946: 1, 0.14806929244790806: 1,
0.14779428611165926: 1, 0.14617557772392265: 1, 0.14536559071855892: 1,
0.14305643122229011: 1, 0.14107267574758858: 1, 0.13961745799585218: 1,
0.1395896230755626: 1, 0.13940912588636753: 1, 0.13898809680578053: 1,
0.1365161765633385: 1, 0.1359968103123439: 1, 0.13290806734505112: 1,
0.13256048184768898: 1, 0.13195725333416125: 1, 0.1316542864023354: 1,
0.13153888902690364: 1, 0.13011799015153783: 1, 0.12914207250556553: 1,
0.12833801195207828: 1, 0.12823576155753644: 1, 0.12740681876482868: 1,
0.12711156645585847: 1, 0.12696117254180134: 1, 0.12688282910228063: 1,
0.12615593333451816: 1, 0.12590791537554247: 1, 0.12589651234643492: 1,
0.1257086664563: 1, 0.1243015847731725: 1, 0.12281926861272781: 1, 0.12
```

```
128299264349006: 1, 0.12114874946604202: 1, 0.12065637566607268: 1, 0.1
206069658861622: 1, 0.12030731697742982: 1, 0.11996736405165002: 1, 0.1
1853169524872022: 1, 0.11852232137984617: 1, 0.11849442909528361: 1, 0.
11713497955316353: 1, 0.11700848645825244: 1, 0.11678217748346252: 1,
0.11375417619240798: 1, 0.11336478357101645: 1, 0.11305548055495081: 1,
0.1129900150197311: 1, 0.11271458911910907: 1, 0.11214960291389296: 1,
0.11109984803199267: 1, 0.1107865191738894: 1, 0.11058515462080609: 1,
0.10938505207206992: 1, 0.10933109314468623: 1, 0.10861275609646981: 1,
0.1086109347142838 7: 1, 0.1083741120271672: 1, 0.10819294485751217: 1,
0.10782517499429789: 1, 0.10776774709752711: 1, 0.10702299692834309: 1,
0.10700695223099407: 1, 0.1065320994860266: 1, 0.10641134958778589: 1,
0.10581044882878135: 1, 0.10559634705395302: 1, 0.1049733016252095: 1,
0.10448808437565468: 1, 0.10426495545922909: 1, 0.10354365968654032: 1,
0.10346044523007618: 1, 0.10330034588416369: 1, 0.10277790017591816: 1,
0.10271843750118592: 1, 0.10233972446685913: 1, 0.10224379583429065: 1,
0.10216492923223683: 1, 0.10203438604614036: 1, 0.10172766417867432: 1,
0.1016643166593028: 1, 0.1012326003152948 7: 1, 0.100369768579148: 1, 0.
09982388231902069: 1, 0.09967850136828979: 1, 0.09953206323817326: 1,
0.0988742951075605: 1, 0.09853564752122215: 1, 0.09810439433603096: 1,
0.09797547093286602: 1, 0.09783462841116887: 1, 0.09756719151899787: 1,
0.09717286152737653: 1, 0.09693350952007826: 1, 0.09660098387796942: 1,
0.0965736392149951: 1, 0.0958923830166986: 1, 0.09566697549958995: 1,
0.09476262165480367: 1, 0.09434330513101687: 1, 0.09414751347555342: 1,
0.09401128740151057: 1, 0.09374425105785747: 1, 0.09362595103667978: 1,
0.09336218422157831: 1, 0.09331309079634688: 1, 0.09291603509733368: 1,
0.09270271667481733: 1, 0.09244307687723005: 1, 0.09201219985511375: 1,
0.09174885304415352: 1, 0.09147435934151743: 1, 0.09127998996857828: 1,
0.09115454850625462: 1, 0.09107746322321114: 1, 0.09095162636382376: 1,
0.09094244148031888: 1, 0.09093755170673684: 1, 0.09092123217864265: 1,
0.09071078399489177: 1, 0.09064984362667353: 1, 0.09052228451235783: 1,
0.08989940021175824: 1, 0.08943032288695654: 1, 0.08942886787008536: 1,
0.08719194109915886: 1, 0.08693644003682825: 1, 0.08663671411715602: 1,
0.08633415970127292: 1, 0.08581589158250165: 1, 0.08547018651433176: 1,
0.08500474754121018: 1, 0.08437340790389516: 1, 0.08373479292128859: 1,
0.08333155738172685: 1, 0.08203930581542107: 1, 0.08183462893308277: 1,
0.08038228147942485: 1, 0.08032740400408113: 1, 0.08019254900860005: 1,
0.07996871230765296: 1, 0.07985032896215706: 1, 0.07975084919962197: 1,
0.07923564338418786: 1, 0.07816559929755197: 1, 0.07739505269175417: 1,
0.0770215392178507: 1, 0.07682764487456399: 1, 0.07648242447410346: 1,
0.0764182873738883: 1, 0.07636180745307097: 1, 0.0762382081890318: 1,
0.076130794974754: 1, 0.07544980940443007: 1, 0.0753452975176547 3: 1,
0.07530916069880528: 1, 0.07516731591635638: 1, 0.07509192482392728: 1,
0.07408229057727213: 1, 0.07400446861057067: 1, 0.0735437900937563: 1,
0.07342178127673449: 1, 0.0731073925830413: 1, 0.0730710103232779: 1,
0.07294585555386499: 1, 0.0727760482792413: 1, 0.0725958444900923 5: 1,
0.07238116367271381: 1, 0.07196008246103153: 1, 0.07165313579556506: 1,
0.0712949133911959: 1, 0.0706338389252504 4: 1, 0.0705110887901071 3: 1,
0.07040685355954579: 1, 0.07013051834391726: 1, 0.06944612343754637: 1,
0.06938331615738803: 1, 0.06934815989113438: 1, 0.06906046175140308: 1,
0.06876492769012504: 1, 0.06863247556239951: 1, 0.06858598447175562: 1,
0.06853607917667164: 1, 0.06819133484914155: 1, 0.06802050441135689: 1,
0.06794114037960561: 1, 0.06785408448982724: 1, 0.06658546726560802: 1,
0.06493933569711097: 1, 0.06457297128467406: 1, 0.06457103625278277: 1,
0.06452730816245641: 1, 0.06420343015518695: 1, 0.06419758825422897: 1,
0.06302099456952587: 1, 0.06272814006649151: 1, 0.06193899778523330 5:
1, 0.06189269942916262: 1, 0.06169145115898528: 1, 0.06134311120937846:
1, 0.06107225474898131: 1, 0.06073442224192411 6: 1, 0.06054127476168867
6: 1, 0.06042034215544486: 1, 0.06033792779046192 5: 1, 0.0602086673052 1
621: 1, 0.06005595068398566 4: 1, 0.05981696377610748: 1, 0.059602414980
66997: 1, 0.05950013149321619 5: 1, 0.05938511498411327: 1, 0.0592349674
7707207: 1, 0.05875100076394728: 1, 0.05864950312211148: 1, 0.058472631
28544179: 1, 0.05829511742805311 4: 1, 0.05768449178350443: 1, 0.0575895
0945663842: 1, 0.057555559490227 6: 1, 0.05746084895042559 4: 1, 0.057205
10387823661: 1, 0.05709867753694172: 1, 0.05687979420816316 5: 1, 0.0568
099832341225 5: 1, 0.05676088818749448: 1, 0.05668044656865 4: 1, 0.05643
144323307875 5: 1, 0.05615219293564034 7: 1, 0.05579737425789227: 1, 0.05
566349462965433 6: 1, 0.05550766516804 2: 1, 0.05516476010969781: 1, 0.05
515457853308665 4: 1, 0.05512724948077993: 1, 0.05510102008159155: 1, 0.
0549389556928572 3: 1, 0.05491582066090419: 1, 0.05475321002687473 5: 1,
0.05461439498558049 6: 1, 0.05402062944097294: 1, 0.0539409919379602: 1,
0.05384509648027943 6: 1, 0.05350386310008123 4: 1, 0.05309387879789042:
1, 0.0528643811619352 3: 1, 0.05284980163092349 6: 1, 0.05275423849479 68
9: 1, 0.05274399197903667 5: 1, 0.05263570257820674: 1, 0.0525506290292 8
5825: 1, 0.05170692430352321 5: 1, 0.05124461490662015 5: 1, 0.051180131 9
56495935: 1, 0.05113633152063836 6: 1, 0.05098291570723511 4: 1, 0.050832
268174768304: 1, 0.05081407414209386: 1, 0.05079419569705788 4: 1, 0.050
52887390092319: 1, 0.05042064452489483: 1, 0.05014382456808025: 1, 0.04
996104396213859: 1, 0.04976868306256487: 1, 0.04925382901673023: 1, 0.0
```

```
48937174360445305: 1, 0.04886395990139336: 1, 0.048818616685663355: 1,
0.04855907264694686: 1, 0.04812197334151612: 1, 0.04795952280269493: 1,
0.04788719457440078: 1, 0.04786237604992055: 1, 0.047820701441797384:
1, 0.047486259056457134: 1, 0.047455629548066175: 1, 0.047136777231104
5: 1, 0.047015118514523: 1, 0.04678188976855782: 1, 0.0466060472911239
3: 1, 0.045991435757888346: 1, 0.04577825162303506: 1, 0.04573706289246
0253: 1, 0.04564749186185053: 1, 0.04554633468226045: 1, 0.045529546447
05567: 1, 0.04415664130291744: 1, 0.04398300879923088: 1, 0.04396573508
287065: 1, 0.04384140457637173: 1, 0.043753906326611545: 1, 0.043745482
281493456: 1, 0.04372640015151265: 1, 0.0436630796064226: 1, 0.04363161
1394630954: 1, 0.04348432625719137: 1, 0.04341854559813331: 1, 0.042799
60039814806: 1, 0.04271011217040506: 1, 0.042607931935963725: 1, 0.0424
9293195120137: 1, 0.042471121320763694: 1, 0.042312337279878556: 1, 0.0
4207567139746807: 1, 0.04204951672722543: 1, 0.041958803595968285: 1,
0.04148883504589402: 1, 0.04139526357024405: 1, 0.04136845878525431: 1,
0.041132899884457466: 1, 0.040955306748115435: 1, 0.04076662325933495:
1, 0.04006029923791489: 1, 0.039939054607590486: 1, 0.03983792440049805
5: 1, 0.03974305864620184: 1, 0.03971229011050721: 1, 0.039668036455160
32: 1, 0.03961537679601285: 1, 0.039599676431109336: 1, 0.0394943556585
3321: 1, 0.0394518508271226: 1, 0.039219873930789434: 1, 0.038993732466
003406: 1, 0.03889945806809303: 1, 0.03876591158010802: 1, 0.0382210712
34999175: 1, 0.03812043699218887: 1, 0.03794416921111047: 1, 0.03777594
298970201: 1, 0.037750397861589256: 1, 0.037408720980331454: 1, 0.03711
174885575096: 1, 0.03695601407166796: 1, 0.03670988240717369: 1, 0.0366
9631858318854: 1, 0.03669073910484234: 1, 0.03643489608163277: 1, 0.036
351504505724455: 1, 0.036308243224346726: 1, 0.035931903294392345: 1,
0.035770754850653785: 1, 0.03571267828962793: 1, 0.035632572221356705:
1, 0.0355080702930982: 1, 0.035415317831210796: 1, 0.03536079946938317:
1, 0.035269567716587645: 1, 0.035269213601657724: 1, 0.0350364677380018
14: 1, 0.034829948052556826: 1, 0.034366203167249296: 1, 0.034358977228
129084: 1, 0.03425658910255102: 1, 0.03414571468541568: 1, 0.0341363437
40201316: 1, 0.03409645393415615: 1, 0.033929422884720284: 1, 0.0334595
0791170262: 1, 0.03337988476667457: 1, 0.03327345042760689: 1, 0.033157
99387655613: 1, 0.032635742827201046: 1, 0.032469667848555483: 1, 0.032
42970384540689: 1, 0.032411923295868575: 1, 0.032399525321205894: 1, 0.
032375125244005404: 1, 0.03221280073730234: 1, 0.03217944810234939: 1,
0.03215724662315257: 1, 0.03215540926914696: 1, 0.03196297232230214: 1,
0.03179236806691178: 1, 0.031370706159836406: 1, 0.03133472846175363:
1, 0.03129937367000141: 1, 0.031199320488499244: 1, 0.03109374505887506
```

```
7: 1, 0.0308758336837597 97: 1, 0.030857863774517 34: 1, 0.03085003202825
```

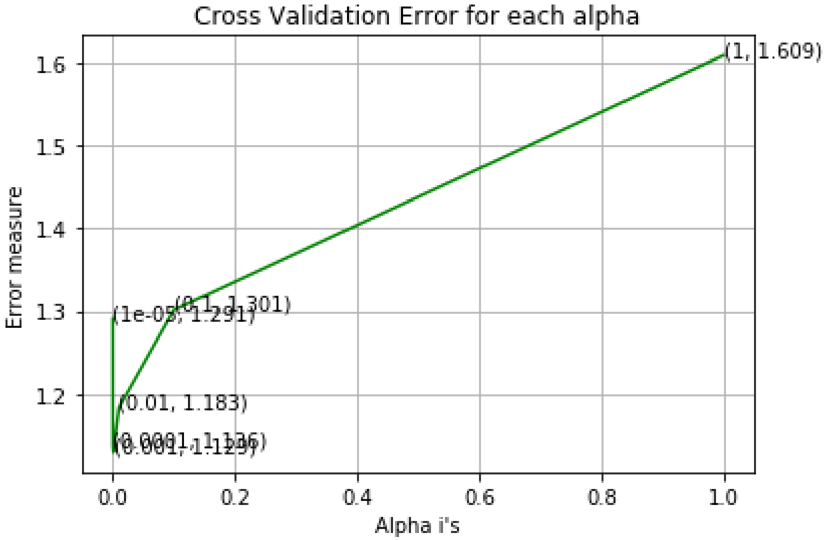
```
8177610696: 1, 0.0179322911816134: 1, 0.017783074452815868: 1, 0.017254
13190500498: 1, 0.01712835275354158: 1, 0.01709693976521024: 1, 0.01688
9424284697592: 1, 0.01685692583634627: 1, 0.01668987100434885: 1, 0.016
234833924277742: 1, 0.0160235778855855: 1, 0.01596036561957233: 1, 0.01
5869711298621556: 1, 0.015858771978449808: 1, 0.01574428388904523: 1,
0.015622307424049228: 1, 0.015546540395863876: 1, 0.015481339121600497:
1, 0.015458686670549577: 1, 0.015432980592456607: 1, 0.015083805257772З
1: 1, 0.014848997740955628: 1, 0.014654617601766999: 1, 0.0144722366095
60984: 1, 0.014431976059285688: 1, 0.014428841572446293: 1, 0.014396979
604351802: 1, 0.014231078384726493: 1, 0.014200900520766203: 1, 0.01419
0129557965465: 1, 0.014153729682979918: 1, 0.014107860808269689: 1, 0.0
14081032445397615: 1, 0.01402336809209615: 1, 0.013955809703184б: 1, 0.
013851191348471535: 1, 0.013709032116748579: 1, 0.013510664549234632:
1, 0.013251400948885113: 1, 0.012733560184457876: 1, 0.012437235961214５
8: 1, 0.012410550050843759: 1, 0.012297414095248104: 1, 0.0122267710553
4821: 1, 0.012108896716125962: 1, 0.012002866633198118: 1, 0.0119909536
67218569: 1, 0.011662107278650656: 1, 0.011262878416978355: 1, 0.011091
602505112196: 1, 0.010581277151479l6: 1, 0.009973112553041362: 1, 0.009
879446550578144: 1, 0.009495754794276８2: 1, 0.009072809830492565: 1, 0.
008622673868876713: 1, 0.008604470066323753: 1, 0.008347457786311８9: 1,
0.008318082123543988: 1, 0.007770392210106169５: 1, 0.00651650348686126
1: 1, 0.005001253670099914: 1})
```

```
In [73]: # Train a Logistic regression+Calibration model using text features whi
         cha re on-hot encoded
         alpha = [10 ** x for x in range(-5, 1)]

         # read more about SGDClassifier() at http://scikit-learn.org/stable/mod
         ules/generated/sklearn.linear_model.SGDClassifier.html
         # ------------------------------
         # default parameters
         # SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
         5, fit_intercept=True, max_iter=None, tol=None,
         # shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
         arning_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 S
```

```
tochastic 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.clas
ses_, 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_arra
y[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, predict_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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
For values of alpha =  1e-05 The log loss is: 1.2906403036339138
For values of alpha =  0.0001 The log loss is: 1.1363099270543104
For values of alpha =  0.001 The log loss is: 1.1292792746287146
For values of alpha =  0.01 The log loss is: 1.1832056257958172
For values of alpha =  0.1 The log loss is: 1.3010851196673259
For values of alpha =  1 The log loss is: 1.6086712540971753
```



```
For values of best alpha =  0.001 The train log loss is: 0.638122385022
0423
For values of best alpha =  0.001 The cross validation log loss is: 1.1
For values of best alpha =  0.001 The cross validation log loss is: 1.1
292792746287146
For values of best alpha =  0.001 The test log loss is: 1.0592532621982
15
```

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

Ans. Yes, it seems like!

In [0]:

```
def get_intersec_text(df):
    df_text_vec = TfidfVectorizer(min_df=3)
    df_text_fea = df_text_vec.fit_transform(df['TEXT'])

    df_text_features = top_mean_feats(df_text_fea,
                                      df_text_vec.get_feature_names(),
                                      top_n=1000)['feature'].tolist()

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

In [75]:

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

```
61.2 % of word of test data appeared in train data
61.5 % of word of Cross Validation appeared in train data
```

4. Machine Learning Models

In [0]:

```
#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 probabilit
ies 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 [0]:

```
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 [0]:

```
# 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 = TfidfVectorizer()
    var_count_vec = TfidfVectorizer()
    text_count_vec = TfidfVectorizer(min_df=3)

    gene_vec = gene_count_vec.fit(x_train['Gene'])
    var_vec  = var_count_vec.fit(x_train['Variation'])
    text_vec = text_count_vec.fit(x_train['TEXT'])
    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 = gene_vec.get_feature_names()[v]
            yes_no = True if word == gene else False
            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 = var_vec.get_feature_names()[v-(fea1_len)]
            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 = text_vec.get_feature_names()[v-(fea1_len+fea2_len)]
            yes_no = True if word in text.split() else False
            if yes_no:
                word_present += 1
                print(i, "Text feature [{}] present in test data point [{}]".format(word,yes_no))

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

Stacking the three types of features

In [0]:

```
# 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_onehotCoding))
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(y_train['Class']))

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

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


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
cv_gene_var_responseCoding = np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))

train_x_responseCoding = np.hstack((train_gene_var_responseCoding, train_text_feature_responseCoding))
test_x_responseCoding = np.hstack((test_gene_var_responseCoding, test_t
ext_feature_responseCoding))
cv_x_responseCoding = np.hstack((cv_gene_var_responseCoding, cv_text_fe
ature_responseCoding))
```

In [80]:

```
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 = ", t
est_x_onehotCoding.shape)
print("(number of data points * number of features) in cross validation
 data =", cv_x_onehotCoding.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 55
527)
(number of data points * number of features) in test data =  (665, 5552
7)
(number of data points * number of features) in cross validation data =
(532, 55527)
```

In [81]:

```
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 = ", t
est_x_responseCoding.shape)
print("(number of data points * number of features) in cross validation
 data =", cv_x_responseCoding.shape)
```

```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 2
7)
(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)
```

4.1. Base Line Model

4.1.1. Naive Bayes

4.1.1.1. Hyper parameter tuning

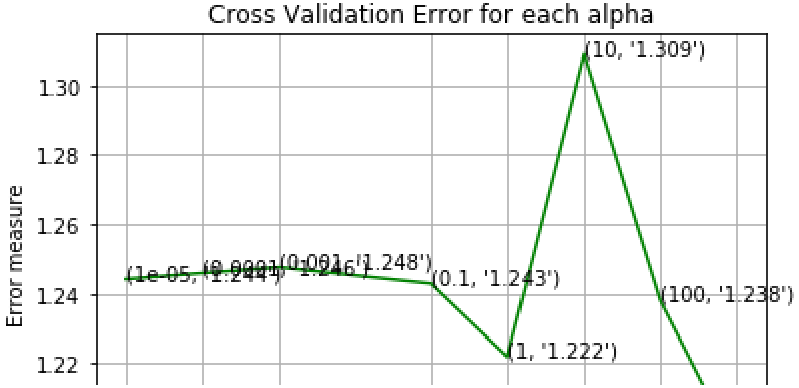
In [83]:

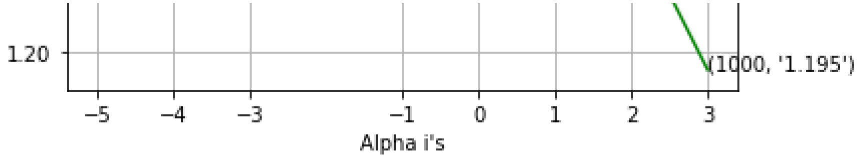
```
# find more about Multinomial Naive base function here http://scikit-le
arn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_pr
ior=None)

# some of methods of MultinomialNB()
# fit(X, y[, sample_weight])    Fit Naive Bayes classifier according to
 X, y
# predict(X)    Perform classification on an array of test vectors X.
# predict_log_proba(X)  Return log-probability estimates for the test v
ector X.
# -----------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/naive-bayes-algorithm-1/
# -----------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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: https://www.appliedaicourse.com/course/applied-ai-course-
```

```
online/lessons/naive-bayes-algorithm-1/
# ----------------------


alpha = [0.00001, 0.0001, 0.001, 0.1, 1, 10, 100,1000]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = MultinomialNB(alpha=i)
    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(np.log10(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)), (np.log10(alpha[i]),cv_log_error_a
rray[i]))
plt.grid()
plt.xticks(np.log10(alpha))
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 = MultinomialNB(alpha=alpha[best_alpha])
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 vali
dation 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 l
oss is:",log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for alpha = 1e-05
Log Loss : 1.2441919287301777
for alpha = 0.0001
Log Loss : 1.24589632132774
for alpha = 0.001
Log Loss : 1.2476088854582121
for alpha = 0.1
Log Loss : 1.2429138292862616
for alpha = 1
Log Loss : 1.2218884269527595
for alpha = 10
Log Loss : 1.3090154859112322
for alpha = 100
Log Loss : 1.238201355475871
for alpha = 1000
Log Loss : 1.194867860769994
```





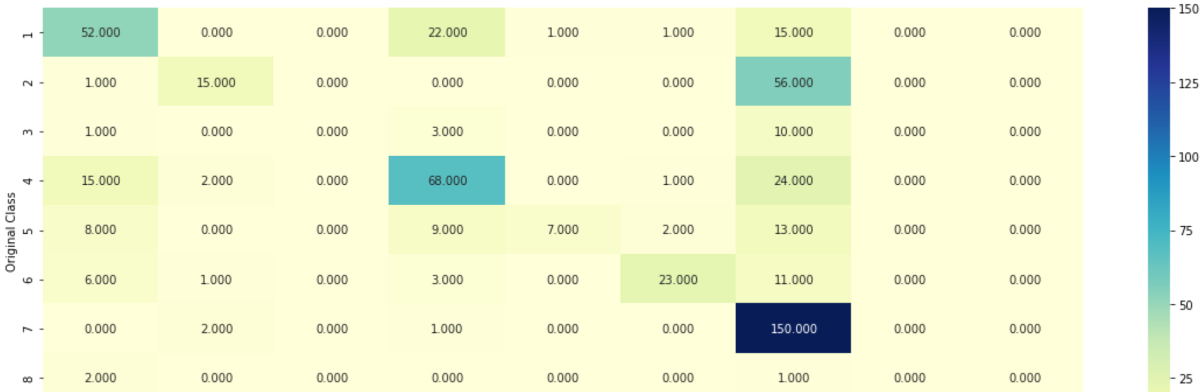
```
For values of best alpha =  1000 The train log loss is: 0.9487894148226409
For values of best alpha =  1000 The cross validation log loss is: 1.194867860769994
For values of best alpha =  1000 The test log loss is: 1.2168228044797074
```

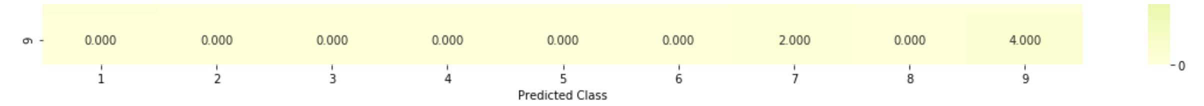
4.1.1.2. Testing the model with best hyper paramters

In [84]:

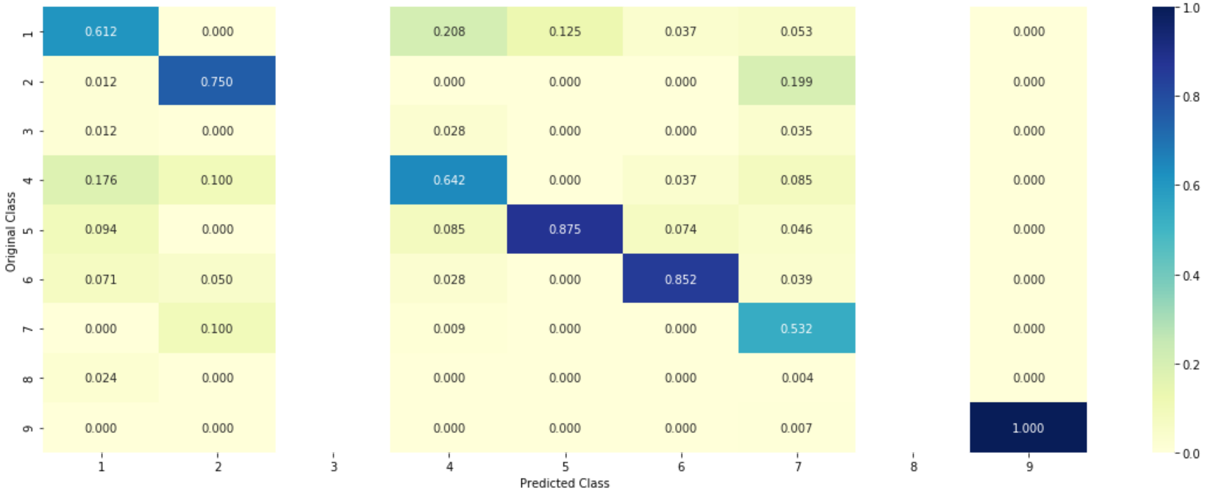
```
# find more about Multinomial Naive base function here http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html
# -------------------------
# default paramters
# sklearn.naive_bayes.MultinomialNB(alpha=1.0, fit_prior=True, class_prior=None)

# some of methods of MultinomialNB()
# fit(X, y[, sample_weight])    Fit Naive Bayes classifier according to X, y
# predict(X)    Perform classification on an array of test vectors X.
# predict_log_proba(X)  Return log-probability estimates for the test vector X.
# -----------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/naive-bayes-algorithm-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, metho
d='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
# ----------------------------

clf = MultinomialNB(alpha=alpha[best_alpha])
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)
# to avoid rounding error while multiplying probabilites we use log-pro
bability estimates
print("Log Loss :",log_loss(cv_y, sig_clf_probs))
print("Number of missclassified point :", np.count_nonzero((sig_clf.pre
dict(cv_x_onehotCoding)- cv_y))/cv_y.shape[0])
plot_confusion_matrix(cv_y, sig_clf.predict(cv_x_onehotCoding.toarray
()))
```

```
Log Loss : 1.194867860769994
Number of missclassified point : 0.40037593984962405

-------------------- Confusion matrix --------------------
```

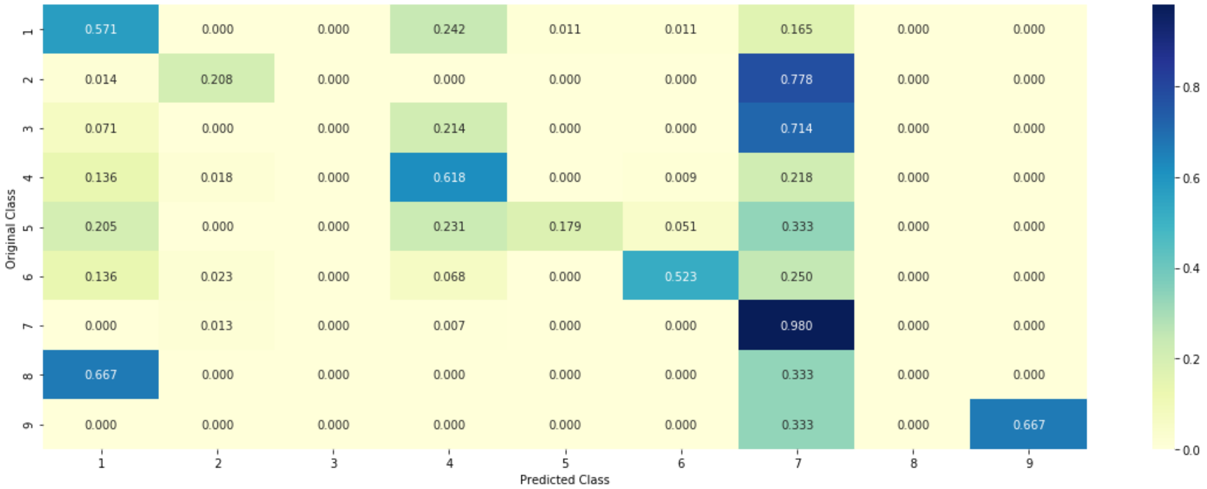




```
-------------------- Precision matrix (Columm Sum=1) ------------------
--
```



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



4.1.1.3. Feature Importance, Correctly classified point

In [85]:

```
test_point_index = 1
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[3.10e-03 3.29e-02 4.00e-04 2.00e-03 6.
40e-03 8.50e-03 9.46e-01 5.00e-04
  2.00e-04]]
Actual Class : 2
--------------------------------------------------
15 Text feature [cells] present in test data point [True]
16 Text feature [activated] present in test data point [True]
17 Text feature [kinase] present in test data point [True]
18 Text feature [activation] present in test data point [True]
19 Text feature [cell] present in test data point [True]
21 Text feature [downstream] present in test data point [True]
22 Text feature [contrast] present in test data point [True]
23 Text feature [factor] present in test data point [True]
24 Text feature [presence] present in test data point [True]
```

```
24 Text feature [presence] present in test data point [True]
25 Text feature [expressing] present in test data point [True]
26 Text feature [phosphorylation] present in test data point [True]
27 Text feature [shown] present in test data point [True]
28 Text feature [growth] present in test data point [True]
29 Text feature [signaling] present in test data point [True]
33 Text feature [however] present in test data point [True]
34 Text feature [inhibitor] present in test data point [True]
35 Text feature [also] present in test data point [True]
36 Text feature [recently] present in test data point [True]
37 Text feature [suggest] present in test data point [True]
38 Text feature [addition] present in test data point [True]
39 Text feature [10] present in test data point [True]
40 Text feature [independent] present in test data point [True]
41 Text feature [1a] present in test data point [True]
42 Text feature [compared] present in test data point [True]
43 Text feature [previously] present in test data point [True]
44 Text feature [treated] present in test data point [True]
45 Text feature [found] present in test data point [True]
46 Text feature [similar] present in test data point [True]
47 Text feature [mechanism] present in test data point [True]
48 Text feature [interestingly] present in test data point [True]
49 Text feature [constitutively] present in test data point [True]
50 Text feature [described] present in test data point [True]
51 Text feature [showed] present in test data point [True]
52 Text feature [increased] present in test data point [True]
53 Text feature [mutations] present in test data point [True]
54 Text feature [potential] present in test data point [True]
55 Text feature [well] present in test data point [True]
56 Text feature [treatment] present in test data point [True]
57 Text feature [3b] present in test data point [True]
58 Text feature [demonstrated] present in test data point [True]
61 Text feature [tyrosine] present in test data point [True]
62 Text feature [serum] present in test data point [True]
63 Text feature [followed] present in test data point [True]
64 Text feature [antibodies] present in test data point [True]
65 Text feature [confirmed] present in test data point [True]
66 Text feature [using] present in test data point [True]
67 Text feature [consistent] present in test data point [True]

68 Text feature [without] present in test data point [True]
```

```
68 Text feature [without] present in test data point [True]
69 Text feature [sensitive] present in test data point [True]
70 Text feature [inhibition] present in test data point [True]
71 Text feature [inhibited] present in test data point [True]
72 Text feature [lines] present in test data point [True]
73 Text feature [absence] present in test data point [True]
74 Text feature [fig] present in test data point [True]
75 Text feature [mutant] present in test data point [True]
76 Text feature [higher] present in test data point [True]
77 Text feature [various] present in test data point [True]
78 Text feature [figure] present in test data point [True]
79 Text feature [observed] present in test data point [True]
80 Text feature [inhibitors] present in test data point [True]
81 Text feature [enhanced] present in test data point [True]
82 Text feature [expression] present in test data point [True]
83 Text feature [furthermore] present in test data point [True]
84 Text feature [expressed] present in test data point [True]
85 Text feature [including] present in test data point [True]
86 Text feature [3a] present in test data point [True]
87 Text feature [total] present in test data point [True]
88 Text feature [detected] present in test data point [True]
89 Text feature [reported] present in test data point [True]
90 Text feature [activating] present in test data point [True]
91 Text feature [obtained] present in test data point [True]
92 Text feature [proliferation] present in test data point [True]
93 Text feature [respectively] present in test data point [True]
94 Text feature [antibody] present in test data point [True]
95 Text feature [mutation] present in test data point [True]
96 Text feature [increase] present in test data point [True]
97 Text feature [performed] present in test data point [True]
98 Text feature [pathways] present in test data point [True]
100 Text feature [may] present in test data point [True]
101 Text feature [two] present in test data point [True]
102 Text feature [concentrations] present in test data point [True]
103 Text feature [identified] present in test data point [True]
104 Text feature [approved] present in test data point [True]
105 Text feature [1b] present in test data point [True]
106 Text feature [confirm] present in test data point [True]
107 Text feature [examined] present in test data point [True]

108 Text feature [induced] present in test data point [True]
```

```
108 Text feature [induced] present in test data point [True]
109 Text feature [activate] present in test data point [True]
110 Text feature [previous] present in test data point [True]
111 Text feature [western] present in test data point [True]
112 Text feature [receptor] present in test data point [True]
115 Text feature [approximately] present in test data point [True]
116 Text feature [3c] present in test data point [True]
117 Text feature [survival] present in test data point [True]
118 Text feature [recent] present in test data point [True]
119 Text feature [role] present in test data point [True]
120 Text feature [small] present in test data point [True]
121 Text feature [mitogen] present in test data point [True]
123 Text feature [hours] present in test data point [True]
125 Text feature [occur] present in test data point [True]
126 Text feature [leading] present in test data point [True]
127 Text feature [whether] present in test data point [True]
128 Text feature [development] present in test data point [True]
129 Text feature [either] present in test data point [True]
130 Text feature [12] present in test data point [True]
131 Text feature [15] present in test data point [True]
132 Text feature [report] present in test data point [True]
133 Text feature [suggesting] present in test data point [True]
134 Text feature [domain] present in test data point [True]
135 Text feature [molecular] present in test data point [True]
136 Text feature [20] present in test data point [True]
137 Text feature [13] present in test data point [True]
138 Text feature [2b] present in test data point [True]
139 Text feature [next] present in test data point [True]
140 Text feature [oncogenic] present in test data point [True]
141 Text feature [due] present in test data point [True]
142 Text feature [4a] present in test data point [True]
143 Text feature [suggests] present in test data point [True]
144 Text feature [revealed] present in test data point [True]
145 Text feature [tumor] present in test data point [True]
146 Text feature [different] present in test data point [True]
147 Text feature [dependent] present in test data point [True]
148 Text feature [study] present in test data point [True]
149 Text feature [together] present in test data point [True]
150 Text feature [results] present in test data point [True]

151 Text feature [led] present in test data point [True]
```

```
151 Text feature [led] present in test data point [True]
152 Text feature [studies] present in test data point [True]
154 Text feature [4b] present in test data point [True]
155 Text feature [findings] present in test data point [True]
156 Text feature [three] present in test data point [True]
157 Text feature [pathway] present in test data point [True]
158 Text feature [discussion] present in test data point [True]
159 Text feature [thus] present in test data point [True]
160 Text feature [despite] present in test data point [True]
161 Text feature [mechanisms] present in test data point [True]
162 Text feature [transduction] present in test data point [True]
163 Text feature [kinases] present in test data point [True]
164 Text feature [resulting] present in test data point [True]
166 Text feature [show] present in test data point [True]
167 Text feature [express] present in test data point [True]
168 Text feature [effective] present in test data point [True]
169 Text feature [new] present in test data point [True]
170 Text feature [promote] present in test data point [True]
171 Text feature [therapeutic] present in test data point [True]
172 Text feature [initial] present in test data point [True]
173 Text feature [lead] present in test data point [True]
174 Text feature [could] present in test data point [True]
175 Text feature [tissue] present in test data point [True]
176 Text feature [assessed] present in test data point [True]
177 Text feature [common] present in test data point [True]
178 Text feature [similarly] present in test data point [True]
179 Text feature [although] present in test data point [True]
180 Text feature [culture] present in test data point [True]
181 Text feature [conditions] present in test data point [True]
182 Text feature [whereas] present in test data point [True]
183 Text feature [identification] present in test data point [True]
184 Text feature [measured] present in test data point [True]
185 Text feature [18] present in test data point [True]
186 Text feature [suggested] present in test data point [True]
187 Text feature [patients] present in test data point [True]
188 Text feature [active] present in test data point [True]
189 Text feature [within] present in test data point [True]
190 Text feature [single] present in test data point [True]
191 Text feature [human] present in test data point [True]

192 Text feature [ligand] present in test data point [True]
```

```
192 Text feature [ligand] present in test data point [True]
193 Text feature [drug] present in test data point [True]
194 Text feature [institutional] present in test data point [True]
195 Text feature [one] present in test data point [True]
196 Text feature [others] present in test data point [True]
197 Text feature [high] present in test data point [True]
198 Text feature [medium] present in test data point [True]
201 Text feature [positive] present in test data point [True]
203 Text feature [sequenced] present in test data point [True]
204 Text feature [analysis] present in test data point [True]
205 Text feature [overexpression] present in test data point [True]
206 Text feature [mutants] present in test data point [True]
207 Text feature [might] present in test data point [True]
208 Text feature [additional] present in test data point [True]
210 Text feature [wt] present in test data point [True]
211 Text feature [result] present in test data point [True]
212 Text feature [regulated] present in test data point [True]
213 Text feature [less] present in test data point [True]
215 Text feature [indicated] present in test data point [True]
216 Text feature [commonly] present in test data point [True]
217 Text feature [indicate] present in test data point [True]
218 Text feature [point] present in test data point [True]
219 Text feature [another] present in test data point [True]
220 Text feature [analyzed] present in test data point [True]
222 Text feature [relative] present in test data point [True]
223 Text feature [specific] present in test data point [True]
225 Text feature [effects] present in test data point [True]
227 Text feature [clinical] present in test data point [True]
228 Text feature [indicating] present in test data point [True]
229 Text feature [present] present in test data point [True]
230 Text feature [according] present in test data point [True]
231 Text feature [subsequently] present in test data point [True]
232 Text feature [several] present in test data point [True]
234 Text feature [mediated] present in test data point [True]
235 Text feature [1c] present in test data point [True]
236 Text feature [oncogene] present in test data point [True]
237 Text feature [transformation] present in test data point [True]
238 Text feature [trials] present in test data point [True]
239 Text feature [determine] present in test data point [True]

240 Text feature [unlike] present in test data point [True]
```

```
240 Text feature [unlike] present in test data point [True]
241 Text feature [directly] present in test data point [True]
242 Text feature [target] present in test data point [True]
243 Text feature [contribute] present in test data point [True]
244 Text feature [mapk] present in test data point [True]
245 Text feature [progression] present in test data point [True]
246 Text feature [generated] present in test data point [True]
247 Text feature [genomic] present in test data point [True]
248 Text feature [table] present in test data point [True]
251 Text feature [sequencing] present in test data point [True]
253 Text feature [leads] present in test data point [True]
254 Text feature [detect] present in test data point [True]
255 Text feature [test] present in test data point [True]
257 Text feature [days] present in test data point [True]
259 Text feature [four] present in test data point [True]
261 Text feature [lung] present in test data point [True]
262 Text feature [blot] present in test data point [True]
263 Text feature [taken] present in test data point [True]
264 Text feature [25] present in test data point [True]
265 Text feature [somatic] present in test data point [True]
266 Text feature [evaluated] present in test data point [True]
267 Text feature [signal] present in test data point [True]
268 Text feature [per] present in test data point [True]
269 Text feature [ml] present in test data point [True]
270 Text feature [novel] present in test data point [True]
271 Text feature [gene] present in test data point [True]
272 Text feature [experiments] present in test data point [True]
273 Text feature [derived] present in test data point [True]
274 Text feature [characterized] present in test data point [True]
275 Text feature [following] present in test data point [True]
276 Text feature [first] present in test data point [True]
277 Text feature [standard] present in test data point [True]
278 Text feature [lower] present in test data point [True]
279 Text feature [2a] present in test data point [True]
280 Text feature [normal] present in test data point [True]
281 Text feature [major] present in test data point [True]
282 Text feature [primary] present in test data point [True]
283 Text feature [prepared] present in test data point [True]
284 Text feature [30] present in test data point [True]

285 Text feature [level] present in test data point [True]
```

```
285 Text feature [level] present in test data point [True]
286 Text feature [targets] present in test data point [True]
287 Text feature [27] present in test data point [True]
288 Text feature [caused] present in test data point [True]
289 Text feature [24] present in test data point [True]
290 Text feature [selected] present in test data point [True]
291 Text feature [tumors] present in test data point [True]
292 Text feature [representative] present in test data point [True]
293 Text feature [later] present in test data point [True]
294 Text feature [vitro] present in test data point [True]
295 Text feature [determined] present in test data point [True]
296 Text feature [targeted] present in test data point [True]
297 Text feature [download] present in test data point [True]
298 Text feature [activity] present in test data point [True]
299 Text feature [part] present in test data point [True]
300 Text feature [time] present in test data point [True]
301 Text feature [introduced] present in test data point [True]
302 Text feature [19] present in test data point [True]
303 Text feature [pcr] present in test data point [True]
305 Text feature [samples] present in test data point [True]
306 Text feature [non] present in test data point [True]
307 Text feature [cancers] present in test data point [True]
308 Text feature [occurred] present in test data point [True]
311 Text feature [therapies] present in test data point [True]
313 Text feature [serine] present in test data point [True]
314 Text feature [provided] present in test data point [True]
316 Text feature [molecule] present in test data point [True]
317 Text feature [resulted] present in test data point [True]
318 Text feature [malignant] present in test data point [True]
319 Text feature [patient] present in test data point [True]
320 Text feature [much] present in test data point [True]
321 Text feature [variety] present in test data point [True]
322 Text feature [stably] present in test data point [True]
323 Text feature [sample] present in test data point [True]
324 Text feature [cause] present in test data point [True]
325 Text feature [regulation] present in test data point [True]
326 Text feature [frequently] present in test data point [True]
327 Text feature [important] present in test data point [True]
328 Text feature [seen] present in test data point [True]

329 Text feature [agent] present in test data point [True]
```

```
329 Text feature [agent] present in test data point [True]
330 Text feature [100] present in test data point [True]
331 Text feature [disease] present in test data point [True]
332 Text feature [harboring] present in test data point [True]
333 Text feature [manner] present in test data point [True]
334 Text feature [highly] present in test data point [True]
335 Text feature [significantly] present in test data point [True]
336 Text feature [support] present in test data point [True]
337 Text feature [protocol] present in test data point [True]
338 Text feature [40] present in test data point [True]
339 Text feature [key] present in test data point [True]
340 Text feature [f3] present in test data point [True]
341 Text feature [open] present in test data point [True]
343 Text feature [still] present in test data point [True]
344 Text feature [collection] present in test data point [True]
345 Text feature [containing] present in test data point [True]
346 Text feature [currently] present in test data point [True]
347 Text feature [play] present in test data point [True]
348 Text feature [ba] present in test data point [True]
349 Text feature [advanced] present in test data point [True]
351 Text feature [therefore] present in test data point [True]
353 Text feature [48] present in test data point [True]
354 Text feature [investigated] present in test data point [True]
355 Text feature [cancer] present in test data point [True]
356 Text feature [rate] present in test data point [True]
357 Text feature [possibility] present in test data point [True]
358 Text feature [free] present in test data point [True]
359 Text feature [fold] present in test data point [True]
360 Text feature [manufacturer] present in test data point [True]
361 Text feature [represent] present in test data point [True]
363 Text feature [activates] present in test data point [True]
364 Text feature [17] present in test data point [True]
365 Text feature [distinct] present in test data point [True]
366 Text feature [oncogenes] present in test data point [True]
367 Text feature [whose] present in test data point [True]
368 Text feature [established] present in test data point [True]
369 Text feature [transform] present in test data point [True]
370 Text feature [transfected] present in test data point [True]
371 Text feature [form] present in test data point [True]

372 Text feature [increasing] present in test data point [True]
```

```
372 Text feature [increasing] present in test data point [True]
373 Text feature [institute] present in test data point [True]
374 Text feature [identify] present in test data point [True]
375 Text feature [negative] present in test data point [True]
376 Text feature [significant] present in test data point [True]
377 Text feature [collected] present in test data point [True]
378 Text feature [receptors] present in test data point [True]
379 Text feature [levels] present in test data point [True]
380 Text feature [represents] present in test data point [True]
381 Text feature [3t3] present in test data point [True]
382 Text feature [driven] present in test data point [True]
383 Text feature [1d] present in test data point [True]
384 Text feature [14] present in test data point [True]
385 Text feature [tissues] present in test data point [True]
391 Text feature [remains] present in test data point [True]
396 Text feature [finding] present in test data point [True]
397 Text feature [even] present in test data point [True]
398 Text feature [harbor] present in test data point [True]
399 Text feature [ng] present in test data point [True]
400 Text feature [remain] present in test data point [True]
401 Text feature [96] present in test data point [True]
402 Text feature [epidermal] present in test data point [True]
403 Text feature [erk] present in test data point [True]
404 Text feature [22] present in test data point [True]
405 Text feature [50] present in test data point [True]
406 Text feature [differences] present in test data point [True]
407 Text feature [s3] present in test data point [True]
408 Text feature [response] present in test data point [True]
409 Text feature [entire] present in test data point [True]
411 Text feature [11] present in test data point [True]
412 Text feature [stable] present in test data point [True]
413 Text feature [strongly] present in test data point [True]
414 Text feature [possible] present in test data point [True]
415 Text feature [introduction] present in test data point [True]
416 Text feature [immunoblotting] present in test data point [True]
417 Text feature [used] present in test data point [True]
418 Text feature [concentration] present in test data point [True]
419 Text feature [highest] present in test data point [True]
420 Text feature [line] present in test data point [True]

421 Text feature [would] present in test data point [True]
```

```
421 Text feature [would] present in test data point [True]
423 Text feature [five] present in test data point [True]
424 Text feature [comparable] present in test data point [True]
425 Text feature [university] present in test data point [True]
427 Text feature [finally] present in test data point [True]
428 Text feature [benefit] present in test data point [True]
429 Text feature [mutated] present in test data point [True]
430 Text feature [decreased] present in test data point [True]
431 Text feature [day] present in test data point [True]
432 Text feature [associated] present in test data point [True]
433 Text feature [independently] present in test data point [True]
434 Text feature [clinically] present in test data point [True]
435 Text feature [type] present in test data point [True]
436 Text feature [powerpoint] present in test data point [True]
437 Text feature [therapy] present in test data point [True]
438 Text feature [appear] present in test data point [True]
439 Text feature [difference] present in test data point [True]
441 Text feature [3e] present in test data point [True]
443 Text feature [subjected] present in test data point [True]
444 Text feature [green] present in test data point [True]
445 Text feature [enhance] present in test data point [True]
447 Text feature [amplification] present in test data point [True]
448 Text feature [software] present in test data point [True]
449 Text feature [plays] present in test data point [True]
450 Text feature [contained] present in test data point [True]
451 Text feature [event] present in test data point [True]
453 Text feature [cases] present in test data point [True]
454 Text feature [almost] present in test data point [True]
455 Text feature [indeed] present in test data point [True]
456 Text feature [control] present in test data point [True]
457 Text feature [demonstrate] present in test data point [True]
458 Text feature [factors] present in test data point [True]
459 Text feature [conformation] present in test data point [True]
460 Text feature [doses] present in test data point [True]
461 Text feature [resistant] present in test data point [True]
462 Text feature [upon] present in test data point [True]
463 Text feature [carcinoma] present in test data point [True]
464 Text feature [agents] present in test data point [True]
466 Text feature [observation] present in test data point [True]

467 Text feature [prism] present in test data point [True]
```

```
467 Text feature [prism] present in test data point [True]
468 Text feature [complete] present in test data point [True]
469 Text feature [4c] present in test data point [True]
470 Text feature [transfection] present in test data point [True]
471 Text feature [transformed] present in test data point [True]
472 Text feature [yet] present in test data point [True]
474 Text feature [frozen] present in test data point [True]
475 Text feature [potent] present in test data point [True]
476 Text feature [developed] present in test data point [True]
477 Text feature [data] present in test data point [True]
478 Text feature [required] present in test data point [True]
479 Text feature [respond] present in test data point [True]
481 Text feature [plates] present in test data point [True]
482 Text feature [us] present in test data point [True]
483 Text feature [phosphorylated] present in test data point [True]
484 Text feature [failed] present in test data point [True]
487 Text feature [sensitivity] present in test data point [True]
488 Text feature [importance] present in test data point [True]
489 Text feature [endogenous] present in test data point [True]
491 Text feature [et] present in test data point [True]
492 Text feature [extracted] present in test data point [True]
493 Text feature [fact] present in test data point [True]
494 Text feature [confer] present in test data point [True]
495 Text feature [hospital] present in test data point [True]
496 Text feature [located] present in test data point [True]
499 Text feature [numbers] present in test data point [True]
500 Text feature [28] present in test data point [True]
501 Text feature [able] present in test data point [True]
502 Text feature [center] present in test data point [True]
503 Text feature [16] present in test data point [True]
504 Text feature [al] present in test data point [True]
505 Text feature [requires] present in test data point [True]
506 Text feature [ca] present in test data point [True]
507 Text feature [alone] present in test data point [True]
508 Text feature [initially] present in test data point [True]
510 Text feature [greater] present in test data point [True]
511 Text feature [understanding] present in test data point [True]
512 Text feature [range] present in test data point [True]
513 Text feature [via] present in test data point [True]

514 Text feature [low] present in test data point [True]
```

```
514 Text feature [low] present in test data point [True]
515 Text feature [s5] present in test data point [True]
516 Text feature [respective] present in test data point [True]
517 Text feature [often] present in test data point [True]
518 Text feature [s1] present in test data point [True]
519 Text feature [mouse] present in test data point [True]
520 Text feature [completely] present in test data point [True]
521 Text feature [remained] present in test data point [True]
522 Text feature [exhibited] present in test data point [True]
523 Text feature [status] present in test data point [True]
524 Text feature [generation] present in test data point [True]
525 Text feature [multiple] present in test data point [True]
526 Text feature [coding] present in test data point [True]
527 Text feature [150] present in test data point [True]
529 Text feature [combination] present in test data point [True]
530 Text feature [extracellular] present in test data point [True]
531 Text feature [nm] present in test data point [True]
534 Text feature [occurs] present in test data point [True]
535 Text feature [intracellular] present in test data point [True]
537 Text feature [72] present in test data point [True]
539 Text feature [reverse] present in test data point [True]
540 Text feature [trial] present in test data point [True]
541 Text feature [date] present in test data point [True]
542 Text feature [viability] present in test data point [True]
543 Text feature [80] present in test data point [True]
544 Text feature [constructs] present in test data point [True]
546 Text feature [provide] present in test data point [True]
547 Text feature [neither] present in test data point [True]
549 Text feature [whole] present in test data point [True]
550 Text feature [order] present in test data point [True]
552 Text feature [block] present in test data point [True]
553 Text feature [biological] present in test data point [True]
554 Text feature [effectively] present in test data point [True]
555 Text feature [responsible] present in test data point [True]
556 Text feature [regulates] present in test data point [True]
557 Text feature [number] present in test data point [True]
558 Text feature [transforming] present in test data point [True]
559 Text feature [sufficient] present in test data point [True]
560 Text feature [example] present in test data point [True]

561 Text feature [period] present in test data point [True]
```

```
561 Text feature [period] present in test data point [True]
564 Text feature [stage] present in test data point [True]
565 Text feature [direct] present in test data point [True]
566 Text feature [separated] present in test data point [True]
567 Text feature [targeting] present in test data point [True]
568 Text feature [wild] present in test data point [True]
569 Text feature [reaction] present in test data point [True]
571 Text feature [32] present in test data point [True]
573 Text feature [board] present in test data point [True]
575 Text feature [33] present in test data point [True]
576 Text feature [corresponding] present in test data point [True]
577 Text feature [effect] present in test data point [True]
578 Text feature [protein] present in test data point [True]
579 Text feature [paraffin] present in test data point [True]
580 Text feature [involving] present in test data point [True]
582 Text feature [transient] present in test data point [True]
583 Text feature [subsequent] present in test data point [True]
584 Text feature [since] present in test data point [True]
585 Text feature [identical] present in test data point [True]
586 Text feature [agar] present in test data point [True]
587 Text feature [isolated] present in test data point [True]
588 Text feature [acid] present in test data point [True]
589 Text feature [summary] present in test data point [True]
590 Text feature [observations] present in test data point [True]
591 Text feature [size] present in test data point [True]
592 Text feature [confirming] present in test data point [True]
594 Text feature [reports] present in test data point [True]
595 Text feature [majority] present in test data point [True]
596 Text feature [discovered] present in test data point [True]
597 Text feature [selective] present in test data point [True]
598 Text feature [appears] present in test data point [True]
599 Text feature [critical] present in test data point [True]
601 Text feature [course] present in test data point [True]
602 Text feature [initiated] present in test data point [True]
604 Text feature [investigate] present in test data point [True]
606 Text feature [appeared] present in test data point [True]
607 Text feature [glutamine] present in test data point [True]
608 Text feature [90] present in test data point [True]
609 Text feature [account] present in test data point [True]

610 Text feature [approval] present in test data point [True]
```

```
610 Text feature [approval] present in test data point [True]
612 Text feature [importantly] present in test data point [True]
615 Text feature [metastatic] present in test data point [True]
616 Text feature [regulate] present in test data point [True]
618 Text feature [briefly] present in test data point [True]
619 Text feature [case] present in test data point [True]
621 Text feature [thereby] present in test data point [True]
622 Text feature [primers] present in test data point [True]
624 Text feature [involved] present in test data point [True]
625 Text feature [overall] present in test data point [True]
626 Text feature [cdna] present in test data point [True]
627 Text feature [preclinical] present in test data point [True]
628 Text feature [frequent] present in test data point [True]
629 Text feature [particularly] present in test data point [True]
630 Text feature [proliferate] present in test data point [True]
631 Text feature [subset] present in test data point [True]
632 Text feature [1e] present in test data point [True]
633 Text feature [phase] present in test data point [True]
634 Text feature [adjacent] present in test data point [True]
636 Text feature [proteins] present in test data point [True]
637 Text feature [prolonged] present in test data point [True]
638 Text feature [solid] present in test data point [True]
639 Text feature [fixed] present in test data point [True]
640 Text feature [exons] present in test data point [True]
641 Text feature [4d] present in test data point [True]
642 Text feature [necessary] present in test data point [True]
643 Text feature [left] present in test data point [True]
645 Text feature [carried] present in test data point [True]
646 Text feature [exposure] present in test data point [True]
647 Text feature [least] present in test data point [True]
648 Text feature [vivo] present in test data point [True]
649 Text feature [akt] present in test data point [True]
651 Text feature [designed] present in test data point [True]
652 Text feature [toward] present in test data point [True]
659 Text feature [il] present in test data point [True]
660 Text feature [forms] present in test data point [True]
661 Text feature [shows] present in test data point [True]
663 Text feature [nsclc] present in test data point [True]
664 Text feature [23] present in test data point [True]

666 Text feature [develop] present in test data point [True]
```

```
666 Text feature [develop] present in test data point [True]
667 Text feature [find] present in test data point [True]
668 Text feature [transiently] present in test data point [True]
669 Text feature [statistical] present in test data point [True]
670 Text feature [notably] present in test data point [True]
672 Text feature [future] present in test data point [True]
673 Text feature [review] present in test data point [True]
674 Text feature [blocks] present in test data point [True]
676 Text feature [long] present in test data point [True]
677 Text feature [sanger] present in test data point [True]
678 Text feature [promoting] present in test data point [True]
680 Text feature [sections] present in test data point [True]
681 Text feature [additionally] present in test data point [True]
682 Text feature [limited] present in test data point [True]
683 Text feature [render] present in test data point [True]
686 Text feature [full] present in test data point [True]
689 Text feature [hybridization] present in test data point [True]
691 Text feature [frequency] present in test data point [True]
692 Text feature [egfr] present in test data point [True]
693 Text feature [six] present in test data point [True]
694 Text feature [cascade] present in test data point [True]
695 Text feature [mean] present in test data point [True]
696 Text feature [informed] present in test data point [True]
698 Text feature [current] present in test data point [True]
699 Text feature [model] present in test data point [True]
701 Text feature [37] present in test data point [True]
703 Text feature [include] present in test data point [True]
704 Text feature [harbored] present in test data point [True]
705 Text feature [expected] present in test data point [True]
706 Text feature [slightly] present in test data point [True]
708 Text feature [view] present in test data point [True]
709 Text feature [early] present in test data point [True]
710 Text feature [many] present in test data point [True]
712 Text feature [biosystems] present in test data point [True]
713 Text feature [consent] present in test data point [True]
714 Text feature [drugs] present in test data point [True]
716 Text feature [exon] present in test data point [True]
717 Text feature [origin] present in test data point [True]
719 Text feature [months] present in test data point [True]

721 Text feature [sites] present in test data point [True]
```

```
721 Text feature [sites] present in test data point [True]
725 Text feature [21] present in test data point [True]
726 Text feature [note] present in test data point [True]
727 Text feature [treating] present in test data point [True]
732 Text feature [explain] present in test data point [True]
734 Text feature [insensitive] present in test data point [True]
736 Text feature [bearing] present in test data point [True]
737 Text feature [hypothesized] present in test data point [True]
738 Text feature [versus] present in test data point [True]
739 Text feature [5a] present in test data point [True]
740 Text feature [context] present in test data point [True]
741 Text feature [median] present in test data point [True]
744 Text feature [amplified] present in test data point [True]
745 Text feature [experimental] present in test data point [True]
746 Text feature [inhibitory] present in test data point [True]
747 Text feature [substitution] present in test data point [True]
748 Text feature [glycine] present in test data point [True]
749 Text feature [values] present in test data point [True]
751 Text feature [spectrum] present in test data point [True]
753 Text feature [share] present in test data point [True]
755 Text feature [formation] present in test data point [True]
756 Text feature [properties] present in test data point [True]
757 Text feature [regions] present in test data point [True]
759 Text feature [smaller] present in test data point [True]
760 Text feature [tested] present in test data point [True]
761 Text feature [evaluation] present in test data point [True]
762 Text feature [inhibit] present in test data point [True]
763 Text feature [every] present in test data point [True]
765 Text feature [supports] present in test data point [True]
768 Text feature [wide] present in test data point [True]
770 Text feature [red] present in test data point [True]
771 Text feature [needed] present in test data point [True]
772 Text feature [represented] present in test data point [True]
774 Text feature [2000] present in test data point [True]
775 Text feature [schematic] present in test data point [True]
776 Text feature [accordance] present in test data point [True]
778 Text feature [medical] present in test data point [True]
779 Text feature [predominantly] present in test data point [True]
780 Text feature [design] present in test data point [True]

781 Text feature [unclear] present in test data point [True]
```

```
781 Text feature [nuclear] present in test data point [True]
782 Text feature [late] present in test data point [True]
783 Text feature [500] present in test data point [True]
784 Text feature [anchorage] present in test data point [True]
785 Text feature [short] present in test data point [True]
786 Text feature [progressed] present in test data point [True]
787 Text feature [ic50] present in test data point [True]
789 Text feature [third] present in test data point [True]
790 Text feature [comparison] present in test data point [True]
794 Text feature [adenocarcinoma] present in test data point [True]
795 Text feature [nucleotide] present in test data point [True]
798 Text feature [embedded] present in test data point [True]
799 Text feature [inactive] present in test data point [True]
801 Text feature [induce] present in test data point [True]
802 Text feature [seems] present in test data point [True]
804 Text feature [numerous] present in test data point [True]
807 Text feature [use] present in test data point [True]
808 Text feature [implicated] present in test data point [True]
809 Text feature [end] present in test data point [True]
810 Text feature [promotes] present in test data point [True]
811 Text feature [specimens] present in test data point [True]
812 Text feature [region] present in test data point [True]
813 Text feature [strategies] present in test data point [True]
814 Text feature [bank] present in test data point [True]
815 Text feature [matched] present in test data point [True]
816 Text feature [correlate] present in test data point [True]
817 Text feature [29] present in test data point [True]
819 Text feature [genes] present in test data point [True]
820 Text feature [equivalent] present in test data point [True]
821 Text feature [establish] present in test data point [True]
822 Text feature [biochemical] present in test data point [True]
824 Text feature [known] present in test data point [True]
825 Text feature [therapeutics] present in test data point [True]
826 Text feature [supported] present in test data point [True]
827 Text feature [prognosis] present in test data point [True]
828 Text feature [amino] present in test data point [True]
829 Text feature [mg] present in test data point [True]
830 Text feature [panel] present in test data point [True]
831 Text feature [components] present in test data point [True]

832 Text feature [metastasis] present in test data point [True]
```

```
832 Text feature [metastasis] present in test data point [True]
833 Text feature [life] present in test data point [True]
834 Text feature [see] present in test data point [True]
835 Text feature [like] present in test data point [True]
837 Text feature [week] present in test data point [True]
838 Text feature [gels] present in test data point [True]
841 Text feature [materials] present in test data point [True]
842 Text feature [hand] present in test data point [True]
843 Text feature [75] present in test data point [True]
844 Text feature [certain] present in test data point [True]
846 Text feature [blocking] present in test data point [True]
847 Text feature [instructions] present in test data point [True]
852 Text feature [events] present in test data point [True]
853 Text feature [preferentially] present in test data point [True]
854 Text feature [dual] present in test data point [True]
855 Text feature [resistance] present in test data point [True]
856 Text feature [larger] present in test data point [True]
858 Text feature [soft] present in test data point [True]
861 Text feature [poor] present in test data point [True]
862 Text feature [screen] present in test data point [True]
863 Text feature [120] present in test data point [True]
864 Text feature [laboratory] present in test data point [True]
865 Text feature [involve] present in test data point [True]
866 Text feature [primarily] present in test data point [True]
867 Text feature [among] present in test data point [True]
868 Text feature [site] present in test data point [True]
869 Text feature [44] present in test data point [True]
870 Text feature [acquired] present in test data point [True]
871 Text feature [250] present in test data point [True]
872 Text feature [conventional] present in test data point [True]
873 Text feature [selection] present in test data point [True]
874 Text feature [capable] present in test data point [True]
875 Text feature [best] present in test data point [True]
877 Text feature [exhibit] present in test data point [True]
878 Text feature [immediately] present in test data point [True]
881 Text feature [experiment] present in test data point [True]
887 Text feature [responses] present in test data point [True]
890 Text feature [latter] present in test data point [True]
892 Text feature [05] present in test data point [True]

893 Text feature [curves] present in test data point [True]
```

```
893 Text feature [curves] present in test data point [True]
894 Text feature [essential] present in test data point [True]
895 Text feature [contain] present in test data point [True]
896 Text feature [bone] present in test data point [True]
897 Text feature [rank] present in test data point [True]
899 Text feature [indicates] present in test data point [True]
900 Text feature [clearly] present in test data point [True]
901 Text feature [pattern] present in test data point [True]
902 Text feature [right] present in test data point [True]
903 Text feature [alterations] present in test data point [True]
904 Text feature [detection] present in test data point [True]
905 Text feature [triplicate] present in test data point [True]
906 Text feature [groups] present in test data point [True]
907 Text feature [listed] present in test data point [True]
908 Text feature [gastrointestinal] present in test data point [True]
910 Text feature [noted] present in test data point [True]
911 Text feature [done] present in test data point [True]
912 Text feature [epithelial] present in test data point [True]
913 Text feature [summarized] present in test data point [True]
914 Text feature [except] present in test data point [True]
915 Text feature [exclusive] present in test data point [True]
916 Text feature [products] present in test data point [True]
917 Text feature [marker] present in test data point [True]
918 Text feature [explanation] present in test data point [True]
919 Text feature [bars] present in test data point [True]
920 Text feature [obvious] present in test data point [True]
921 Text feature [transfer] present in test data point [True]
922 Text feature [blue] present in test data point [True]
923 Text feature [year] present in test data point [True]
924 Text feature [options] present in test data point [True]
926 Text feature [differential] present in test data point [True]
927 Text feature [understand] present in test data point [True]
928 Text feature [system] present in test data point [True]
929 Text feature [applied] present in test data point [True]
930 Text feature [interesting] present in test data point [True]
931 Text feature [histological] present in test data point [True]
932 Text feature [closely] present in test data point [True]
934 Text feature [dose] present in test data point [True]
935 Text feature [plus] present in test data point [True]

938 Text feature [must] present in test data point [True]
```

```
938 Text feature [must] present in test data point [True]
939 Text feature [chemotherapy] present in test data point [True]
940 Text feature [included] present in test data point [True]
941 Text feature [adenocarcinomas] present in test data point [True]
942 Text feature [times] present in test data point [True]
943 Text feature [conclusion] present in test data point [True]
944 Text feature [rates] present in test data point [True]
945 Text feature [implications] present in test data point [True]
947 Text feature [length] present in test data point [True]
948 Text feature [exogenous] present in test data point [True]
949 Text feature [frame] present in test data point [True]
950 Text feature [profile] present in test data point [True]
951 Text feature [near] present in test data point [True]
953 Text feature [types] present in test data point [True]
955 Text feature [particular] present in test data point [True]
956 Text feature [mainly] present in test data point [True]
959 Text feature [alternative] present in test data point [True]
960 Text feature [given] present in test data point [True]
961 Text feature [health] present in test data point [True]
962 Text feature [ongoing] present in test data point [True]
965 Text feature [procedures] present in test data point [True]
968 Text feature [characterization] present in test data point [True]
969 Text feature [analyses] present in test data point [True]
970 Text feature [sets] present in test data point [True]
972 Text feature [substantial] present in test data point [True]
974 Text feature [monoclonal] present in test data point [True]
975 Text feature [understood] present in test data point [True]
976 Text feature [old] present in test data point [True]
977 Text feature [image] present in test data point [True]
979 Text feature [potentially] present in test data point [True]
980 Text feature [residues] present in test data point [True]
982 Text feature [tumorigenesis] present in test data point [True]
983 Text feature [primer] present in test data point [True]
986 Text feature [mutually] present in test data point [True]
987 Text feature [specifically] present in test data point [True]
988 Text feature [codons] present in test data point [True]
989 Text feature [interleukin] present in test data point [True]
990 Text feature [ability] present in test data point [True]
991 Text feature [deletion] present in test data point [True]

993 Text feature [36] present in test data point [True]
```

```
993 Text feature [50] present in test data point [True]
994 Text feature [elucidated] present in test data point [True]
995 Text feature [codon] present in test data point [True]
996 Text feature [national] present in test data point [True]
999 Text feature [main] present in test data point [True]
Out of the top  1000  features  793 are present in query point
```

4.1.1.4. Feature Importance, Incorrectly classified point

In [86]:

```
test_point_index = 55
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.1486 0.1306 0.0093 0.226  0.0436 0.0
407 0.3957 0.0029 0.0026]]
Actual Class : 4
--------------------------------------------------
15 Text feature [cells] present in test data point [True]
18 Text feature [activation] present in test data point [True]
19 Text feature [cell] present in test data point [True]
24 Text feature [presence] present in test data point [True]
27 Text feature [shown] present in test data point [True]
29 Text feature [signaling] present in test data point [True]
33 Text feature [however] present in test data point [True]
35 Text feature [also] present in test data point [True]
36 Text feature [recently] present in test data point [True]
38 Text feature [addition] present in test data point [True]
```

```
39 Text feature [10] present in test data point [True]
41 Text feature [1a] present in test data point [True]
42 Text feature [compared] present in test data point [True]
43 Text feature [previously] present in test data point [True]
51 Text feature [showed] present in test data point [True]
53 Text feature [mutations] present in test data point [True]
54 Text feature [potential] present in test data point [True]
56 Text feature [treatment] present in test data point [True]
60 Text feature [constitutive] present in test data point [True]
65 Text feature [confirmed] present in test data point [True]
66 Text feature [using] present in test data point [True]
67 Text feature [consistent] present in test data point [True]
68 Text feature [without] present in test data point [True]
72 Text feature [lines] present in test data point [True]
77 Text feature [various] present in test data point [True]
78 Text feature [figure] present in test data point [True]
79 Text feature [observed] present in test data point [True]
82 Text feature [expression] present in test data point [True]
84 Text feature [expressed] present in test data point [True]
85 Text feature [including] present in test data point [True]
87 Text feature [total] present in test data point [True]
89 Text feature [reported] present in test data point [True]
95 Text feature [mutation] present in test data point [True]
96 Text feature [increase] present in test data point [True]
97 Text feature [performed] present in test data point [True]
98 Text feature [pathways] present in test data point [True]
100 Text feature [may] present in test data point [True]
101 Text feature [two] present in test data point [True]
103 Text feature [identified] present in test data point [True]
105 Text feature [1b] present in test data point [True]
108 Text feature [induced] present in test data point [True]
110 Text feature [previous] present in test data point [True]
112 Text feature [receptor] present in test data point [True]
115 Text feature [approximately] present in test data point [True]
117 Text feature [survival] present in test data point [True]
118 Text feature [recent] present in test data point [True]
127 Text feature [whether] present in test data point [True]
128 Text feature [development] present in test data point [True]

129 Text feature [either] present in test data point [True]
```

```
130 Text feature [12] present in test data point [True]
136 Text feature [20] present in test data point [True]
137 Text feature [13] present in test data point [True]
139 Text feature [next] present in test data point [True]
141 Text feature [due] present in test data point [True]
144 Text feature [revealed] present in test data point [True]
145 Text feature [tumor] present in test data point [True]
149 Text feature [together] present in test data point [True]
150 Text feature [results] present in test data point [True]
152 Text feature [studies] present in test data point [True]
155 Text feature [findings] present in test data point [True]
156 Text feature [three] present in test data point [True]
157 Text feature [pathway] present in test data point [True]
160 Text feature [despite] present in test data point [True]
161 Text feature [mechanisms] present in test data point [True]
164 Text feature [resulting] present in test data point [True]
174 Text feature [could] present in test data point [True]
177 Text feature [common] present in test data point [True]
179 Text feature [although] present in test data point [True]
182 Text feature [whereas] present in test data point [True]
183 Text feature [identification] present in test data point [True]
187 Text feature [patients] present in test data point [True]
189 Text feature [within] present in test data point [True]
191 Text feature [human] present in test data point [True]
195 Text feature [one] present in test data point [True]
197 Text feature [high] present in test data point [True]
201 Text feature [positive] present in test data point [True]
202 Text feature [technology] present in test data point [True]
204 Text feature [analysis] present in test data point [True]
207 Text feature [might] present in test data point [True]
213 Text feature [less] present in test data point [True]
215 Text feature [indicated] present in test data point [True]
216 Text feature [commonly] present in test data point [True]
217 Text feature [indicate] present in test data point [True]
220 Text feature [analyzed] present in test data point [True]
223 Text feature [specific] present in test data point [True]
229 Text feature [present] present in test data point [True]
232 Text feature [several] present in test data point [True]

237 Text feature [transformation] present in test data point [True]
```

```
243 Text feature [contribute] present in test data point [True]
246 Text feature [generated] present in test data point [True]
247 Text feature [genomic] present in test data point [True]
248 Text feature [table] present in test data point [True]
251 Text feature [sequencing] present in test data point [True]
265 Text feature [somatic] present in test data point [True]
270 Text feature [novel] present in test data point [True]
271 Text feature [gene] present in test data point [True]
272 Text feature [experiments] present in test data point [True]
273 Text feature [derived] present in test data point [True]
274 Text feature [characterized] present in test data point [True]
278 Text feature [lower] present in test data point [True]
280 Text feature [normal] present in test data point [True]
282 Text feature [primary] present in test data point [True]
285 Text feature [level] present in test data point [True]
288 Text feature [caused] present in test data point [True]
289 Text feature [24] present in test data point [True]
295 Text feature [determined] present in test data point [True]
298 Text feature [activity] present in test data point [True]
299 Text feature [part] present in test data point [True]
301 Text feature [introduced] present in test data point [True]
303 Text feature [pcr] present in test data point [True]
305 Text feature [samples] present in test data point [True]
306 Text feature [non] present in test data point [True]
310 Text feature [moreover] present in test data point [True]
314 Text feature [provided] present in test data point [True]
318 Text feature [malignant] present in test data point [True]
326 Text feature [frequently] present in test data point [True]
331 Text feature [disease] present in test data point [True]
335 Text feature [significantly] present in test data point [True]
343 Text feature [still] present in test data point [True]
355 Text feature [cancer] present in test data point [True]
356 Text feature [rate] present in test data point [True]
361 Text feature [represent] present in test data point [True]
365 Text feature [distinct] present in test data point [True]
368 Text feature [established] present in test data point [True]
373 Text feature [institute] present in test data point [True]
374 Text feature [identify] present in test data point [True]

375 Text feature [negative] present in test data point [True]
```

```
376 Text feature [significant] present in test data point [True]
379 Text feature [levels] present in test data point [True]
384 Text feature [14] present in test data point [True]
392 Text feature [cellular] present in test data point [True]
400 Text feature [remain] present in test data point [True]
406 Text feature [differences] present in test data point [True]
407 Text feature [s3] present in test data point [True]
411 Text feature [11] present in test data point [True]
414 Text feature [possible] present in test data point [True]
417 Text feature [used] present in test data point [True]
420 Text feature [line] present in test data point [True]
423 Text feature [five] present in test data point [True]
429 Text feature [mutated] present in test data point [True]
432 Text feature [associated] present in test data point [True]
435 Text feature [type] present in test data point [True]
444 Text feature [green] present in test data point [True]
453 Text feature [cases] present in test data point [True]
456 Text feature [control] present in test data point [True]
457 Text feature [demonstrate] present in test data point [True]
477 Text feature [data] present in test data point [True]
499 Text feature [numbers] present in test data point [True]
502 Text feature [center] present in test data point [True]
511 Text feature [understanding] present in test data point [True]
512 Text feature [range] present in test data point [True]
514 Text feature [low] present in test data point [True]
515 Text feature [s5] present in test data point [True]
518 Text feature [s1] present in test data point [True]
523 Text feature [status] present in test data point [True]
525 Text feature [multiple] present in test data point [True]
529 Text feature [combination] present in test data point [True]
537 Text feature [72] present in test data point [True]
539 Text feature [reverse] present in test data point [True]
546 Text feature [provide] present in test data point [True]
549 Text feature [whole] present in test data point [True]
555 Text feature [responsible] present in test data point [True]
557 Text feature [number] present in test data point [True]
560 Text feature [example] present in test data point [True]
564 Text feature [stage] present in test data point [True]

568 Text feature [wild] present in test data point [True]
```

```
569 Text feature [reaction] present in test data point [True]
577 Text feature [effect] present in test data point [True]
578 Text feature [protein] present in test data point [True]
595 Text feature [majority] present in test data point [True]
597 Text feature [selective] present in test data point [True]
611 Text feature [staining] present in test data point [True]
624 Text feature [involved] present in test data point [True]
625 Text feature [overall] present in test data point [True]
636 Text feature [proteins] present in test data point [True]
637 Text feature [prolonged] present in test data point [True]
643 Text feature [left] present in test data point [True]
647 Text feature [least] present in test data point [True]
650 Text feature [extent] present in test data point [True]
662 Text feature [consistently] present in test data point [True]
677 Text feature [sanger] present in test data point [True]
682 Text feature [limited] present in test data point [True]
686 Text feature [full] present in test data point [True]
690 Text feature [marked] present in test data point [True]
691 Text feature [frequency] present in test data point [True]
697 Text feature [reduced] present in test data point [True]
703 Text feature [include] present in test data point [True]
717 Text feature [origin] present in test data point [True]
722 Text feature [virus] present in test data point [True]
731 Text feature [encoding] present in test data point [True]
732 Text feature [explain] present in test data point [True]
751 Text feature [spectrum] present in test data point [True]
757 Text feature [regions] present in test data point [True]
765 Text feature [supports] present in test data point [True]
783 Text feature [500] present in test data point [True]
790 Text feature [comparison] present in test data point [True]
805 Text feature [membrane] present in test data point [True]
809 Text feature [end] present in test data point [True]
819 Text feature [genes] present in test data point [True]
821 Text feature [establish] present in test data point [True]
824 Text feature [known] present in test data point [True]
830 Text feature [panel] present in test data point [True]
852 Text feature [events] present in test data point [True]
867 Text feature [among] present in test data point [True]

868 Text feature [site] present in test data point [True]
```

```
869 Text feature [44] present in test data point [True]
884 Text feature [encodes] present in test data point [True]
892 Text feature [05] present in test data point [True]
895 Text feature [contain] present in test data point [True]
899 Text feature [indicates] present in test data point [True]
900 Text feature [clearly] present in test data point [True]
901 Text feature [pattern] present in test data point [True]
902 Text feature [right] present in test data point [True]
904 Text feature [detection] present in test data point [True]
913 Text feature [summarized] present in test data point [True]
926 Text feature [differential] present in test data point [True]
943 Text feature [conclusion] present in test data point [True]
949 Text feature [frame] present in test data point [True]
953 Text feature [types] present in test data point [True]
954 Text feature [gain] present in test data point [True]
959 Text feature [alternative] present in test data point [True]
969 Text feature [analyses] present in test data point [True]
970 Text feature [sets] present in test data point [True]
977 Text feature [image] present in test data point [True]
987 Text feature [specifically] present in test data point [True]
991 Text feature [deletion] present in test data point [True]
995 Text feature [codon] present in test data point [True]
Out of the top  1000  features  227 are present in query point
```

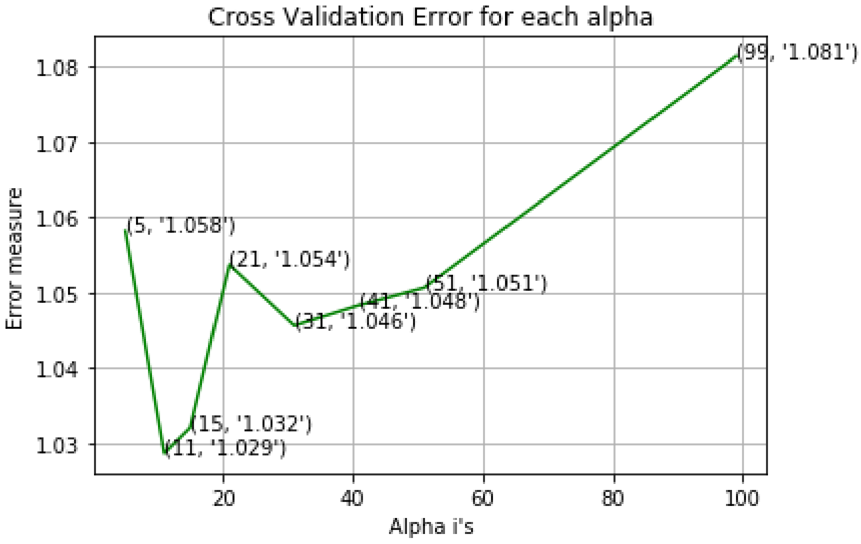
4.2. K Nearest Neighbour Classification

4.2.1. Hyper parameter tuning

In [87]:

```
# find more about KNeighborsClassifier()
# here http://scikit-learn.org/stable/modules/generated/sklearn.neighbo
rs.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights=’uniform’, algorithm=’aut
o’, leaf_size=30, p=2,
# metric=’minkowski’, metric_params=None, n_jobs=1, **kwargs)
# methods of
# fit(X, y) : Fit the model using X as training data and y as target va
lues
# predict(X):Predict the class labels for the provided data
# predict_proba(X):Return probability estimates for the test data X.
#-------------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/k-nearest-neighbors-geometric-intuition-with-a-toy-examp
le-1/
#-------------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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 = [5, 11, 15, 21, 31, 41, 51, 99]
cv_log_error_array = []
for i in alpha:
    print("for alpha =", i)
    clf = KNeighborsClassifier(n_neighbors=i)
    clf.fit(train_x_responseCoding, train_y)
    sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
    sig_clf.fit(train_x_responseCoding, train_y)
    sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
    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 = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
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_responseCoding)
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_responseCoding)
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 = 5

Log Loss : 1.0581898499383993
for alpha = 11
Log Loss : 1.028651787574871
for alpha = 15
Log Loss : 1.032050829954371
for alpha = 21
Log Loss : 1.0536903442039476
for alpha = 31
Log Loss : 1.045605619491946
for alpha = 41
Log Loss : 1.0483008763032136
for alpha = 51
Log Loss : 1.0506358917895318
for alpha = 99
Log Loss : 1.081345902653926
```



```
For values of best alpha =  11 The train log loss is: 0.65052539590497
For values of best alpha =  11 The cross validation log loss is: 1.0286
51787574871
```

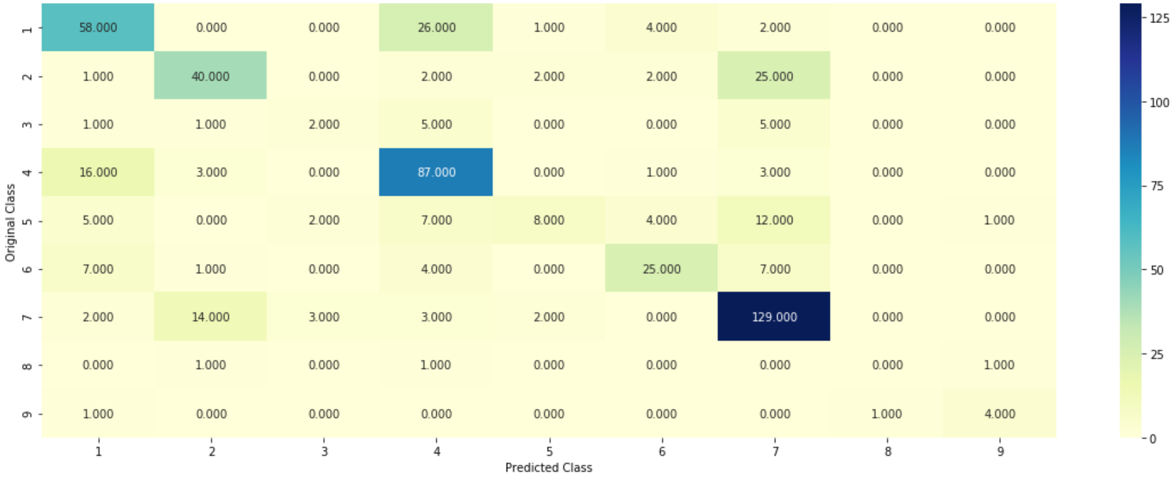
```
For values of best alpha =  11 The test log loss is: 1.0264733906932348
```

4.2.2. Testing the model with best hyper paramters

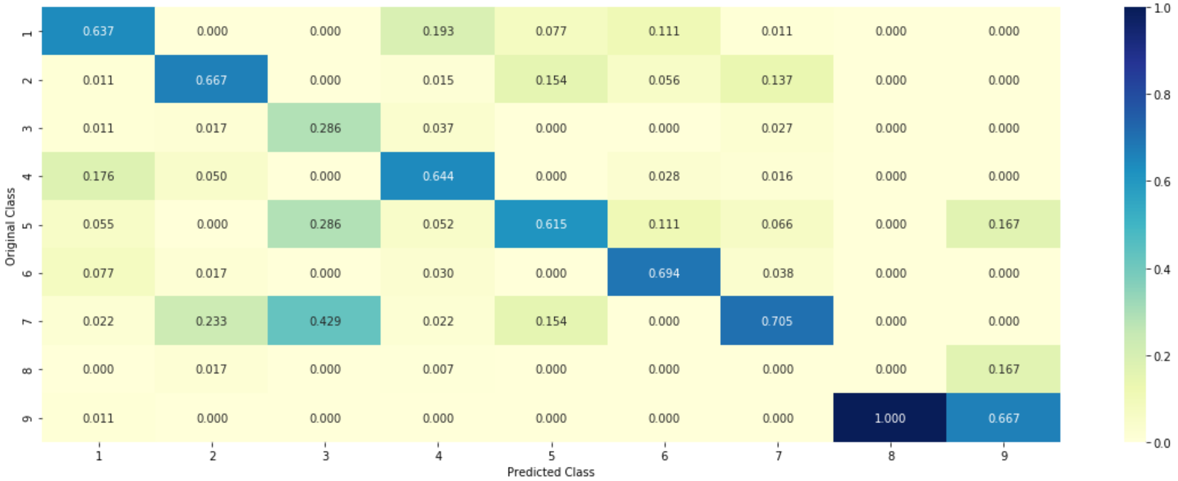
In [88]:

```
# find more about KNeighborsClassifier()
# here http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
# -------------------------
# default parameter
# KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2,
# metric='minkowski', metric_params=None, n_jobs=1, **kwargs)

# methods of
# fit(X, y) : Fit the model using X as training data and y as target values
# predict(X):Predict the class labels for the provided data
# predict_proba(X):Return probability estimates for the test data X.
#--------------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/k-nearest-neighbors-geometric-intuition-with-a-toy-example-1/
#--------------------------------------
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y, cv_x_responseCoding, cv_y, clf)
```

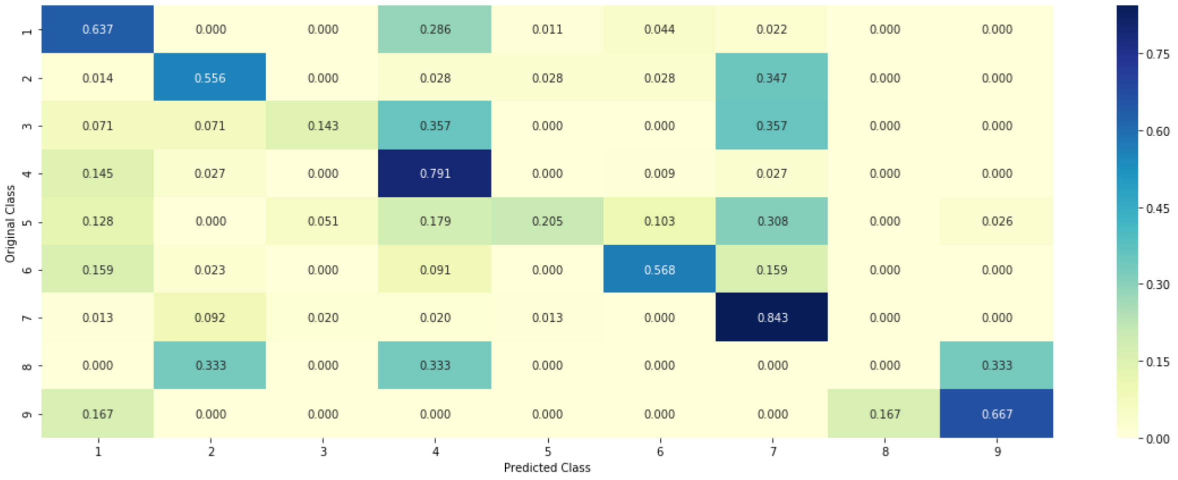
```
Log loss : 1.028651787574871
Number of mis-classified points : 0.33646616541353386

-------------------- Confusion matrix --------------------
```



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



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



4.2.3.Sample Query point -1

In [89]:

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

test_point_index = 1
predicted_cls = sig_clf.predict(test_x_responseCoding[0].reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", test_y[test_point_index])
neighbors = clf.kneighbors(test_x_responseCoding[test_point_index].resh
ape(1, -1), alpha[best_alpha])
print("The ",alpha[best_alpha]," nearest neighbours of the test points
 belongs to classes",train_y[neighbors[1][0]])
print("Fequency of nearest points :",Counter(train_y[neighbors[1][0]]))
```

```
Predicted Class : 7
Actual Class : 2
The  11  nearest neighbours of the test points belongs to classes [7 7
7 7 7 7 7 7 7 7 7]
Fequency of nearest points : Counter({7: 11})
```

4.2.4. Sample Query Point-2

In [90]:

```
clf = KNeighborsClassifier(n_neighbors=alpha[best_alpha])
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

test_point_index = 100

predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index]
.reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Actual Class :", test_y[test_point_index])
neighbors = clf.kneighbors(test_x_responseCoding[test_point_index].resh
ape(1, -1), alpha[best_alpha])
print("the k value for knn is",alpha[best_alpha],"and the nearest neigh
bours of the test points belongs to classes",train_y[neighbors[1][0]])
print("Fequency of nearest points :",Counter(train_y[neighbors[1][0]]))
```

```
Predicted Class : 4
Actual Class : 4
the k value for knn is 11 and the nearest neighbours of the test points
belongs to classes [4 2 4 4 1 4 4 4 4 4 4]
Fequency of nearest points : Counter({4: 9, 2: 1, 1: 1})
```

4.3. Logistic Regression

4.3.1. With Class balancing

4.3.1.1. Hyper paramter tuning

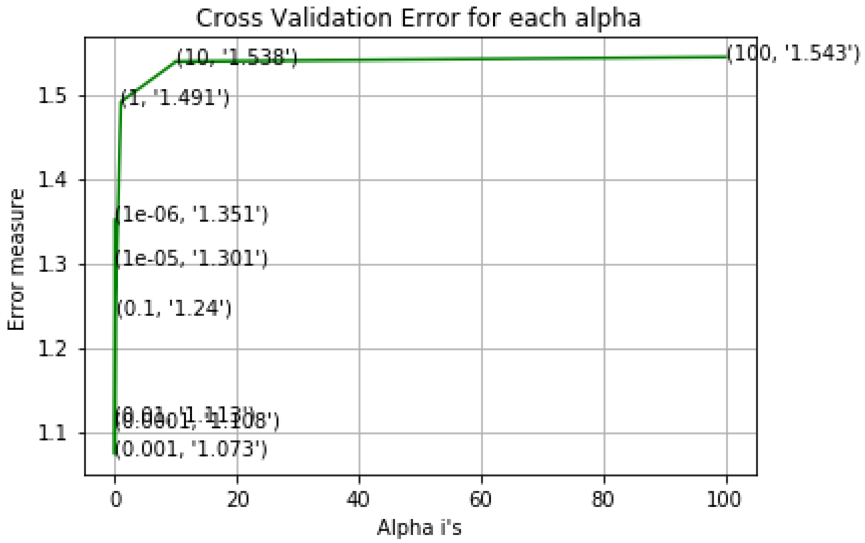
In [91]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_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 S
tochastic 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.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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], p
enalty='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 : 1.3513773498460369
for alpha = 1e-05
Log Loss : 1.3008479426043587
for alpha = 0.0001
Log Loss : 1.1080853241445796
for alpha = 0.001
Log Loss : 1.0730789682675586
for alpha = 0.01
Log Loss : 1.1131168087915757
for alpha = 0.1
Log Loss : 1.2395318186884294
for alpha = 1
Log Loss : 1.4908769917345757
for alpha = 10
Log Loss : 1.5383972814312759
for alpha = 100
Log Loss : 1.5434799269987975
```



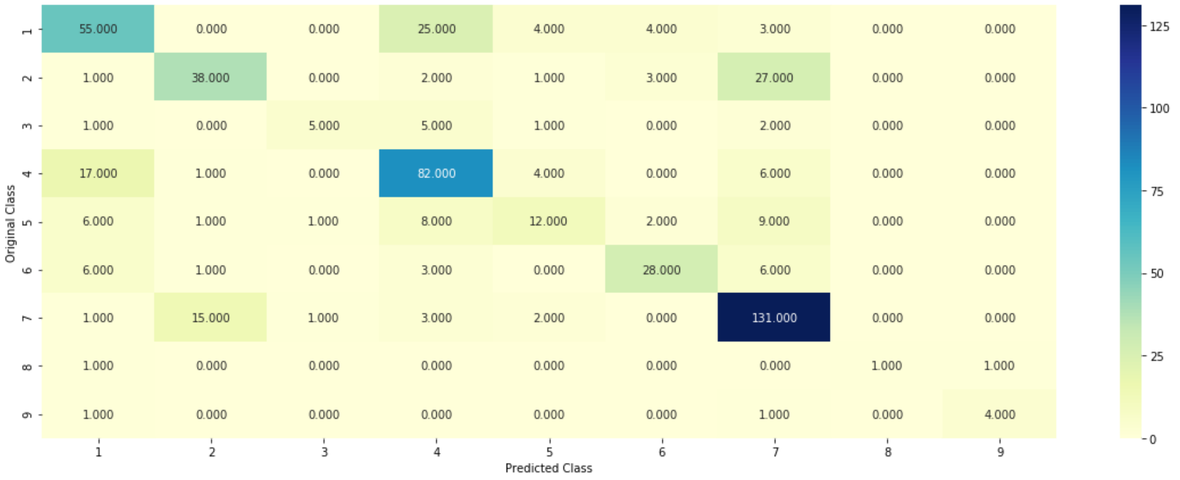
```
For values of best alpha =  0.001 The train log loss is: 0.5510171396
060384
For values of best alpha =  0.001 The cross validation log loss is:
1.0730789682675586
For values of best alpha =  0.001 The test log loss is: 0.99470930035
44093
```

4.3.1.2. Testing the model with best hyper paramters

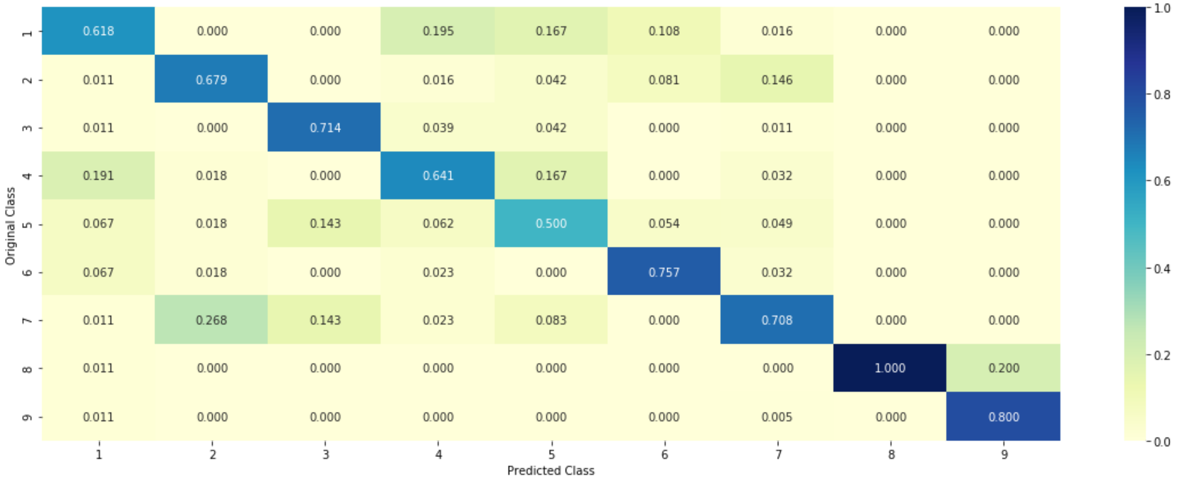
In [92]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss=’hinge’, penalty=’l2’, alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_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)
```

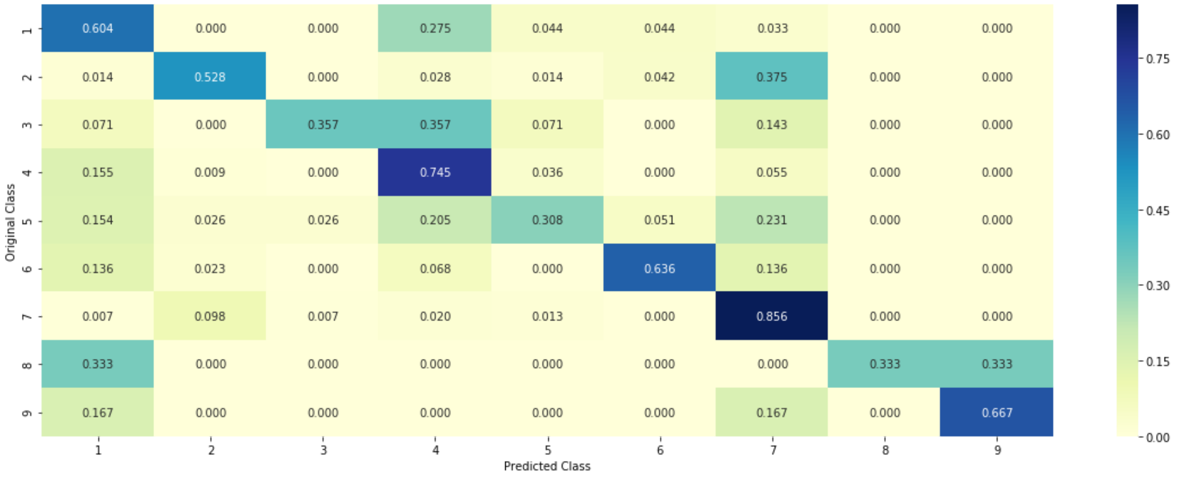
```
Log loss : 1.0730789682675586
Number of mis-classified points : 0.3308270676691729

-------------------- Confusion matrix --------------------
```



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



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



4.3.1.3. Feature Importance

In [0]:

```
def get_imp_feature_names(text, indices, removed_ind = []):
    word_present = 0
    tabulte_list = []
    incresingorder_ind = 0
    for i in indices:
        if i < train_gene_feature_onehotCoding.shape[1]:
            tabulte_list.append([incresingorder_ind, "Gene", "Yes"])
        elif i< 18:
            tabulte_list.append([incresingorder_ind,"Variation", "Yes"
])
        if ((i > 17) & (i not in removed_ind)) :
            word = train_text_features[i]
            yes_no = True if word in text.split() else False
            if yes_no:
                word_present += 1
            tabulte_list.append([incresingorder_ind,train_text_features
[i], yes_no])
        incresingorder_ind += 1
    print(word_present, "most importent features are present in our que
ry point")
    print("-"*50)
    print("The features that are most importent of the ",predicted_cls[
0]," class:")
    print (tabulate(tabulte_list, headers=["Index",'Feature name', 'Pre
sent or Not']))
```

4.3.1.3.1. Correctly Classified point

In [94]:

```
# from tabulate import tabulate
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
enalty='l2', loss='log', random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[2.000e-03 2.360e-02 7.000e-04 1.000e-0
3 2.100e-03 1.600e-03 9.666e-01
  2.000e-03 4.000e-04]]
Actual Class : 2
--------------------------------------------------
22 Text feature [activated] present in test data point [True]
24 Text feature [3t3] present in test data point [True]
31 Text feature [constitutively] present in test data point [True]
38 Text feature [proliferate] present in test data point [True]
47 Text feature [serum] present in test data point [True]
48 Text feature [oncogene] present in test data point [True]
58 Text feature [mitogen] present in test data point [True]
59 Text feature [oncogenes] present in test data point [True]
62 Text feature [stat] present in test data point [True]
63 Text feature [activation] present in test data point [True]
70 Text feature [agar] present in test data point [True]
81 Text feature [ligand] present in test data point [True]
91 Text feature [downstream] present in test data point [True]
105 Text feature [ba] present in test data point [True]
107 Text feature [activate] present in test data point [True]
111 Text feature [transform] present in test data point [True]
112 Text feature [f3] present in test data point [True]
116 Text feature [egfrs] present in test data point [True]
169 Text feature [mapk] present in test data point [True]
170 Text feature [trough] present in test data point [True]
173 Text feature [behaviors] present in test data point [True]
176 Text feature [expressing] present in test data point [True]
178 Text feature [receptors] present in test data point [True]
183 Text feature [interventions] present in test data point [True]
200 Text feature [il] present in test data point [True]
203 Text feature [transforming] present in test data point [True]
211 Text feature [hcc827] present in test data point [True]
239 Text feature [erk] present in test data point [True]
255 Text feature [transformation] present in test data point [True]
```

```
276 Text feature [manual] present in test data point [True]
296 Text feature [activating] present in test data point [True]
298 Text feature [surgically] present in test data point [True]
311 Text feature [kinase] present in test data point [True]
333 Text feature [stems] present in test data point [True]
342 Text feature [deparaffinization] present in test data point [True]
362 Text feature [inhibited] present in test data point [True]
370 Text feature [h3255] present in test data point [True]
375 Text feature [rarely] present in test data point [True]
376 Text feature [tk] present in test data point [True]
378 Text feature [soft] present in test data point [True]
392 Text feature [p753inss] present in test data point [True]
404 Text feature [phosphorylation] present in test data point [True]
430 Text feature [photographed] present in test data point [True]
438 Text feature [776] present in test data point [True]
443 Text feature [akt] present in test data point [True]
460 Text feature [epidermal] present in test data point [True]
471 Text feature [interleukin] present in test data point [True]
478 Text feature [exon18] present in test data point [True]
508 Text feature [557] present in test data point [True]
517 Text feature [displace] present in test data point [True]
524 Text feature [afatinib] present in test data point [True]
539 Text feature [transformed] present in test data point [True]
548 Text feature [doses] present in test data point [True]
565 Text feature [s9b] present in test data point [True]
570 Text feature [hki] present in test data point [True]
701 Text feature [intrinsic] present in test data point [True]
736 Text feature [superimposes] present in test data point [True]
742 Text feature [tyrosine] present in test data point [True]
746 Text feature [dell747] present in test data point [True]
747 Text feature [activates] present in test data point [True]
757 Text feature [adenocarcinomas] present in test data point [True]
787 Text feature [requisite] present in test data point [True]
788 Text feature [overexpression] present in test data point [True]
789 Text feature [played] present in test data point [True]
818 Text feature [glutamic] present in test data point [True]
822 Text feature [receptor] present in test data point [True]
835 Text feature [ic50s] present in test data point [True]
851 Text feature [oncogenic] present in test data point [True]
```

```
861 Text feature [adenocarcinoma] present in test data point [True]
898 Text feature [observational] present in test data point [True]
929 Text feature [lrea] present in test data point [True]
959 Text feature [erlotinib] present in test data point [True]
960 Text feature [s752] present in test data point [True]
962 Text feature [146] present in test data point [True]
991 Text feature [handful] present in test data point [True]
995 Text feature [guanine] present in test data point [True]
Out of the top  1000  features  76 are present in query point
```

4.3.1.3.2. Incorrectly Classified point

In [95]:

```
test_point_index = 55
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.1496 0.1327 0.0122 0.4338 0.0494 0.0
267 0.1835 0.0069 0.0052]]
Actual Class : 4
--------------------------------------------------
272 Text feature [to0] present in test data point [True]
626 Text feature [inactivating] present in test data point [True]
862 Text feature [families] present in test data point [True]
Out of the top  1000  features  3 are present in query point
```

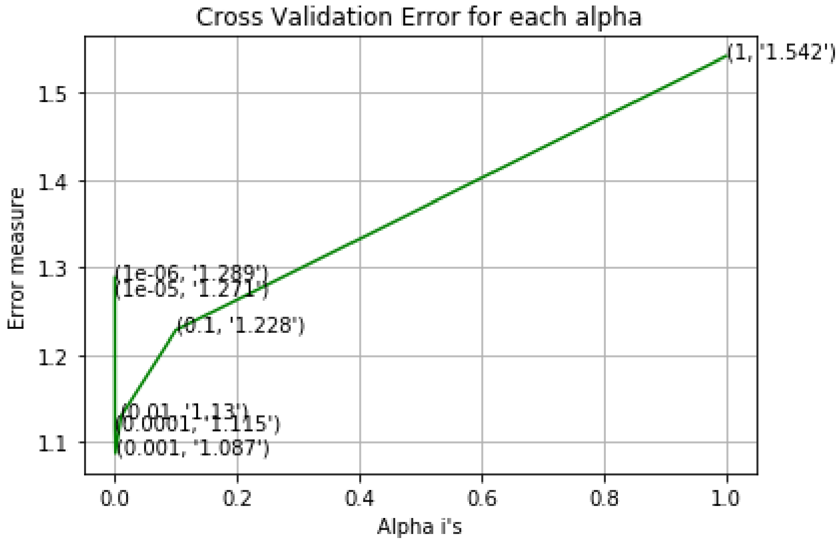
4.3.2. Without Class balancing

4.3.2.1. Hyper paramter tuning

In [96]:

```
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_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 S
tochastic 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.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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, 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 : 1.2886494374393582
for alpha = 1e-05
Log Loss : 1.2706147675753412
for alpha = 0.0001
Log Loss : 1.1150237205353215
for alpha = 0.001
Log Loss : 1.087453249546352
for alpha = 0.01
Log Loss : 1.1298235657511246
for alpha = 0.1
Log Loss : 1.2283260461917693
for alpha = 1
Log Loss : 1.541784299398985
```

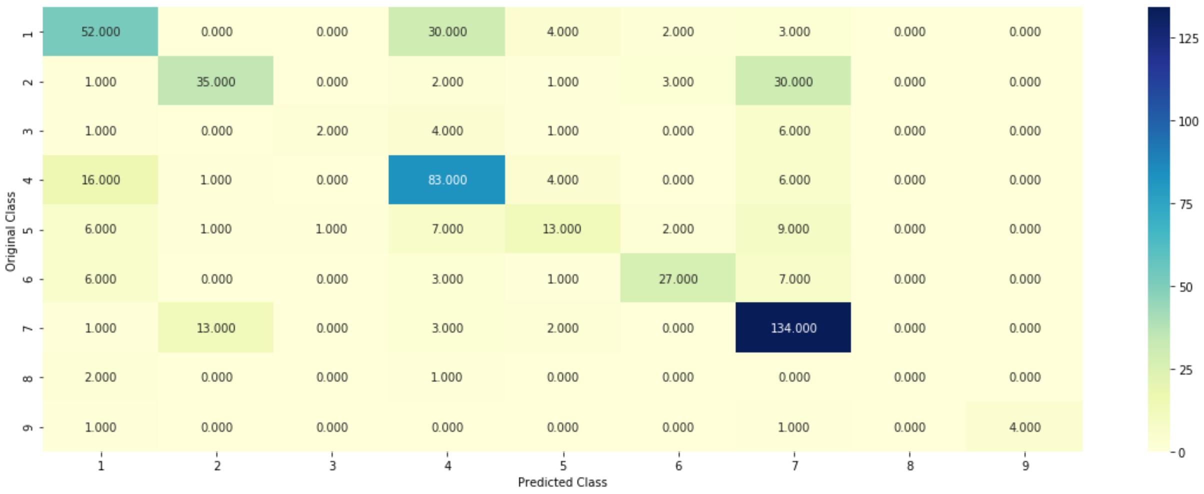


```
For values of best alpha =  0.001 The train log loss is: 0.545541979995
5438
For values of best alpha =  0.001 The cross validation log loss is: 1.0
87453249546352
For values of best alpha =  0.001 The test log loss is: 1.0187243699953
155
```

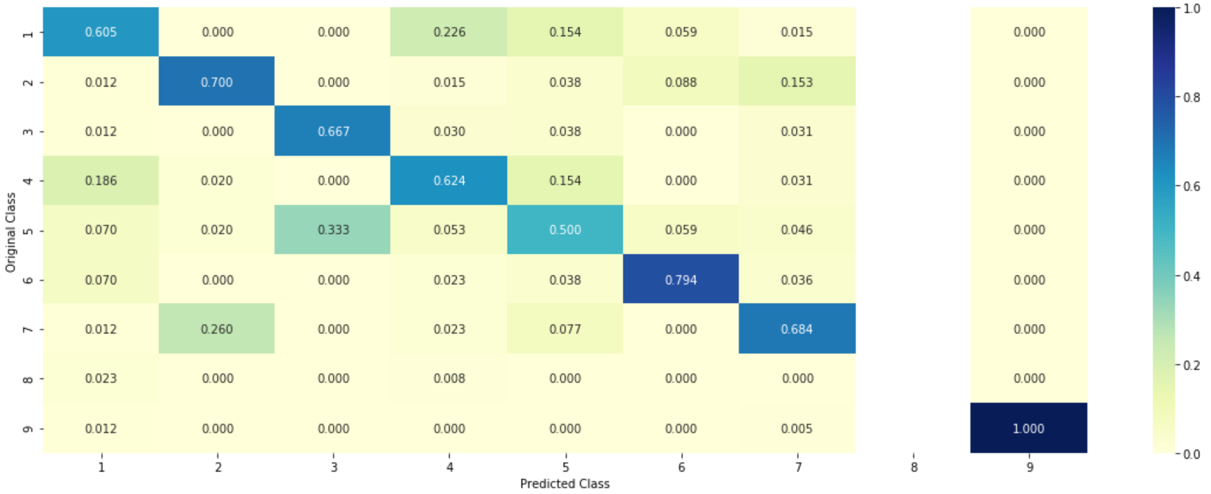
4.3.2.2. Testing model with best hyper parameters

In [97]:
```
# read more about SGDClassifier() at http://scikit-learn.org/stable/mod
ules/generated/sklearn.linear_model.SGDClassifier.html
# ------------------------------
# default parameters
# SGDClassifier(loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.1
5, fit_intercept=True, max_iter=None, tol=None,
# shuffle=True, verbose=0, epsilon=0.1, n_jobs=1, random_state=None, le
arning_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 S
tochastic 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_o
nehotCoding, cv_y, clf)
```

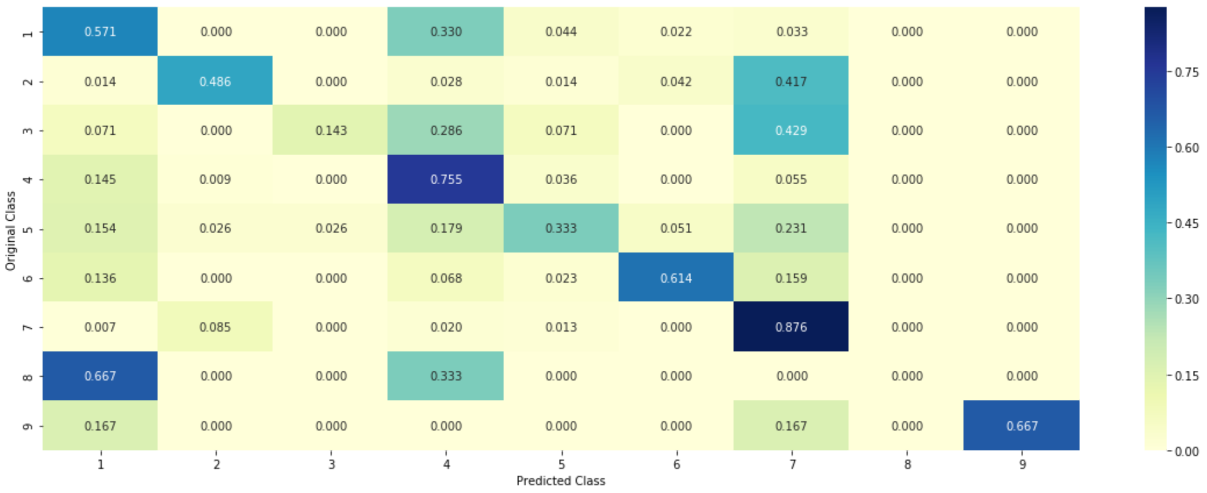
```
Log loss : 1.087453249546352
Number of mis-classified points : 0.34210526315789475

-------------------- Confusion matrix --------------------
```



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



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



4.3.2.3. Feature Importance, Correctly Classified point

```
In [98]: clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='log',
         random_state=42)
         clf.fit(train_x_onehotCoding,train_y)
         test_point_index = 100
         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_index]),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],
                              x_test['TEXT'].iloc[test_point_index],
                              x_test['Gene'].iloc[test_point_index],
                              x_test['Variation'].iloc[test_point_index],
                              no_feature)
```

```
Predicted Class : 2
Predicted Class Probabilities: [[0.1901 0.4998 0.0135 0.1589 0.0286 0.0
246 0.0752 0.005  0.0043]]
Actual Class : 4
--------------------------------------------------
100 Text feature [s4m] present in test data point [True]
160 Text feature [s4q] present in test data point [True]
161 Text feature [s4r] present in test data point [True]
208 Text feature [q33] present in test data point [True]
235 Text feature [glomulin] present in test data point [True]
239 Text feature [s4s] present in test data point [True]
248 Text feature [therapy] present in test data point [True]
362 Text feature [s4l] present in test data point [True]
431 Text feature [glomuvenous] present in test data point [True]
```

```
451 Text feature [glomuvenous] present in test data point [True]
458 Text feature [s4o] present in test data point [True]
465 Text feature [pc3] present in test data point [True]
472 Text feature [s4j] present in test data point [True]
503 Text feature [5i] present in test data point [True]
510 Text feature [bashir] present in test data point [True]
540 Text feature [4n] present in test data point [True]
574 Text feature [pin1] present in test data point [True]
587 Text feature [4m] present in test data point [True]
590 Text feature [wd40] present in test data point [True]
612 Text feature [malyukova] present in test data point [True]
614 Text feature [treatment] present in test data point [True]
629 Text feature [inuzuka] present in test data point [True]
688 Text feature [promising] present in test data point [True]
713 Text feature [5j] present in test data point [True]
737 Text feature [wertz] present in test data point [True]
806 Text feature [s4i] present in test data point [True]
826 Text feature [failure] present in test data point [True]
915 Text feature [isomerizes] present in test data point [True]
Out of the top  1000  features  27 are present in query point
```

4.3.2.4. Feature Importance, Inorrectly Classified point

In [99]:
```
test_point_index = 1
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[2.400e-03 2.560e-02 5.000e-04 1.300e-0
3 1.800e-03 1.400e-03 9.666e-01
  3.000e-04 0.000e+00]]
Actual Class : 2
--------------------------------------------------
43 Text feature [activated] present in test data point [True]
45 Text feature [3t3] present in test data point [True]
67 Text feature [constitutively] present in test data point [True]
94 Text feature [proliferate] present in test data point [True]
98 Text feature [oncogene] present in test data point [True]
102 Text feature [activation] present in test data point [True]
103 Text feature [serum] present in test data point [True]
110 Text feature [agar] present in test data point [True]
111 Text feature [mitogen] present in test data point [True]
117 Text feature [oncogenes] present in test data point [True]
122 Text feature [stat] present in test data point [True]
124 Text feature [downstream] present in test data point [True]
127 Text feature [activate] present in test data point [True]
132 Text feature [expressing] present in test data point [True]
135 Text feature [ba] present in test data point [True]
139 Text feature [ligand] present in test data point [True]
140 Text feature [f3] present in test data point [True]
157 Text feature [transform] present in test data point [True]
193 Text feature [transformation] present in test data point [True]
199 Text feature [kinase] present in test data point [True]
205 Text feature [transforming] present in test data point [True]
211 Text feature [egfrs] present in test data point [True]
232 Text feature [activating] present in test data point [True]
235 Text feature [trough] present in test data point [True]
238 Text feature [interventions] present in test data point [True]
240 Text feature [inhibited] present in test data point [True]
244 Text feature [il] present in test data point [True]
246 Text feature [phosphorylation] present in test data point [True]
256 Text feature [mapk] present in test data point [True]
264 Text feature [hcc827] present in test data point [True]
287 Text feature [manual] present in test data point [True]
293 Text feature [behaviors] present in test data point [True]
294 Text feature [receptors] present in test data point [True]
```

```
305 Text feature [erk] present in test data point [True]
312 Text feature [soft] present in test data point [True]
327 Text feature [surgically] present in test data point [True]
337 Text feature [akt] present in test data point [True]
415 Text feature [interleukin] present in test data point [True]
439 Text feature [rarely] present in test data point [True]
451 Text feature [transformed] present in test data point [True]
475 Text feature [deparaffinization] present in test data point [True]
478 Text feature [tk] present in test data point [True]
479 Text feature [epidermal] present in test data point [True]
512 Text feature [afatinib] present in test data point [True]
518 Text feature [doses] present in test data point [True]
523 Text feature [tyrosine] present in test data point [True]
533 Text feature [oncogenic] present in test data point [True]
536 Text feature [overexpression] present in test data point [True]
538 Text feature [exon18] present in test data point [True]
545 Text feature [776] present in test data point [True]
552 Text feature [557] present in test data point [True]
569 Text feature [activates] present in test data point [True]
590 Text feature [stems] present in test data point [True]
592 Text feature [photographed] present in test data point [True]
629 Text feature [s9b] present in test data point [True]
632 Text feature [h3255] present in test data point [True]
634 Text feature [blot] present in test data point [True]
645 Text feature [displace] present in test data point [True]
653 Text feature [adenocarcinoma] present in test data point [True]
686 Text feature [hki] present in test data point [True]
689 Text feature [receptor] present in test data point [True]
695 Text feature [glutamic] present in test data point [True]
707 Text feature [superimposes] present in test data point [True]
726 Text feature [p753inss] present in test data point [True]
742 Text feature [observational] present in test data point [True]
744 Text feature [braf] present in test data point [True]
745 Text feature [signaling] present in test data point [True]
792 Text feature [requisite] present in test data point [True]
794 Text feature [intrinsic] present in test data point [True]
810 Text feature [inhibitor] present in test data point [True]
811 Text feature [played] present in test data point [True]
813 Text feature [factor] present in test data point [True]
```

```
835 Text feature [guanine] present in test data point [True]
859 Text feature [2239] present in test data point [True]
870 Text feature [driven] present in test data point [True]
961 Text feature [conventional] present in test data point [True]
964 Text feature [erlotinib] present in test data point [True]
998 Text feature [adenocarcinomas] present in test data point [True]
Out of the top  1000  features  78 are present in query point
```

4.4. Linear Support Vector Machines

4.4.1. Hyper paramter tuning

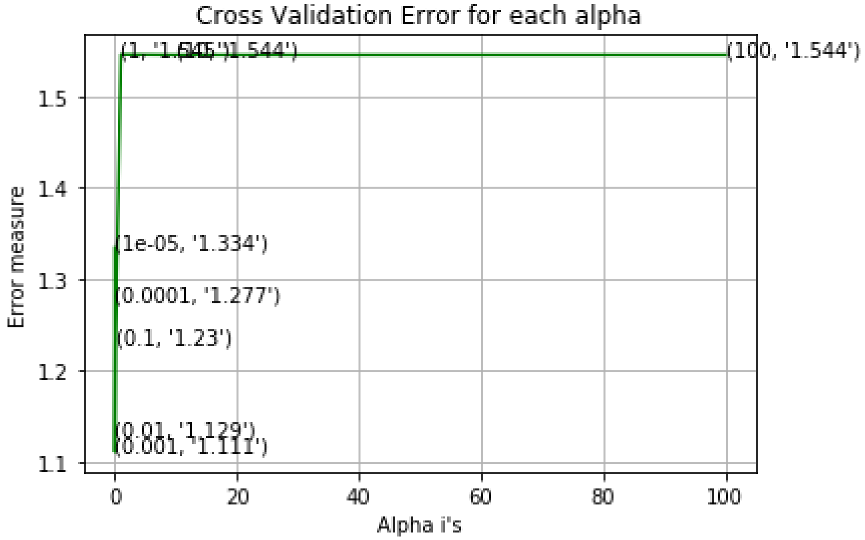
In [100]:

```
# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

# --------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', random_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight])    Fit the SVM model according to the given training data.
# predict(X)    Perform classification on samples in X.
# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/mathematical-derivation-copy-8/
# --------------------------------



# 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, metho
d='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(-5, 3)]
cv_log_error_array = []
for i in alpha:
    print("for C =", i)
#     clf = SVC(C=i,kernel='linear',probability=True, class_weight='bal
anced')
    clf = SGDClassifier( class_weight='balanced', alpha=i, penalty='l2'
, loss='hinge', 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()
```

```python
best_alpha = np.argmin(cv_log_error_array)
# clf = SVC(C=i,kernel='linear',probability=True, class_weight='balanced')
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], penalty='l2', loss='hinge', 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 C = 1e-05
Log Loss : 1.3339426096207683
for C = 0.0001
Log Loss : 1.2765716758633658
for C = 0.001
Log Loss : 1.111257872246351
for C = 0.01
Log Loss : 1.128665272473871
for C = 0.1
Log Loss : 1.2303411853345119
for C = 1
```

```
Log Loss : 1.5448164003187426
for C = 10
Log Loss : 1.5443359985945422
for C = 100
Log Loss : 1.5443358842428505
```



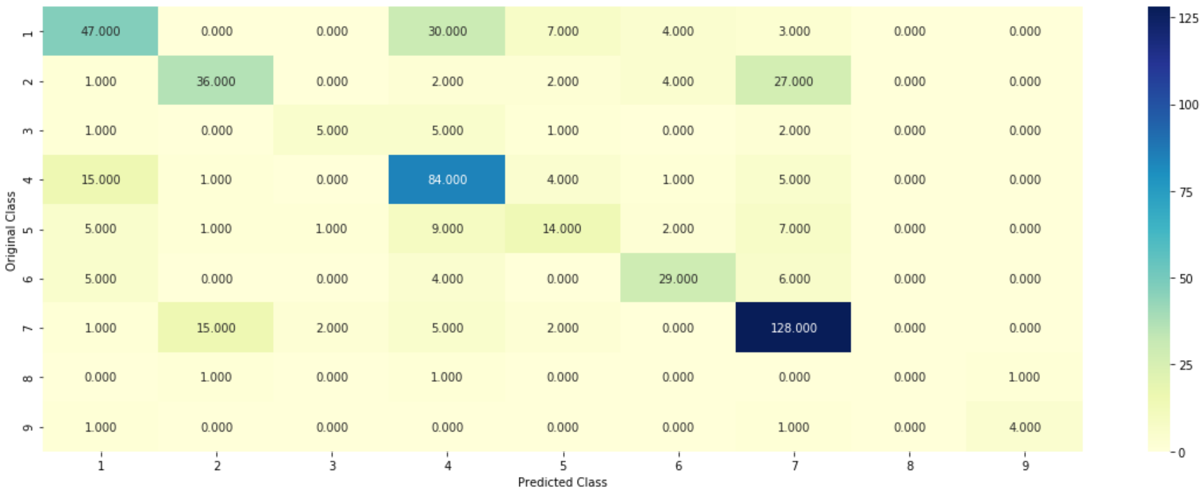
```
For values of best alpha =  0.001 The train log loss is: 0.587853712603
1901
For values of best alpha =  0.001 The cross validation log loss is: 1.1
11257872246351
For values of best alpha =  0.001 The test log loss is: 1.0837300854668
97
```

4.4.2. Testing model with best hyper parameters

In [101]:

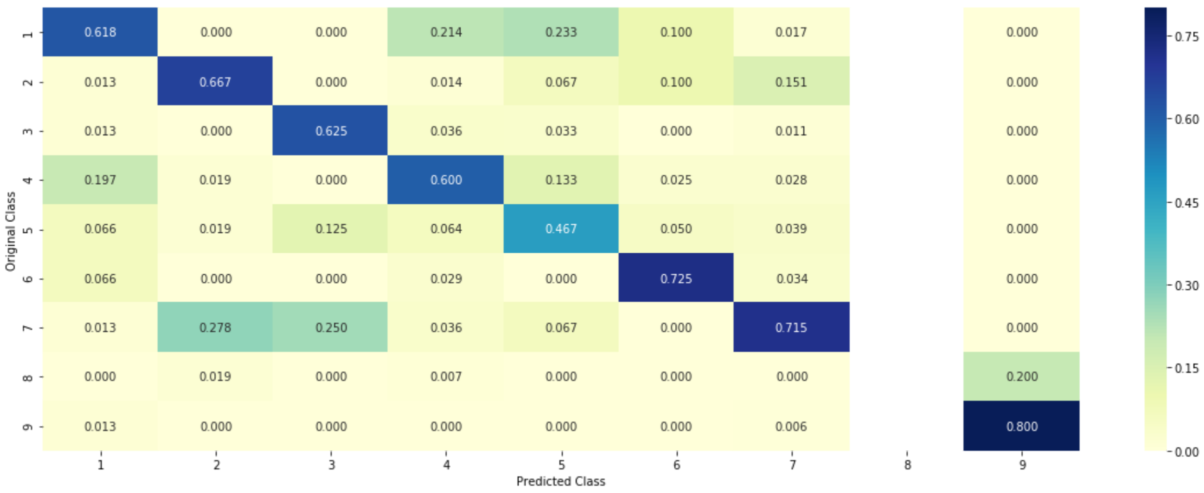
```
# read more about support vector machines with linear kernals here htt
p://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

# --------------------------------
# default parameters
# SVC(C=1.0, kernel=’rbf’, degree=3, gamma=’auto’, coef0=0.0, shrinking
=True, probability=False, tol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decisi
on_function_shape='ovr', random_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight])    Fit the SVM model according to the give
n training data.
# predict(X)    Perform classification on samples in X.
# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/mathematical-derivation-copy-8/
# --------------------------------



# clf = SVC(C=alpha[best_alpha],kernel='linear',probability=True, class
_weight='balanced')
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge'
, random_state=42,class_weight='balanced')
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_on
ehotCoding,cv_y, clf)
```

```
Log loss : 1.111257872246351
Number of mis-classified points : 0.34774436090225563

-------------------- Confusion matrix --------------------
```

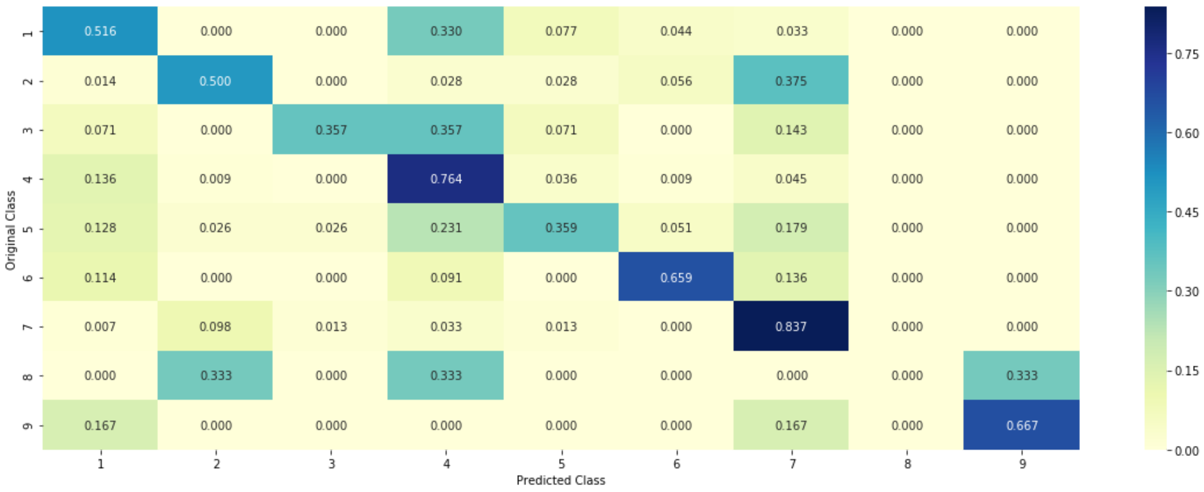


	1	2	3	4	5	6	7	8	9
1	47.000	0.000	0.000	30.000	7.000	4.000	3.000	0.000	0.000
2	1.000	36.000	0.000	2.000	2.000	4.000	27.000	0.000	0.000
3	1.000	0.000	5.000	5.000	1.000	0.000	2.000	0.000	0.000
4	15.000	1.000	0.000	84.000	4.000	1.000	5.000	0.000	0.000
5	5.000	1.000	1.000	9.000	14.000	2.000	7.000	0.000	0.000
6	5.000	0.000	0.000	4.000	0.000	29.000	6.000	0.000	0.000
7	1.000	15.000	2.000	5.000	2.000	0.000	128.000	0.000	0.000
8	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000
9	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	4.000

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



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



4.3.3. Feature Importance

4.3.3.1. For Correctly classified point

In [102]:

```
clf = SGDClassifier(alpha=alpha[best_alpha], penalty='l2', loss='hinge'
, random_state=42)
clf.fit(train_x_onehotCoding,train_y)
test_point_index = 1
# test_point_index = 100
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0229 0.036  0.0034 0.0209 0.0158 0.0
091 0.8858 0.0034 0.0028]]
Actual Class : 2
--------------------------------------------------
176 Text feature [activated] present in test data point [True]
187 Text feature [3t3] present in test data point [True]
```

```
229 Text feature [braf] present in test data point [True]
256 Text feature [oncogene] present in test data point [True]
290 Text feature [constitutively] present in test data point [True]
350 Text feature [oncogenes] present in test data point [True]
355 Text feature [s768i] present in test data point [True]
356 Text feature [agar] present in test data point [True]
358 Text feature [proliferate] present in test data point [True]
386 Text feature [trough] present in test data point [True]
388 Text feature [activation] present in test data point [True]
396 Text feature [surgically] present in test data point [True]
402 Text feature [interventions] present in test data point [True]
405 Text feature [hki] present in test data point [True]
412 Text feature [stat] present in test data point [True]
413 Text feature [activate] present in test data point [True]
437 Text feature [deparaffinization] present in test data point [True]
443 Text feature [transformation] present in test data point [True]
445 Text feature [expressing] present in test data point [True]
446 Text feature [ba] present in test data point [True]
448 Text feature [f3] present in test data point [True]
458 Text feature [egfrs] present in test data point [True]
464 Text feature [hcc827] present in test data point [True]
493 Text feature [downstream] present in test data point [True]
506 Text feature [serum] present in test data point [True]
552 Text feature [transforming] present in test data point [True]
558 Text feature [inhibited] present in test data point [True]
593 Text feature [manual] present in test data point [True]
597 Text feature [heterogeneous] present in test data point [True]
602 Text feature [136] present in test data point [True]
605 Text feature [spectrumgreen] present in test data point [True]
615 Text feature [afatinib] present in test data point [True]
659 Text feature [transform] present in test data point [True]
680 Text feature [adenocarcinoma] present in test data point [True]
698 Text feature [ligand] present in test data point [True]
705 Text feature [driven] present in test data point [True]
708 Text feature [ipass] present in test data point [True]
723 Text feature [146] present in test data point [True]
727 Text feature [2239] present in test data point [True]
751 Text feature [behaviors] present in test data point [True]
772 Text feature [rarer] present in test data point [True]
```

```
795 Text feature [kinase] present in test data point [True]
811 Text feature [month] present in test data point [True]
812 Text feature [asv] present in test data point [True]
824 Text feature [776] present in test data point [True]
833 Text feature [superimposes] present in test data point [True]
858 Text feature [akt] present in test data point [True]
861 Text feature [blot] present in test data point [True]
864 Text feature [guanine] present in test data point [True]
885 Text feature [il] present in test data point [True]
894 Text feature [disfavours] present in test data point [True]
905 Text feature [272] present in test data point [True]
908 Text feature [soft] present in test data point [True]
919 Text feature [transformed] present in test data point [True]
923 Text feature [phosphorylation] present in test data point [True]
958 Text feature [observational] present in test data point [True]
967 Text feature [overexpression] present in test data point [True]
972 Text feature [3a] present in test data point [True]
976 Text feature [activates] present in test data point [True]
983 Text feature [rarely] present in test data point [True]
985 Text feature [exon18] present in test data point [True]
Out of the top  1000  features  61 are present in query point
```

4.3.3.2. For Incorrectly classified point

In [103]:

```python
test_point_index = 31
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_index]),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],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 9
Predicted Class Probabilities: [[0.0312 0.0292 0.0064 0.0319 0.0173 0.0
14  0.3062 0.0046 0.5592]]
Actual Class : 7
--------------------------------------------------
Out of the top  1000  features  0 are present in query point
```

4.5 Random Forest Classifier

4.5.1. Hyper paramter tuning (With One hot Encoding)

In [104]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the give
n training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
```

```
online/lessons/random-forest-and-their-construction-2/
# --------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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 = [100,200,500,1000,2000]
max_depth = [5, 10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini',
max_depth=j, random_state=42, n_jobs=-1)
        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()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ra
```

```
vel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/2)],max_depth[int(i%2)],str(txt)), (featur
es[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 = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], cri
terion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42,
 n_jobs=-1)
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 estimator = ',
      alpha[int(best_alpha/2)],
      "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 estimator = ',
      alpha[int(best_alpha/2)],
      "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 estimator = ',
      alpha[int(best_alpha/2)],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

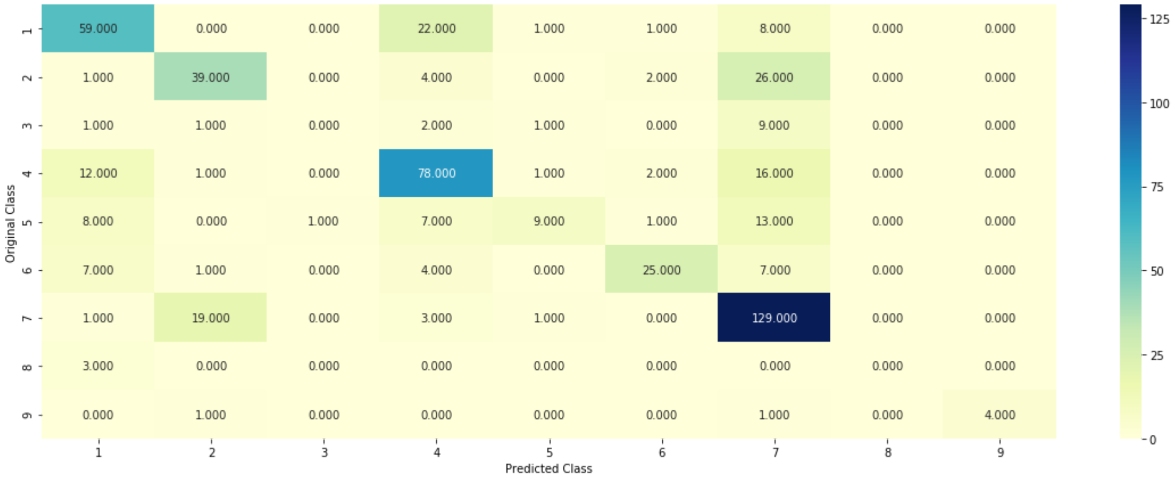
```
for n_estimators = 100 and max depth =  5
Log Loss : 1.2269755900877062
for n_estimators = 100 and max depth =  10
for n_estimators = 100 and max depth =  10
Log Loss : 1.1889784452028322
for n_estimators = 200 and max depth =  5
Log Loss : 1.2149804131913267
for n_estimators = 200 and max depth =  10

Log Loss : 1.1800898401230269
for n_estimators = 500 and max depth =  5
Log Loss : 1.201600568238576
for n_estimators = 500 and max depth =  10
Log Loss : 1.1685088697399197
for n_estimators = 1000 and max depth =  5
Log Loss : 1.2034523742080965
for n_estimators = 1000 and max depth =  10
Log Loss : 1.1668315552775221
for n_estimators = 2000 and max depth =  5
Log Loss : 1.2026633433767522
for n_estimators = 2000 and max depth =  10
Log Loss : 1.1653689130366385
For values of best estimator =  2000 The train log loss is: 0.662489287
8373628
For values of best estimator =  2000 The cross validation log loss is:
1.1653689130366387
For values of best estimator =  2000 The test log loss is: 1.1407954648
805678
```

4.5.2. Testing model with best hyper parameters (One Hot Encoding)

In [105]:

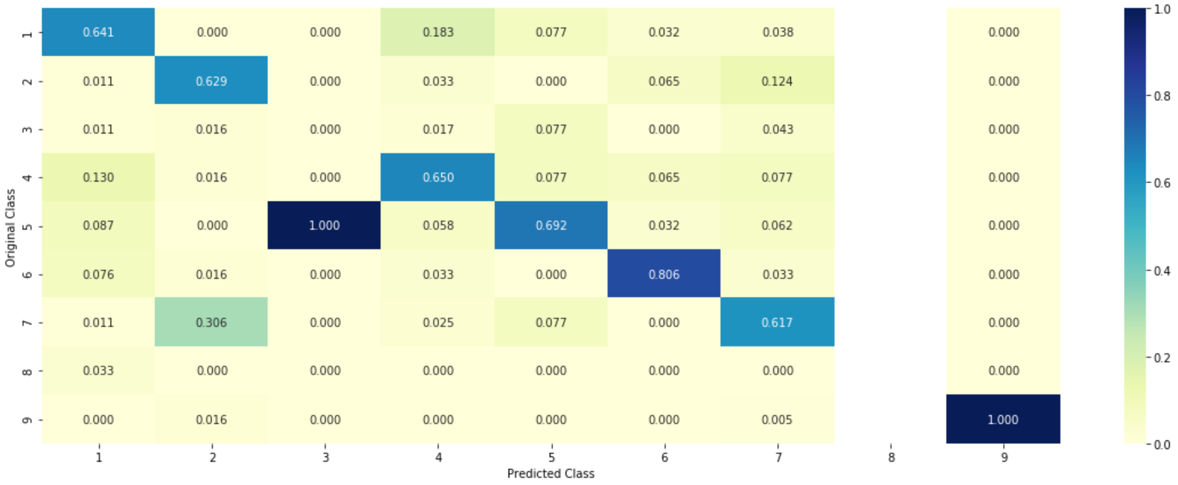
```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)
# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the give
n training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], cri
terion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42,
 n_jobs=-1)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y,cv_x_on
ehotCoding,cv_y, clf)
```

```
Log loss : 1.1653689130366385
Number of mis-classified points : 0.35526315789473684

-------------------- Confusion matrix --------------------
```



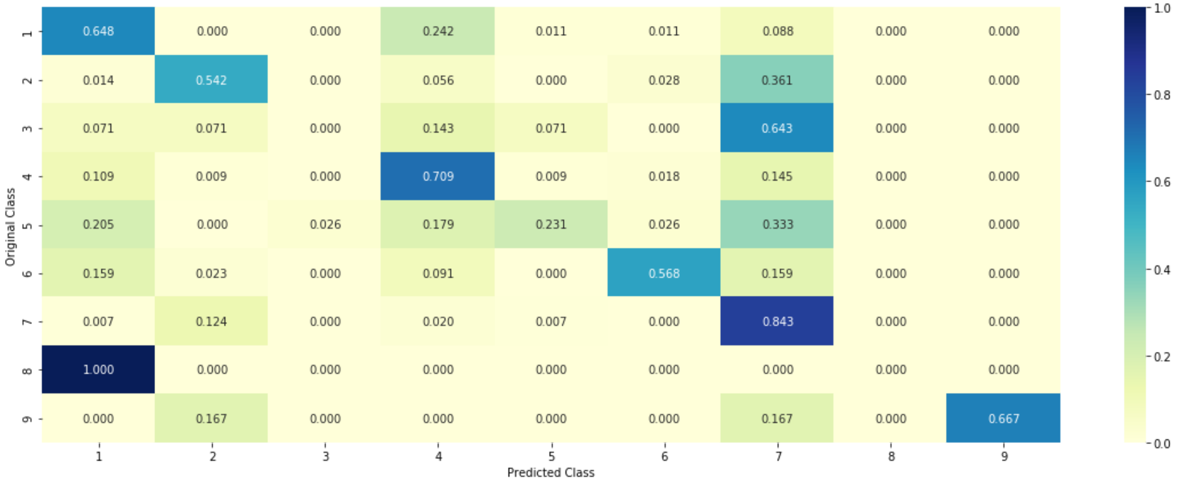
	1	2	3	4	5	6	7	8	9
1	59.000	0.000	0.000	22.000	1.000	1.000	8.000	0.000	0.000
2	1.000	39.000	0.000	4.000	0.000	2.000	26.000	0.000	0.000
3	1.000	1.000	0.000	2.000	1.000	0.000	9.000	0.000	0.000
4	12.000	1.000	0.000	78.000	1.000	2.000	16.000	0.000	0.000
5	8.000	0.000	1.000	7.000	9.000	1.000	13.000	0.000	0.000
6	7.000	1.000	0.000	4.000	0.000	25.000	7.000	0.000	0.000
7	1.000	19.000	0.000	3.000	1.000	0.000	129.000	0.000	0.000
8	3.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	4.000

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



	1	2	3	4	5	6	7	8	9
1	0.641	0.000	0.000	0.183	0.077	0.032	0.038		0.000
2	0.011	0.629	0.000	0.033	0.000	0.065	0.124		0.000
3	0.011	0.016	0.000	0.017	0.077	0.000	0.043		0.000
4	0.130	0.016	0.000	0.650	0.077	0.065	0.077		0.000
5	0.087	0.000	1.000	0.058	0.692	0.032	0.062		0.000
6	0.076	0.016	0.000	0.033	0.000	0.806	0.033		0.000
7	0.011	0.306	0.000	0.025	0.077	0.000	0.617		0.000
8	0.033	0.000	0.000	0.000	0.000	0.000	0.000		0.000
9	0.000	0.016	0.000	0.000	0.000	0.000	0.005		1.000

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



	1	2	3	4	5	6	7	8	9
1	0.648	0.000	0.000	0.242	0.011	0.011	0.088	0.000	0.000
2	0.014	0.542	0.000	0.056	0.000	0.028	0.361	0.000	0.000
3	0.071	0.071	0.000	0.143	0.071	0.000	0.643	0.000	0.000
4	0.109	0.009	0.000	0.709	0.009	0.018	0.145	0.000	0.000
5	0.205	0.000	0.026	0.179	0.231	0.026	0.333	0.000	0.000
6	0.159	0.023	0.000	0.091	0.000	0.568	0.159	0.000	0.000
7	0.007	0.124	0.000	0.020	0.007	0.000	0.843	0.000	0.000
8	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.167	0.000	0.000	0.000	0.000	0.167	0.000	0.667

4.5.3. Feature Importance

4.5.3.1. Correctly Classified point

In [106]:

```
# test_point_index = 10
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/2)], cri
terion='gini', max_depth=max_depth[int(best_alpha%2)], random_state=42,
 n_jobs=-1)
clf.fit(train_x_onehotCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_onehotCoding, train_y)

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

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0294 0.1609 0.0136 0.0203 0.0289 0.0
279 0.7121 0.0038 0.0032]]
Actual Class : 2
--------------------------------------------------
0 Text feature [kinase] present in test data point [True]
1 Text feature [activating] present in test data point [True]
2 Text feature [inhibitors] present in test data point [True]
3 Text feature [oncogenic] present in test data point [True]
4 Text feature [activation] present in test data point [True]
5 Text feature [inhibitor] present in test data point [True]
6 Text feature [tyrosine] present in test data point [True]
7 Text feature [phosphorylation] present in test data point [True]
8 Text feature [suppressor] present in test data point [True]
```

```
9 Text feature [missense] present in test data point [True]
10 Text feature [treatment] present in test data point [True]
12 Text feature [signaling] present in test data point [True]
14 Text feature [activated] present in test data point [True]
15 Text feature [function] present in test data point [True]
16 Text feature [downstream] present in test data point [True]
17 Text feature [growth] present in test data point [True]
18 Text feature [kinases] present in test data point [True]
19 Text feature [receptor] present in test data point [True]
20 Text feature [therapy] present in test data point [True]
21 Text feature [treated] present in test data point [True]
22 Text feature [loss] present in test data point [True]
23 Text feature [erk] present in test data point [True]
24 Text feature [akt] present in test data point [True]
26 Text feature [functional] present in test data point [True]
27 Text feature [amplification] present in test data point [True]
28 Text feature [protein] present in test data point [True]
29 Text feature [transforming] present in test data point [True]
30 Text feature [f3] present in test data point [True]
31 Text feature [constitutively] present in test data point [True]
32 Text feature [stability] present in test data point [True]
33 Text feature [extracellular] present in test data point [True]
34 Text feature [activate] present in test data point [True]
35 Text feature [drug] present in test data point [True]
37 Text feature [factor] present in test data point [True]
38 Text feature [therapeutic] present in test data point [True]
39 Text feature [cells] present in test data point [True]
40 Text feature [expressing] present in test data point [True]
43 Text feature [proliferation] present in test data point [True]
44 Text feature [months] present in test data point [True]
45 Text feature [serum] present in test data point [True]
46 Text feature [3t3] present in test data point [True]
47 Text feature [ba] present in test data point [True]
50 Text feature [resistance] present in test data point [True]
51 Text feature [mek] present in test data point [True]
53 Text feature [inhibited] present in test data point [True]
55 Text feature [nsclc] present in test data point [True]
57 Text feature [oncogene] present in test data point [True]
58 Text feature [proteins] present in test data point [True]
```

```
60 Text feature [ic50] present in test data point [True]
61 Text feature [potential] present in test data point [True]
62 Text feature [ligand] present in test data point [True]
63 Text feature [predicted] present in test data point [True]
64 Text feature [lines] present in test data point [True]
65 Text feature [phosphorylated] present in test data point [True]
67 Text feature [trials] present in test data point [True]
68 Text feature [patients] present in test data point [True]
69 Text feature [advanced] present in test data point [True]
70 Text feature [mitogen] present in test data point [True]
72 Text feature [cell] present in test data point [True]
73 Text feature [egfr] present in test data point [True]
74 Text feature [respond] present in test data point [True]
76 Text feature [trial] present in test data point [True]
78 Text feature [survival] present in test data point [True]
79 Text feature [clinical] present in test data point [True]
83 Text feature [inhibition] present in test data point [True]
87 Text feature [il] present in test data point [True]
89 Text feature [efficacy] present in test data point [True]
92 Text feature [response] present in test data point [True]
93 Text feature [resistant] present in test data point [True]
94 Text feature [effective] present in test data point [True]
95 Text feature [transformation] present in test data point [True]
96 Text feature [sensitive] present in test data point [True]
99 Text feature [patient] present in test data point [True]
100 Text feature [likelihood] present in test data point [True]
101 Text feature [active] present in test data point [True]
102 Text feature [pathway] present in test data point [True]
103 Text feature [expression] present in test data point [True]
106 Text feature [atp] present in test data point [True]
107 Text feature [acquired] present in test data point [True]
111 Text feature [sensitivity] present in test data point [True]
112 Text feature [unclassified] present in test data point [True]
114 Text feature [mapk] present in test data point [True]
115 Text feature [affect] present in test data point [True]
119 Text feature [dose] present in test data point [True]
120 Text feature [dna] present in test data point [True]
121 Text feature [sequence] present in test data point [True]
123 Text feature [presence] present in test data point [True]
```

```
124 Text feature [binding] present in test data point [True]
125 Text feature [kit] present in test data point [True]
126 Text feature [classified] present in test data point [True]
128 Text feature [tki] present in test data point [True]
129 Text feature [harboring] present in test data point [True]
131 Text feature [terminal] present in test data point [True]
133 Text feature [transform] present in test data point [True]
134 Text feature [information] present in test data point [True]
135 Text feature [weeks] present in test data point [True]
136 Text feature [inactivation] present in test data point [True]
138 Text feature [assays] present in test data point [True]
140 Text feature [alignments] present in test data point [True]
141 Text feature [core] present in test data point [True]
143 Text feature [ability] present in test data point [True]
144 Text feature [gastrointestinal] present in test data point [True]
145 Text feature [transfected] present in test data point [True]
146 Text feature [sequencing] present in test data point [True]
148 Text feature [useful] present in test data point [True]
149 Text feature [clinically] present in test data point [True]
150 Text feature [daily] present in test data point [True]
152 Text feature [odds] present in test data point [True]
153 Text feature [preclinical] present in test data point [True]
155 Text feature [tagged] present in test data point [True]
156 Text feature [length] present in test data point [True]
161 Text feature [days] present in test data point [True]
162 Text feature [assay] present in test data point [True]
163 Text feature [achieved] present in test data point [True]
164 Text feature [tumors] present in test data point [True]
166 Text feature [known] present in test data point [True]
168 Text feature [progression] present in test data point [True]
169 Text feature [conserved] present in test data point [True]
170 Text feature [activity] present in test data point [True]
171 Text feature [expected] present in test data point [True]
172 Text feature [independent] present in test data point [True]
173 Text feature [benefit] present in test data point [True]
174 Text feature [responses] present in test data point [True]
176 Text feature [mutants] present in test data point [True]
179 Text feature [used] present in test data point [True]
181 Text feature [contrast] present in test data point [True]
```

```
182 Text feature [lung] present in test data point [True]
184 Text feature [combined] present in test data point [True]
185 Text feature [concentrations] present in test data point [True]
186 Text feature [molecular] present in test data point [True]
187 Text feature [use] present in test data point [True]
188 Text feature [predictions] present in test data point [True]
189 Text feature [tkis] present in test data point [True]
190 Text feature [hybridization] present in test data point [True]
191 Text feature [median] present in test data point [True]
192 Text feature [anchorage] present in test data point [True]
194 Text feature [recently] present in test data point [True]
195 Text feature [26] present in test data point [True]
196 Text feature [nih] present in test data point [True]
197 Text feature [receptors] present in test data point [True]
200 Text feature [antibodies] present in test data point [True]
203 Text feature [wild] present in test data point [True]
204 Text feature [likely] present in test data point [True]
205 Text feature [signal] present in test data point [True]
207 Text feature [deletion] present in test data point [True]
208 Text feature [arrest] present in test data point [True]
209 Text feature [biopsy] present in test data point [True]
210 Text feature [hours] present in test data point [True]
211 Text feature [based] present in test data point [True]
213 Text feature [study] present in test data point [True]
214 Text feature [57] present in test data point [True]
216 Text feature [interaction] present in test data point [True]
217 Text feature [interleukin] present in test data point [True]
218 Text feature [classification] present in test data point [True]
219 Text feature [controls] present in test data point [True]
223 Text feature [14] present in test data point [True]
225 Text feature [absence] present in test data point [True]
226 Text feature [pathways] present in test data point [True]
231 Text feature [personal] present in test data point [True]
233 Text feature [primary] present in test data point [True]
234 Text feature [metastatic] present in test data point [True]
235 Text feature [epidermal] present in test data point [True]
238 Text feature [mechanism] present in test data point [True]
242 Text feature [risk] present in test data point [True]
243 Text feature [partial] present in test data point [True]
```

```
245 Text feature [surface] present in test data point [True]
246 Text feature [therapeutics] present in test data point [True]
247 Text feature [35] present in test data point [True]
248 Text feature [21] present in test data point [True]
250 Text feature [type] present in test data point [True]
251 Text feature [enhanced] present in test data point [True]
253 Text feature [predictive] present in test data point [True]
254 Text feature [domain] present in test data point [True]
256 Text feature [mutant] present in test data point [True]
257 Text feature [49] present in test data point [True]
258 Text feature [1a] present in test data point [True]
260 Text feature [12] present in test data point [True]
265 Text feature [catalytic] present in test data point [True]
266 Text feature [favor] present in test data point [True]
268 Text feature [evolutionary] present in test data point [True]
269 Text feature [family] present in test data point [True]
270 Text feature [methods] present in test data point [True]
271 Text feature [106] present in test data point [True]
273 Text feature [transduction] present in test data point [True]
275 Text feature [doses] present in test data point [True]
276 Text feature [experiments] present in test data point [True]
277 Text feature [large] present in test data point [True]
278 Text feature [32] present in test data point [True]
280 Text feature [probability] present in test data point [True]
282 Text feature [previously] present in test data point [True]
284 Text feature [approved] present in test data point [True]
292 Text feature [changes] present in test data point [True]
293 Text feature [novel] present in test data point [True]
294 Text feature [increased] present in test data point [True]
296 Text feature [2b] present in test data point [True]
298 Text feature [stop] present in test data point [True]
299 Text feature [therapies] present in test data point [True]
300 Text feature [soft] present in test data point [True]
301 Text feature [region] present in test data point [True]
302 Text feature [duration] present in test data point [True]
303 Text feature [acid] present in test data point [True]
305 Text feature [although] present in test data point [True]
306 Text feature [pi3k] present in test data point [True]
307 Text feature [general] present in test data point [True]
```

```
308 Text feature [gene] present in test data point [True]
309 Text feature [induced] present in test data point [True]
311 Text feature [database] present in test data point [True]
313 Text feature [potentially] present in test data point [True]
314 Text feature [compared] present in test data point [True]
315 Text feature [alk] present in test data point [True]
316 Text feature [significance] present in test data point [True]
317 Text feature [22] present in test data point [True]
318 Text feature [western] present in test data point [True]
319 Text feature [next] present in test data point [True]
320 Text feature [demonstrated] present in test data point [True]
321 Text feature [well] present in test data point [True]
322 Text feature [shown] present in test data point [True]
323 Text feature [structural] present in test data point [True]
327 Text feature [however] present in test data point [True]
328 Text feature [system] present in test data point [True]
329 Text feature [harbored] present in test data point [True]
331 Text feature [potent] present in test data point [True]
332 Text feature [obtained] present in test data point [True]
334 Text feature [cancer] present in test data point [True]
335 Text feature [vitro] present in test data point [True]
337 Text feature [indicated] present in test data point [True]
339 Text feature [positive] present in test data point [True]
340 Text feature [11] present in test data point [True]
342 Text feature [data] present in test data point [True]
343 Text feature [tested] present in test data point [True]
344 Text feature [2a] present in test data point [True]
347 Text feature [stably] present in test data point [True]
348 Text feature [substrate] present in test data point [True]
349 Text feature [essential] present in test data point [True]
350 Text feature [pretreatment] present in test data point [True]
352 Text feature [results] present in test data point [True]
353 Text feature [43] present in test data point [True]
354 Text feature [higher] present in test data point [True]
359 Text feature [rearrangements] present in test data point [True]
361 Text feature [addition] present in test data point [True]
362 Text feature [10] present in test data point [True]
366 Text feature [wt] present in test data point [True]
367 Text feature [involved] present in test data point [True]
```

```
369 Text feature [time] present in test data point [True]
371 Text feature [medium] present in test data point [True]
372 Text feature [105] present in test data point [True]
373 Text feature [residues] present in test data point [True]
375 Text feature [control] present in test data point [True]
377 Text feature [18] present in test data point [True]
379 Text feature [within] present in test data point [True]
381 Text feature [oncogenes] present in test data point [True]
382 Text feature [table] present in test data point [True]
384 Text feature [whether] present in test data point [True]
385 Text feature [line] present in test data point [True]
386 Text feature [studied] present in test data point [True]
387 Text feature [signals] present in test data point [True]
388 Text feature [one] present in test data point [True]
389 Text feature [interacts] present in test data point [True]
390 Text feature [interact] present in test data point [True]
393 Text feature [potency] present in test data point [True]
395 Text feature [introduction] present in test data point [True]
396 Text feature [23] present in test data point [True]
397 Text feature [inhibitory] present in test data point [True]
399 Text feature [history] present in test data point [True]
401 Text feature [rearrangement] present in test data point [True]
403 Text feature [discussion] present in test data point [True]
406 Text feature [mg] present in test data point [True]
408 Text feature [testing] present in test data point [True]
410 Text feature [basis] present in test data point [True]
411 Text feature [culture] present in test data point [True]
412 Text feature [significant] present in test data point [True]
413 Text feature [molecule] present in test data point [True]
415 Text feature [40] present in test data point [True]
416 Text feature [predict] present in test data point [True]
418 Text feature [24] present in test data point [True]
421 Text feature [gefitinib] present in test data point [True]
422 Text feature [showed] present in test data point [True]
423 Text feature [determine] present in test data point [True]
425 Text feature [mutations] present in test data point [True]
427 Text feature [figure] present in test data point [True]
428 Text feature [revealed] present in test data point [True]
429 Text feature [01] present in test data point [True]
```

```
430 Text feature [mek1] present in test data point [True]
432 Text feature [ml] present in test data point [True]
433 Text feature [targeted] present in test data point [True]
435 Text feature [truncated] present in test data point [True]
437 Text feature [antibody] present in test data point [True]
438 Text feature [promote] present in test data point [True]
440 Text feature [first] present in test data point [True]
441 Text feature [analysis] present in test data point [True]
442 Text feature [leading] present in test data point [True]
443 Text feature [domains] present in test data point [True]
444 Text feature [repeat] present in test data point [True]
447 Text feature [identified] present in test data point [True]
449 Text feature [multiple] present in test data point [True]
450 Text feature [substitutions] present in test data point [True]
451 Text feature [mutagenesis] present in test data point [True]
454 Text feature [available] present in test data point [True]
455 Text feature [role] present in test data point [True]
456 Text feature [plates] present in test data point [True]
457 Text feature [highly] present in test data point [True]
459 Text feature [relative] present in test data point [True]
460 Text feature [focused] present in test data point [True]
461 Text feature [braf] present in test data point [True]
463 Text feature [small] present in test data point [True]
464 Text feature [33] present in test data point [True]
465 Text feature [genetic] present in test data point [True]
466 Text feature [intrinsic] present in test data point [True]
467 Text feature [certain] present in test data point [True]
468 Text feature [tissue] present in test data point [True]
469 Text feature [human] present in test data point [True]
470 Text feature [genes] present in test data point [True]
472 Text feature [tumor] present in test data point [True]
473 Text feature [day] present in test data point [True]
474 Text feature [expressed] present in test data point [True]
476 Text feature [majority] present in test data point [True]
477 Text feature [two] present in test data point [True]
478 Text feature [model] present in test data point [True]
480 Text feature [breast] present in test data point [True]
481 Text feature [common] present in test data point [True]
483 Text feature [15] present in test data point [True]
```

```
484 Text feature [developed] present in test data point [True]
486 Text feature [group] present in test data point [True]
488 Text feature [studies] present in test data point [True]
489 Text feature [drugs] present in test data point [True]
490 Text feature [occurrence] present in test data point [True]
491 Text feature [whereas] present in test data point [True]
494 Text feature [13] present in test data point [True]
496 Text feature [intermediate] present in test data point [True]
497 Text feature [amino] present in test data point [True]
498 Text feature [kras] present in test data point [True]
499 Text feature [3b] present in test data point [True]
500 Text feature [sites] present in test data point [True]
501 Text feature [lobe] present in test data point [True]
502 Text feature [noted] present in test data point [True]
503 Text feature [old] present in test data point [True]
504 Text feature [structure] present in test data point [True]
505 Text feature [transfection] present in test data point [True]
508 Text feature [17] present in test data point [True]
509 Text feature [samples] present in test data point [True]
511 Text feature [47] present in test data point [True]
512 Text feature [paraffin] present in test data point [True]
513 Text feature [interestingly] present in test data point [True]
514 Text feature [specific] present in test data point [True]
515 Text feature [given] present in test data point [True]
516 Text feature [received] present in test data point [True]
517 Text feature [high] present in test data point [True]
518 Text feature [effects] present in test data point [True]
519 Text feature [four] present in test data point [True]
520 Text feature [rate] present in test data point [True]
521 Text feature [suggest] present in test data point [True]
522 Text feature [involving] present in test data point [True]
525 Text feature [using] present in test data point [True]
526 Text feature [complete] present in test data point [True]
527 Text feature [site] present in test data point [True]
528 Text feature [16] present in test data point [True]
530 Text feature [number] present in test data point [True]
531 Text feature [possible] present in test data point [True]
534 Text feature [together] present in test data point [True]
535 Text feature [serine] present in test data point [True]
```

```
537 Text feature [exon] present in test data point [True]
539 Text feature [also] present in test data point [True]
540 Text feature [assessment] present in test data point [True]
543 Text feature [according] present in test data point [True]
544 Text feature [agar] present in test data point [True]
545 Text feature [set] present in test data point [True]
547 Text feature [respectively] present in test data point [True]
548 Text feature [important] present in test data point [True]
550 Text feature [development] present in test data point [True]
551 Text feature [37] present in test data point [True]
552 Text feature [transcription] present in test data point [True]
554 Text feature [could] present in test data point [True]
555 Text feature [allele] present in test data point [True]
556 Text feature [cancers] present in test data point [True]
559 Text feature [28] present in test data point [True]
560 Text feature [contains] present in test data point [True]
561 Text feature [defined] present in test data point [True]
562 Text feature [targeting] present in test data point [True]
563 Text feature [individuals] present in test data point [True]
565 Text feature [purified] present in test data point [True]
567 Text feature [additional] present in test data point [True]
568 Text feature [including] present in test data point [True]
569 Text feature [liver] present in test data point [True]
570 Text feature [initial] present in test data point [True]
571 Text feature [represent] present in test data point [True]
572 Text feature [confirmed] present in test data point [True]
575 Text feature [examined] present in test data point [True]
576 Text feature [observed] present in test data point [True]
577 Text feature [institutional] present in test data point [True]
578 Text feature [19] present in test data point [True]
579 Text feature [transmembrane] present in test data point [True]
580 Text feature [identify] present in test data point [True]
581 Text feature [described] present in test data point [True]
582 Text feature [new] present in test data point [True]
584 Text feature [regions] present in test data point [True]
585 Text feature [total] present in test data point [True]
586 Text feature [status] present in test data point [True]
587 Text feature [alterations] present in test data point [True]
588 Text feature [phosphatidylinositol] present in test data point [Tru
e]
589 Text feature [ascertained] present in test data point [True]
592 Text feature [screen] present in test data point [True]
593 Text feature [full] present in test data point [True]
594 Text feature [fig] present in test data point [True]
596 Text feature [regulation] present in test data point [True]
598 Text feature [scores] present in test data point [True]
599 Text feature [remaining] present in test data point [True]
600 Text feature [3a] present in test data point [True]
602 Text feature [containing] present in test data point [True]
603 Text feature [highlights] present in test data point [True]
604 Text feature [mm] present in test data point [True]
606 Text feature [negative] present in test data point [True]
607 Text feature [somatic] present in test data point [True]
609 Text feature [context] present in test data point [True]
610 Text feature [none] present in test data point [True]
612 Text feature [pocket] present in test data point [True]
613 Text feature [associated] present in test data point [True]
614 Text feature [evaluated] present in test data point [True]
615 Text feature [measure] present in test data point [True]
617 Text feature [driven] present in test data point [True]
619 Text feature [mutated] present in test data point [True]
621 Text feature [variation] present in test data point [True]
622 Text feature [present] present in test data point [True]
623 Text feature [others] present in test data point [True]
625 Text feature [free] present in test data point [True]
626 Text feature [improved] present in test data point [True]
627 Text feature [homozygous] present in test data point [True]
628 Text feature [conformation] present in test data point [True]
629 Text feature [effect] present in test data point [True]
630 Text feature [altered] present in test data point [True]
631 Text feature [findings] present in test data point [True]
633 Text feature [reported] present in test data point [True]
634 Text feature [japan] present in test data point [True]
635 Text feature [differences] present in test data point [True]
636 Text feature [year] present in test data point [True]
638 Text feature [76] present in test data point [True]
639 Text feature [resulting] present in test data point [True]
640 Text feature [performed] present in test data point [True]
```

```
642 Text feature [deletions] present in test data point [True]
645 Text feature [specimens] present in test data point [True]
646 Text feature [characterized] present in test data point [True]
648 Text feature [substitution] present in test data point [True]
649 Text feature [published] present in test data point [True]
650 Text feature [comparison] present in test data point [True]
651 Text feature [endogenous] present in test data point [True]
652 Text feature [genomic] present in test data point [True]
653 Text feature [79] present in test data point [True]
654 Text feature [harbor] present in test data point [True]
655 Text feature [frequency] present in test data point [True]
657 Text feature [60] present in test data point [True]
659 Text feature [confirm] present in test data point [True]
661 Text feature [analyses] present in test data point [True]
662 Text feature [association] present in test data point [True]
664 Text feature [consistent] present in test data point [True]
665 Text feature [confer] present in test data point [True]
666 Text feature [identification] present in test data point [True]
667 Text feature [models] present in test data point [True]
668 Text feature [calculated] present in test data point [True]
669 Text feature [fold] present in test data point [True]
670 Text feature [indicate] present in test data point [True]
671 Text feature [mutation] present in test data point [True]
672 Text feature [consists] present in test data point [True]
673 Text feature [epithelial] present in test data point [True]
674 Text feature [sequenced] present in test data point [True]
675 Text feature [due] present in test data point [True]
676 Text feature [25] present in test data point [True]
677 Text feature [detection] present in test data point [True]
678 Text feature [led] present in test data point [True]
679 Text feature [change] present in test data point [True]
681 Text feature [point] present in test data point [True]
683 Text feature [included] present in test data point [True]
684 Text feature [ca] present in test data point [True]
685 Text feature [second] present in test data point [True]
687 Text feature [generation] present in test data point [True]
688 Text feature [31] present in test data point [True]
689 Text feature [recurrent] present in test data point [True]
690 Text feature [possibly] present in test data point [True]
```

```
691 Text feature [acids] present in test data point [True]
692 Text feature [strong] present in test data point [True]
693 Text feature [related] present in test data point [True]
694 Text feature [generated] present in test data point [True]
696 Text feature [distribution] present in test data point [True]
699 Text feature [case] present in test data point [True]
700 Text feature [plasma] present in test data point [True]
702 Text feature [position] present in test data point [True]
703 Text feature [occur] present in test data point [True]
705 Text feature [induce] present in test data point [True]
706 Text feature [similar] present in test data point [True]
707 Text feature [constructs] present in test data point [True]
708 Text feature [pcr] present in test data point [True]
709 Text feature [another] present in test data point [True]
711 Text feature [required] present in test data point [True]
712 Text feature [genome] present in test data point [True]
715 Text feature [double] present in test data point [True]
719 Text feature [range] present in test data point [True]
721 Text feature [promoter] present in test data point [True]
724 Text feature [probably] present in test data point [True]
725 Text feature [64] present in test data point [True]
726 Text feature [relevant] present in test data point [True]
727 Text feature [complex] present in test data point [True]
729 Text feature [rare] present in test data point [True]
730 Text feature [42] present in test data point [True]
731 Text feature [show] present in test data point [True]
733 Text feature [progressed] present in test data point [True]
734 Text feature [cause] present in test data point [True]
735 Text feature [amplified] present in test data point [True]
736 Text feature [detected] present in test data point [True]
737 Text feature [across] present in test data point [True]
741 Text feature [level] present in test data point [True]
746 Text feature [et] present in test data point [True]
747 Text feature [suggested] present in test data point [True]
748 Text feature [event] present in test data point [True]
749 Text feature [ng] present in test data point [True]
750 Text feature [different] present in test data point [True]
751 Text feature [assessed] present in test data point [True]
752 Text feature [liquid] present in test data point [True]
```

```
753 Text feature [54] present in test data point [True]
755 Text feature [0001] present in test data point [True]
756 Text feature [three] present in test data point [True]
757 Text feature [30] present in test data point [True]
759 Text feature [distinct] present in test data point [True]
761 Text feature [five] present in test data point [True]
762 Text feature [criteria] present in test data point [True]
763 Text feature [terminus] present in test data point [True]
764 Text feature [evidence] present in test data point [True]
766 Text feature [vivo] present in test data point [True]
768 Text feature [lead] present in test data point [True]
769 Text feature [percentage] present in test data point [True]
770 Text feature [150] present in test data point [True]
771 Text feature [predisposition] present in test data point [True]
772 Text feature [mediated] present in test data point [True]
773 Text feature [even] present in test data point [True]
775 Text feature [thus] present in test data point [True]
776 Text feature [single] present in test data point [True]
777 Text feature [alter] present in test data point [True]
778 Text feature [mrna] present in test data point [True]
779 Text feature [27] present in test data point [True]
780 Text feature [measured] present in test data point [True]
781 Text feature [transformed] present in test data point [True]
783 Text feature [98] present in test data point [True]
784 Text feature [corresponding] present in test data point [True]
785 Text feature [screening] present in test data point [True]
786 Text feature [suggesting] present in test data point [True]
789 Text feature [loop] present in test data point [True]
790 Text feature [48] present in test data point [True]
791 Text feature [salt] present in test data point [True]
792 Text feature [therefore] present in test data point [True]
793 Text feature [activates] present in test data point [True]
794 Text feature [comparable] present in test data point [True]
795 Text feature [unable] present in test data point [True]
796 Text feature [directly] present in test data point [True]
798 Text feature [target] present in test data point [True]
799 Text feature [found] present in test data point [True]
800 Text feature [many] present in test data point [True]
802 Text feature [determined] present in test data point [True]
```

```
804 Text feature [agarose] present in test data point [True]
805 Text feature [overall] present in test data point [True]
806 Text feature [approximately] present in test data point [True]
807 Text feature [approach] present in test data point [True]
808 Text feature [indeed] present in test data point [True]
809 Text feature [evaluate] present in test data point [True]
810 Text feature [test] present in test data point [True]
811 Text feature [may] present in test data point [True]
812 Text feature [features] present in test data point [True]
815 Text feature [various] present in test data point [True]
819 Text feature [51] present in test data point [True]
820 Text feature [analyzed] present in test data point [True]
824 Text feature [lower] present in test data point [True]
825 Text feature [presented] present in test data point [True]
826 Text feature [method] present in test data point [True]
827 Text feature [population] present in test data point [True]
829 Text feature [followed] present in test data point [True]
830 Text feature [97] present in test data point [True]
836 Text feature [intracellular] present in test data point [True]
837 Text feature [derived] present in test data point [True]
838 Text feature [least] present in test data point [True]
841 Text feature [side] present in test data point [True]
842 Text feature [conferred] present in test data point [True]
843 Text feature [visualized] present in test data point [True]
844 Text feature [importance] present in test data point [True]
845 Text feature [correlation] present in test data point [True]
846 Text feature [20] present in test data point [True]
847 Text feature [immunoblotting] present in test data point [True]
848 Text feature [silico] present in test data point [True]
849 Text feature [less] present in test data point [True]
850 Text feature [rates] present in test data point [True]
851 Text feature [sequences] present in test data point [True]
853 Text feature [report] present in test data point [True]
854 Text feature [standard] present in test data point [True]
856 Text feature [documented] present in test data point [True]
857 Text feature [nm] present in test data point [True]
858 Text feature [histopathology] present in test data point [True]
859 Text feature [address] present in test data point [True]
860 Text feature [among] present in test data point [True]
```

```
861 Text feature [particularly] present in test data point [True]
864 Text feature [fully] present in test data point [True]
865 Text feature [34] present in test data point [True]
867 Text feature [unlike] present in test data point [True]
868 Text feature [several] present in test data point [True]
869 Text feature [powerful] present in test data point [True]
870 Text feature [72] present in test data point [True]
871 Text feature [functionally] present in test data point [True]
873 Text feature [3c] present in test data point [True]
874 Text feature [review] present in test data point [True]
875 Text feature [significantly] present in test data point [True]
876 Text feature [reports] present in test data point [True]
879 Text feature [overexpression] present in test data point [True]
881 Text feature [malignant] present in test data point [True]
882 Text feature [fish] present in test data point [True]
883 Text feature [low] present in test data point [True]
884 Text feature [sd] present in test data point [True]
885 Text feature [form] present in test data point [True]
887 Text feature [provided] present in test data point [True]
888 Text feature [non] present in test data point [True]
890 Text feature [individual] present in test data point [True]
891 Text feature [disease] present in test data point [True]
892 Text feature [arise] present in test data point [True]
893 Text feature [yielded] present in test data point [True]
897 Text feature [help] present in test data point [True]
898 Text feature [similarly] present in test data point [True]
901 Text feature [inhibit] present in test data point [True]
902 Text feature [conditions] present in test data point [True]
903 Text feature [robust] present in test data point [True]
904 Text feature [ethnic] present in test data point [True]
906 Text feature [plays] present in test data point [True]
908 Text feature [discovery] present in test data point [True]
913 Text feature [manner] present in test data point [True]
914 Text feature [formation] present in test data point [True]
915 Text feature [upon] present in test data point [True]
916 Text feature [100] present in test data point [True]
917 Text feature [indicating] present in test data point [True]
918 Text feature [difference] present in test data point [True]
919 Text feature [38] present in test data point [True]
```

```
920 Text feature [molecules] present in test data point [True]
921 Text feature [increase] present in test data point [True]
923 Text feature [viability] present in test data point [True]
925 Text feature [unknown] present in test data point [True]
927 Text feature [displayed] present in test data point [True]
928 Text feature [059] present in test data point [True]
930 Text feature [recent] present in test data point [True]
931 Text feature [considered] present in test data point [True]
933 Text feature [subsequently] present in test data point [True]
934 Text feature [prepared] present in test data point [True]
936 Text feature [selective] present in test data point [True]
937 Text feature [41] present in test data point [True]
938 Text feature [finally] present in test data point [True]
940 Text feature [cascade] present in test data point [True]
942 Text feature [interactions] present in test data point [True]
943 Text feature [since] present in test data point [True]
947 Text feature [previous] present in test data point [True]
949 Text feature [68] present in test data point [True]
950 Text feature [egf] present in test data point [True]
952 Text feature [4b] present in test data point [True]
955 Text feature [furthermore] present in test data point [True]
956 Text feature [currently] present in test data point [True]
958 Text feature [component] present in test data point [True]
959 Text feature [46] present in test data point [True]
963 Text feature [alleles] present in test data point [True]
964 Text feature [specificity] present in test data point [True]
965 Text feature [example] present in test data point [True]
966 Text feature [critical] present in test data point [True]
967 Text feature [unique] present in test data point [True]
968 Text feature [per] present in test data point [True]
969 Text feature [would] present in test data point [True]
970 Text feature [phase] present in test data point [True]
971 Text feature [roles] present in test data point [True]
972 Text feature [cycle] present in test data point [True]
973 Text feature [direct] present in test data point [True]
974 Text feature [following] present in test data point [True]
976 Text feature [clear] present in test data point [True]
981 Text feature [selection] present in test data point [True]
982 Text feature [67] present in test data point [True]
```

```
983 Text feature [characterization] present in test data point [True]
985 Text feature [lack] present in test data point [True]
986 Text feature [62] present in test data point [True]
988 Text feature [mouse] present in test data point [True]
989 Text feature [limited] present in test data point [True]
990 Text feature [experiment] present in test data point [True]
992 Text feature [56] present in test data point [True]
993 Text feature [express] present in test data point [True]
995 Text feature [result] present in test data point [True]
996 Text feature [characteristics] present in test data point [True]
997 Text feature [pdb] present in test data point [True]
998 Text feature [otherwise] present in test data point [True]
999 Text feature [need] present in test data point [True]
Out of the top  1000  features  684 are present in query point
```

4.5.3.2. Inorrectly Classified point

In [107]:

```
test_point_index = 15
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_index]),4))
print("Actuall Class :", test_y[test_point_index])
indices = np.argsort(-clf.feature_importances_)
print("-"*50)
get_impfeature_names(indices[:no_feature],
                     x_test['TEXT'].iloc[test_point_index],
                     x_test['Gene'].iloc[test_point_index],
                     x_test['Variation'].iloc[test_point_index],
                     no_feature)
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0728 0.1721 0.0216 0.0705 0.0465 0.0
449 0.5603 0.0061 0.0052]]
Actuall Class : 3
--------------------------------------------------
```

```
0 Text feature [kinase] present in test data point [True]
2 Text feature [inhibitors] present in test data point [True]
3 Text feature [oncogenic] present in test data point [True]
4 Text feature [activation] present in test data point [True]
5 Text feature [inhibitor] present in test data point [True]
6 Text feature [tyrosine] present in test data point [True]
7 Text feature [phosphorylation] present in test data point [True]
9 Text feature [missense] present in test data point [True]
10 Text feature [treatment] present in test data point [True]
12 Text feature [signaling] present in test data point [True]
13 Text feature [constitutive] present in test data point [True]
14 Text feature [activated] present in test data point [True]
17 Text feature [growth] present in test data point [True]
19 Text feature [receptor] present in test data point [True]
20 Text feature [therapy] present in test data point [True]
21 Text feature [treated] present in test data point [True]
22 Text feature [loss] present in test data point [True]
23 Text feature [erk] present in test data point [True]
24 Text feature [akt] present in test data point [True]
26 Text feature [functional] present in test data point [True]
27 Text feature [amplification] present in test data point [True]
29 Text feature [transforming] present in test data point [True]
31 Text feature [constitutively] present in test data point [True]
33 Text feature [extracellular] present in test data point [True]
35 Text feature [drug] present in test data point [True]
37 Text feature [factor] present in test data point [True]
38 Text feature [therapeutic] present in test data point [True]
39 Text feature [cells] present in test data point [True]
40 Text feature [expressing] present in test data point [True]
43 Text feature [proliferation] present in test data point [True]
45 Text feature [serum] present in test data point [True]
49 Text feature [phospho] present in test data point [True]
61 Text feature [potential] present in test data point [True]
62 Text feature [ligand] present in test data point [True]
63 Text feature [predicted] present in test data point [True]
64 Text feature [lines] present in test data point [True]
65 Text feature [phosphorylated] present in test data point [True]
67 Text feature [trials] present in test data point [True]
68 Text feature [patients] present in test data point [True]
```

```
72 Text feature [cell] present in test data point [True]
73 Text feature [egfr] present in test data point [True]
74 Text feature [respond] present in test data point [True]
79 Text feature [clinical] present in test data point [True]
83 Text feature [inhibition] present in test data point [True]
84 Text feature [starved] present in test data point [True]
89 Text feature [efficacy] present in test data point [True]
92 Text feature [response] present in test data point [True]
94 Text feature [effective] present in test data point [True]
96 Text feature [sensitive] present in test data point [True]
99 Text feature [patient] present in test data point [True]
101 Text feature [active] present in test data point [True]
103 Text feature [expression] present in test data point [True]
105 Text feature [variant] present in test data point [True]
111 Text feature [sensitivity] present in test data point [True]
115 Text feature [affect] present in test data point [True]
119 Text feature [dose] present in test data point [True]
123 Text feature [presence] present in test data point [True]
124 Text feature [binding] present in test data point [True]
126 Text feature [classified] present in test data point [True]
128 Text feature [tki] present in test data point [True]
131 Text feature [terminal] present in test data point [True]
135 Text feature [weeks] present in test data point [True]
137 Text feature [bind] present in test data point [True]
138 Text feature [assays] present in test data point [True]
142 Text feature [retained] present in test data point [True]
143 Text feature [ability] present in test data point [True]
146 Text feature [sequencing] present in test data point [True]
150 Text feature [daily] present in test data point [True]
156 Text feature [length] present in test data point [True]
157 Text feature [vehicle] present in test data point [True]
161 Text feature [days] present in test data point [True]
162 Text feature [assay] present in test data point [True]
164 Text feature [tumors] present in test data point [True]
169 Text feature [conserved] present in test data point [True]
170 Text feature [activity] present in test data point [True]
171 Text feature [expected] present in test data point [True]
172 Text feature [independent] present in test data point [True]
176 Text feature [mutants] present in test data point [True]
```

```
179 Text feature [used] present in test data point [True]
181 Text feature [contrast] present in test data point [True]
182 Text feature [lung] present in test data point [True]
185 Text feature [concentrations] present in test data point [True]
186 Text feature [molecular] present in test data point [True]
192 Text feature [anchorage] present in test data point [True]
195 Text feature [26] present in test data point [True]
197 Text feature [receptors] present in test data point [True]
198 Text feature [stimulation] present in test data point [True]
200 Text feature [antibodies] present in test data point [True]
203 Text feature [wild] present in test data point [True]
204 Text feature [likely] present in test data point [True]
205 Text feature [signal] present in test data point [True]
207 Text feature [deletion] present in test data point [True]
211 Text feature [based] present in test data point [True]
219 Text feature [controls] present in test data point [True]
223 Text feature [14] present in test data point [True]
225 Text feature [absence] present in test data point [True]
226 Text feature [pathways] present in test data point [True]
233 Text feature [primary] present in test data point [True]
235 Text feature [epidermal] present in test data point [True]
238 Text feature [mechanism] present in test data point [True]
244 Text feature [reduced] present in test data point [True]
245 Text feature [surface] present in test data point [True]
247 Text feature [35] present in test data point [True]
248 Text feature [21] present in test data point [True]
250 Text feature [type] present in test data point [True]
251 Text feature [enhanced] present in test data point [True]
252 Text feature [mice] present in test data point [True]
254 Text feature [domain] present in test data point [True]
256 Text feature [mutant] present in test data point [True]
258 Text feature [1a] present in test data point [True]
260 Text feature [12] present in test data point [True]
262 Text feature [align] present in test data point [True]
273 Text feature [transduction] present in test data point [True]
276 Text feature [experiments] present in test data point [True]
277 Text feature [large] present in test data point [True]
278 Text feature [32] present in test data point [True]
282 Text feature [previously] present in test data point [True]
```

```
285 Text feature [stimulated] present in test data point [True]
289 Text feature [anti] present in test data point [True]
292 Text feature [changes] present in test data point [True]
294 Text feature [increased] present in test data point [True]
296 Text feature [2b] present in test data point [True]
299 Text feature [therapies] present in test data point [True]
300 Text feature [soft] present in test data point [True]
302 Text feature [duration] present in test data point [True]
303 Text feature [acid] present in test data point [True]
305 Text feature [although] present in test data point [True]
309 Text feature [induced] present in test data point [True]
314 Text feature [compared] present in test data point [True]
316 Text feature [significance] present in test data point [True]
317 Text feature [22] present in test data point [True]
318 Text feature [western] present in test data point [True]
319 Text feature [next] present in test data point [True]
320 Text feature [demonstrated] present in test data point [True]
321 Text feature [well] present in test data point [True]
322 Text feature [shown] present in test data point [True]
323 Text feature [structural] present in test data point [True]
327 Text feature [however] present in test data point [True]
332 Text feature [obtained] present in test data point [True]
335 Text feature [vitro] present in test data point [True]
337 Text feature [indicated] present in test data point [True]
340 Text feature [11] present in test data point [True]
342 Text feature [data] present in test data point [True]
343 Text feature [tested] present in test data point [True]
344 Text feature [2a] present in test data point [True]
352 Text feature [results] present in test data point [True]
354 Text feature [higher] present in test data point [True]
361 Text feature [addition] present in test data point [True]
362 Text feature [10] present in test data point [True]
367 Text feature [involved] present in test data point [True]
369 Text feature [time] present in test data point [True]
373 Text feature [residues] present in test data point [True]
375 Text feature [control] present in test data point [True]
377 Text feature [18] present in test data point [True]
379 Text feature [within] present in test data point [True]
384 Text feature [whether] present in test data point [True]
```

```
385 Text feature [line] present in test data point [True]
386 Text feature [studied] present in test data point [True]
387 Text feature [signals] present in test data point [True]
388 Text feature [one] present in test data point [True]
395 Text feature [introduction] present in test data point [True]
396 Text feature [23] present in test data point [True]
403 Text feature [discussion] present in test data point [True]
406 Text feature [mg] present in test data point [True]
407 Text feature [inactivated] present in test data point [True]
408 Text feature [testing] present in test data point [True]
410 Text feature [basis] present in test data point [True]
412 Text feature [significant] present in test data point [True]
413 Text feature [molecule] present in test data point [True]
415 Text feature [40] present in test data point [True]
418 Text feature [24] present in test data point [True]
421 Text feature [gefitinib] present in test data point [True]
422 Text feature [showed] present in test data point [True]
423 Text feature [determine] present in test data point [True]
425 Text feature [mutations] present in test data point [True]
427 Text feature [figure] present in test data point [True]
428 Text feature [revealed] present in test data point [True]
432 Text feature [ml] present in test data point [True]
433 Text feature [targeted] present in test data point [True]
437 Text feature [antibody] present in test data point [True]
440 Text feature [first] present in test data point [True]
441 Text feature [analysis] present in test data point [True]
442 Text feature [leading] present in test data point [True]
443 Text feature [domains] present in test data point [True]
447 Text feature [identified] present in test data point [True]
454 Text feature [available] present in test data point [True]
457 Text feature [highly] present in test data point [True]
459 Text feature [relative] present in test data point [True]
463 Text feature [small] present in test data point [True]
464 Text feature [33] present in test data point [True]
469 Text feature [human] present in test data point [True]
472 Text feature [tumor] present in test data point [True]
473 Text feature [day] present in test data point [True]
474 Text feature [expressed] present in test data point [True]
476 Text feature [majority] present in test data point [True]
```

```
477 Text feature [two] present in test data point [True]
481 Text feature [common] present in test data point [True]
483 Text feature [15] present in test data point [True]
486 Text feature [group] present in test data point [True]
488 Text feature [studies] present in test data point [True]
492 Text feature [lysates] present in test data point [True]
494 Text feature [13] present in test data point [True]
495 Text feature [cultured] present in test data point [True]
497 Text feature [amino] present in test data point [True]
499 Text feature [3b] present in test data point [True]
504 Text feature [structure] present in test data point [True]
508 Text feature [17] present in test data point [True]
509 Text feature [samples] present in test data point [True]
514 Text feature [specific] present in test data point [True]
515 Text feature [given] present in test data point [True]
517 Text feature [high] present in test data point [True]
519 Text feature [four] present in test data point [True]
521 Text feature [suggest] present in test data point [True]
522 Text feature [involving] present in test data point [True]
525 Text feature [using] present in test data point [True]
526 Text feature [complete] present in test data point [True]
527 Text feature [site] present in test data point [True]
528 Text feature [16] present in test data point [True]
530 Text feature [number] present in test data point [True]
531 Text feature [possible] present in test data point [True]
534 Text feature [together] present in test data point [True]
535 Text feature [serine] present in test data point [True]
539 Text feature [also] present in test data point [True]
541 Text feature [basal] present in test data point [True]
544 Text feature [agar] present in test data point [True]
548 Text feature [important] present in test data point [True]
554 Text feature [could] present in test data point [True]
556 Text feature [cancers] present in test data point [True]
559 Text feature [28] present in test data point [True]
560 Text feature [contains] present in test data point [True]
562 Text feature [targeting] present in test data point [True]
568 Text feature [including] present in test data point [True]
570 Text feature [initial] present in test data point [True]
571 Text feature [represent] present in test data point [True]
```

```
572 Text feature [confirmed] present in test data point [True]
576 Text feature [observed] present in test data point [True]
581 Text feature [described] present in test data point [True]
585 Text feature [total] present in test data point [True]
586 Text feature [status] present in test data point [True]
591 Text feature [added] present in test data point [True]
593 Text feature [full] present in test data point [True]
599 Text feature [remaining] present in test data point [True]
600 Text feature [3a] present in test data point [True]
602 Text feature [containing] present in test data point [True]
613 Text feature [associated] present in test data point [True]
619 Text feature [mutated] present in test data point [True]
622 Text feature [present] present in test data point [True]
625 Text feature [free] present in test data point [True]
628 Text feature [conformation] present in test data point [True]
629 Text feature [effect] present in test data point [True]
631 Text feature [findings] present in test data point [True]
633 Text feature [reported] present in test data point [True]
635 Text feature [differences] present in test data point [True]
639 Text feature [resulting] present in test data point [True]
640 Text feature [performed] present in test data point [True]
642 Text feature [deletions] present in test data point [True]
646 Text feature [characterized] present in test data point [True]
648 Text feature [substitution] present in test data point [True]
650 Text feature [comparison] present in test data point [True]
651 Text feature [endogenous] present in test data point [True]
657 Text feature [60] present in test data point [True]
659 Text feature [confirm] present in test data point [True]
662 Text feature [association] present in test data point [True]
664 Text feature [consistent] present in test data point [True]
669 Text feature [fold] present in test data point [True]
670 Text feature [indicate] present in test data point [True]
671 Text feature [mutation] present in test data point [True]
676 Text feature [25] present in test data point [True]
677 Text feature [detection] present in test data point [True]
679 Text feature [change] present in test data point [True]
681 Text feature [point] present in test data point [True]
683 Text feature [included] present in test data point [True]
685 Text feature [second] present in test data point [True]
```

```
687 Text feature [generation] present in test data point [True]
688 Text feature [31] present in test data point [True]
690 Text feature [possibly] present in test data point [True]
691 Text feature [acids] present in test data point [True]
694 Text feature [generated] present in test data point [True]
695 Text feature [marked] present in test data point [True]
699 Text feature [case] present in test data point [True]
702 Text feature [position] present in test data point [True]
703 Text feature [occur] present in test data point [True]
705 Text feature [induce] present in test data point [True]
706 Text feature [similar] present in test data point [True]
709 Text feature [another] present in test data point [True]
711 Text feature [required] present in test data point [True]
719 Text feature [range] present in test data point [True]
720 Text feature [nude] present in test data point [True]
723 Text feature [alignment] present in test data point [True]
731 Text feature [show] present in test data point [True]
732 Text feature [viral] present in test data point [True]
736 Text feature [detected] present in test data point [True]
739 Text feature [elevated] present in test data point [True]
741 Text feature [level] present in test data point [True]
743 Text feature [amount] present in test data point [True]
750 Text feature [different] present in test data point [True]
751 Text feature [assessed] present in test data point [True]
754 Text feature [cellular] present in test data point [True]
755 Text feature [0001] present in test data point [True]
756 Text feature [three] present in test data point [True]
757 Text feature [30] present in test data point [True]
759 Text feature [distinct] present in test data point [True]
760 Text feature [species] present in test data point [True]
766 Text feature [vivo] present in test data point [True]
767 Text feature [phosphotyrosine] present in test data point [True]
768 Text feature [lead] present in test data point [True]
769 Text feature [percentage] present in test data point [True]
776 Text feature [single] present in test data point [True]
779 Text feature [27] present in test data point [True]
784 Text feature [corresponding] present in test data point [True]
786 Text feature [suggesting] present in test data point [True]
787 Text feature [indistinguishable] present in test data point [True]
```

```
789 Text feature [loop] present in test data point [True]
795 Text feature [unable] present in test data point [True]
797 Text feature [central] present in test data point [True]
799 Text feature [found] present in test data point [True]
800 Text feature [many] present in test data point [True]
801 Text feature [lysis] present in test data point [True]
802 Text feature [determined] present in test data point [True]
806 Text feature [approximately] present in test data point [True]
810 Text feature [test] present in test data point [True]
811 Text feature [may] present in test data point [True]
820 Text feature [analyzed] present in test data point [True]
823 Text feature [xenograft] present in test data point [True]
824 Text feature [lower] present in test data point [True]
825 Text feature [presented] present in test data point [True]
827 Text feature [population] present in test data point [True]
829 Text feature [followed] present in test data point [True]
834 Text feature [nucleus] present in test data point [True]
836 Text feature [intracellular] present in test data point [True]
837 Text feature [derived] present in test data point [True]
838 Text feature [least] present in test data point [True]
841 Text feature [side] present in test data point [True]
846 Text feature [20] present in test data point [True]
848 Text feature [silico] present in test data point [True]
849 Text feature [less] present in test data point [True]
851 Text feature [sequences] present in test data point [True]
853 Text feature [report] present in test data point [True]
854 Text feature [standard] present in test data point [True]
856 Text feature [documented] present in test data point [True]
857 Text feature [nm] present in test data point [True]
864 Text feature [fully] present in test data point [True]
865 Text feature [34] present in test data point [True]
866 Text feature [partially] present in test data point [True]
867 Text feature [unlike] present in test data point [True]
868 Text feature [several] present in test data point [True]
873 Text feature [3c] present in test data point [True]
875 Text feature [significantly] present in test data point [True]
876 Text feature [reports] present in test data point [True]
879 Text feature [overexpression] present in test data point [True]
881 Text feature [malignant] present in test data point [True]
```

```
883 Text feature [low] present in test data point [True]
885 Text feature [form] present in test data point [True]
888 Text feature [non] present in test data point [True]
889 Text feature [seeded] present in test data point [True]
892 Text feature [arise] present in test data point [True]
897 Text feature [help] present in test data point [True]
902 Text feature [conditions] present in test data point [True]
903 Text feature [robust] present in test data point [True]
905 Text feature [blotting] present in test data point [True]
913 Text feature [manner] present in test data point [True]
914 Text feature [formation] present in test data point [True]
915 Text feature [upon] present in test data point [True]
917 Text feature [indicating] present in test data point [True]
918 Text feature [difference] present in test data point [True]
921 Text feature [increase] present in test data point [True]
925 Text feature [unknown] present in test data point [True]
927 Text feature [displayed] present in test data point [True]
930 Text feature [recent] present in test data point [True]
933 Text feature [subsequently] present in test data point [True]
945 Text feature [showing] present in test data point [True]
946 Text feature [phenotype] present in test data point [True]
947 Text feature [previous] present in test data point [True]
949 Text feature [68] present in test data point [True]
950 Text feature [egf] present in test data point [True]
952 Text feature [4b] present in test data point [True]
956 Text feature [currently] present in test data point [True]
965 Text feature [example] present in test data point [True]
967 Text feature [unique] present in test data point [True]
968 Text feature [per] present in test data point [True]
969 Text feature [would] present in test data point [True]
970 Text feature [phase] present in test data point [True]
974 Text feature [following] present in test data point [True]
976 Text feature [clear] present in test data point [True]
977 Text feature [substituted] present in test data point [True]
981 Text feature [selection] present in test data point [True]
983 Text feature [characterization] present in test data point [True]
985 Text feature [lack] present in test data point [True]
986 Text feature [62] present in test data point [True]
990 Text feature [experiment] present in test data point [True]
```

```
994 Text feature [consequences] present in test data point [True]
995 Text feature [result] present in test data point [True]
996 Text feature [characteristics] present in test data point [True]
Out of the top  1000  features  393 are present in query point
```

4.5.3. Hyper paramter tuning (With Response Coding)

In [108]:

```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the give
n training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# --------------------------------


# find more about CalibratedClassifierCV here at http://scikit-learn.or
g/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.h
tml
# ----------------------------
# default paramters
# sklearn.calibration.CalibratedClassifierCV(base_estimator=None, metho
d='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 = [10,50,100,200,500,1000]
max_depth = [2,3,5,10]
cv_log_error_array = []
for i in alpha:
    for j in max_depth:
        print("for n_estimators =", i,"and max depth = ", j)
        clf = RandomForestClassifier(n_estimators=i, criterion='gini',
max_depth=j, random_state=42, n_jobs=-1)
        clf.fit(train_x_responseCoding, train_y)
        sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
        sig_clf.fit(train_x_responseCoding, train_y)
        sig_clf_probs = sig_clf.predict_proba(cv_x_responseCoding)
        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()
features = np.dot(np.array(alpha)[:,None],np.array(max_depth)[None]).ra
vel()
ax.plot(features, cv_log_error_array,c='g')
for i, txt in enumerate(np.round(cv_log_error_array,3)):
    ax.annotate((alpha[int(i/4)],max_depth[int(i%4)],str(txt)), (featur
es[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 = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], cri
terion='gini', max_depth=max_depth[int(best_alpha%4)], random_state=42,
 n_jobs=-1)
clf.fit(train_x_responseCoding, train_y)
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)

predict_y = sig_clf.predict_proba(train_x_responseCoding)
print('For values of best alpha = ',
      alpha[int(best_alpha/4)],
      "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_responseCoding)
print('For values of best alpha = ',
      alpha[int(best_alpha/4)],
      "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_responseCoding)
print('For values of best alpha = ',
      alpha[int(best_alpha/4)],
      "The test log loss is:",
      log_loss(y_test, predict_y, labels=clf.classes_, eps=1e-15))
```

```
for n_estimators = 10 and max depth =  2
Log Loss : 2.0182008146197306
for n_estimators = 10 and max depth =  3
Log Loss : 1.8238274802183039
for n_estimators = 10 and max depth =  5
Log Loss : 1.5809792748387719
for n_estimators = 10 and max depth =  10
Log Loss : 1.9456492965962777
for n_estimators = 50 and max depth =  2
Log Loss : 1.692204438079307
for n_estimators = 50 and max depth =  3
```

```
Log Loss : 1.4358881584022634
for n_estimators = 50 and max depth =  5
Log Loss : 1.4334815417208533
for n_estimators = 50 and max depth =  10
Log Loss : 1.689380400500247
for n_estimators = 100 and max depth =  2
Log Loss : 1.566240825260833
for n_estimators = 100 and max depth =  3
Log Loss : 1.5004575437995789
for n_estimators = 100 and max depth =  5
Log Loss : 1.337274195336184
for n_estimators = 100 and max depth =  10
Log Loss : 1.7258293604731398
for n_estimators = 200 and max depth =  2
Log Loss : 1.6491938227930416
for n_estimators = 200 and max depth =  3
Log Loss : 1.5414315201058009
for n_estimators = 200 and max depth =  5
Log Loss : 1.4129134391116438
for n_estimators = 200 and max depth =  10
Log Loss : 1.6860613455971567
for n_estimators = 500 and max depth =  2
Log Loss : 1.7335935756625864
for n_estimators = 500 and max depth =  3
Log Loss : 1.5798057951549236
for n_estimators = 500 and max depth =  5
Log Loss : 1.4076949532959606
for n_estimators = 500 and max depth =  10
Log Loss : 1.709284566298512
for n_estimators = 1000 and max depth =  2
Log Loss : 1.6982219916228383
for n_estimators = 1000 and max depth =  3
Log Loss : 1.5830270033873513
for n_estimators = 1000 and max depth =  5
Log Loss : 1.393089124111703
for n_estimators = 1000 and max depth =  10
Log Loss : 1.6680620742042442
For values of best alpha =  100 The train log loss is: 0.05538723850320
481
For values of best alpha =  100 The cross validation log loss is: 1.337
```

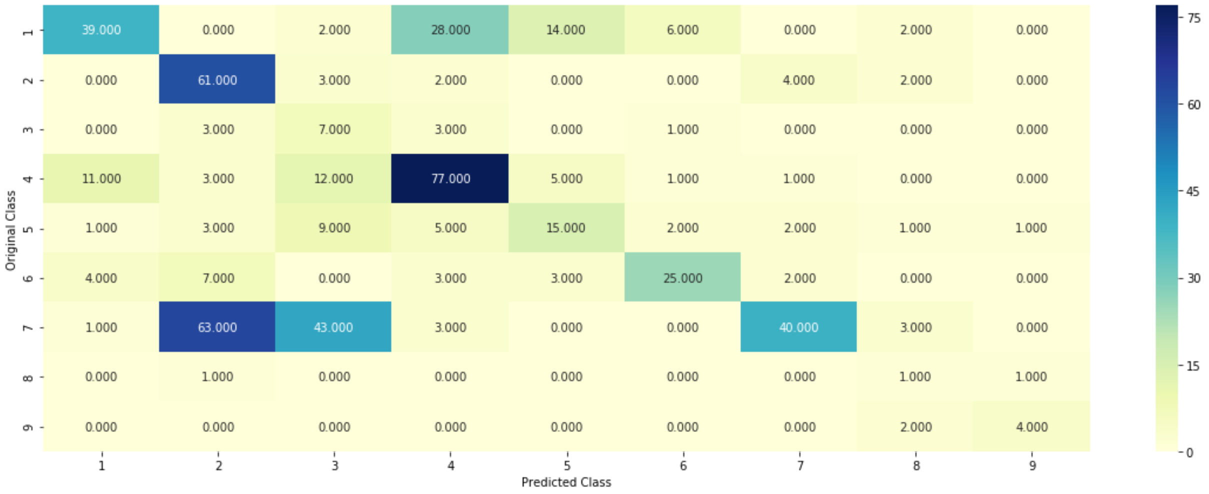
```
2741953361835
For values of best alpha =  100 The test log loss is: 1.342718125993264
8
```

4.5.4. Testing model with best hyper parameters (Response Coding)

In [109]:

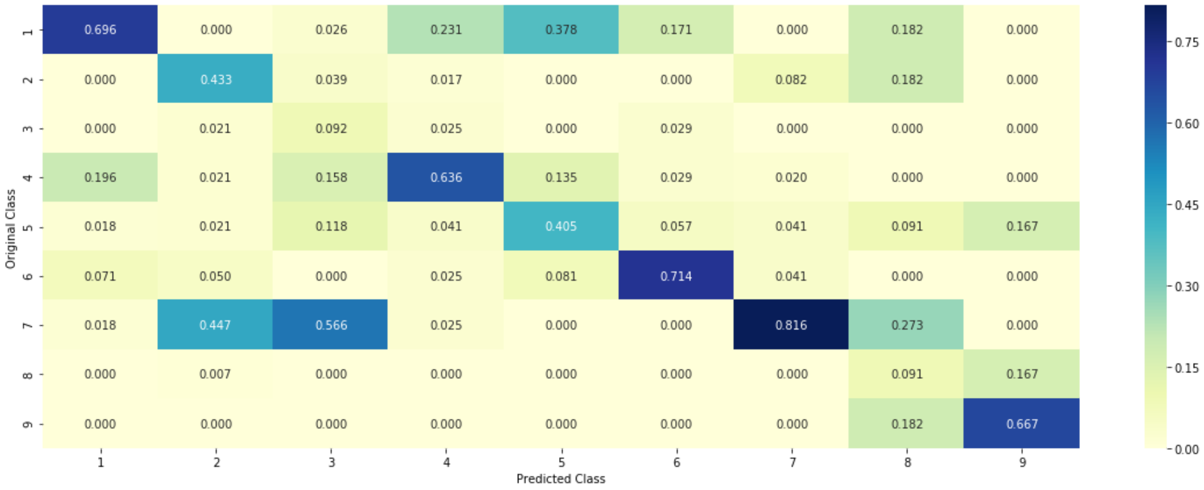
```
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='g
ini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='aut
o', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, r
andom_state=None, verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the give
n training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)     Perform classification on samples in X.

# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-
online/lessons/random-forest-and-their-construction-2/
# --------------------------------

clf = RandomForestClassifier(max_depth=max_depth[int(best_alpha%4)], n_
estimators=alpha[int(best_alpha/4)], criterion='gini', max_features='au
to',random_state=42)
predict_and_plot_confusion_matrix(train_x_responseCoding, train_y,cv_x_
responseCoding,cv_y, clf)
```

```
Log loss : 1.3372741953361835
Number of mis-classified points : 0.4943609022556391
Number of mis-classified points : 0.4943609022556391

-------------------- Confusion matrix --------------------
```



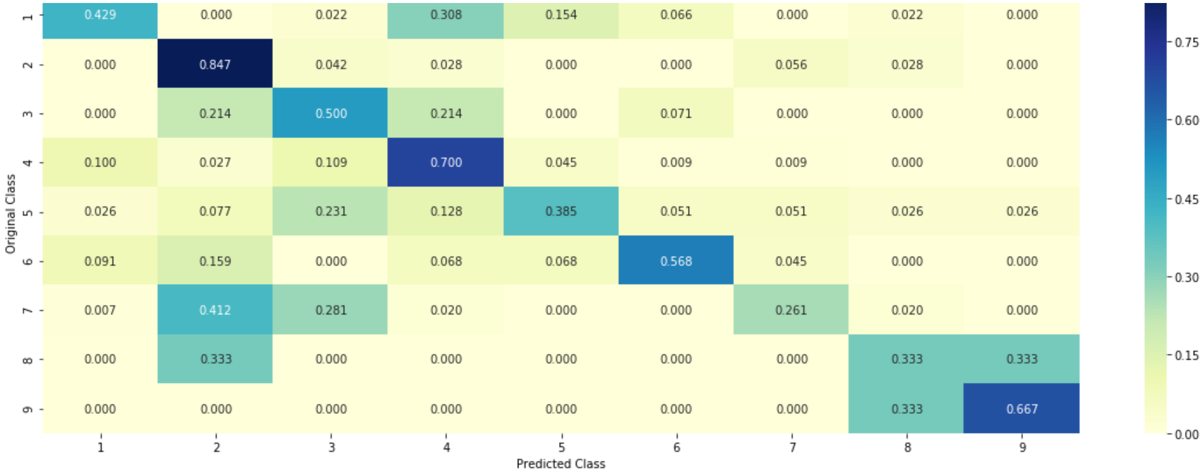
	1	2	3	4	5	6	7	8	9
1	39.000	0.000	2.000	28.000	14.000	6.000	0.000	2.000	0.000
2	0.000	61.000	3.000	2.000	0.000	0.000	4.000	2.000	0.000
3	0.000	3.000	7.000	3.000	0.000	1.000	0.000	0.000	0.000
4	11.000	3.000	12.000	77.000	5.000	1.000	1.000	0.000	0.000
5	1.000	3.000	9.000	5.000	15.000	2.000	2.000	1.000	1.000
6	4.000	7.000	0.000	3.000	3.000	25.000	2.000	0.000	0.000
7	1.000	63.000	43.000	3.000	0.000	0.000	40.000	3.000	0.000
8	0.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2.000	4.000

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



	1	2	3	4	5	6	7	8	9
1	0.696	0.000	0.026	0.231	0.378	0.171	0.000	0.182	0.000
2	0.000	0.433	0.039	0.017	0.000	0.000	0.082	0.182	0.000
3	0.000	0.021	0.092	0.025	0.000	0.029	0.000	0.000	0.000
4	0.196	0.021	0.158	0.636	0.135	0.029	0.020	0.000	0.000
5	0.018	0.021	0.118	0.041	0.405	0.057	0.041	0.091	0.167
6	0.071	0.050	0.000	0.025	0.081	0.714	0.041	0.000	0.000
7	0.018	0.447	0.566	0.025	0.000	0.000	0.816	0.273	0.000
8	0.000	0.007	0.000	0.000	0.000	0.000	0.000	0.091	0.167
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.182	0.667

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



4.5.5. Feature Importance

4.5.5.1. Correctly Classified point

In [110]:

```
clf = RandomForestClassifier(n_estimators=alpha[int(best_alpha/4)], cri
terion='gini', max_depth=max_depth[int(best_alpha%4)], random_state=42,
 n_jobs=-1)
clf.fit(train_x_responseCoding, train_y)
```

```
sig_clf = CalibratedClassifierCV(clf, method="sigmoid")
sig_clf.fit(train_x_responseCoding, train_y)


test_point_index = 100
no_feature = 1000
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index]
.reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(
test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", test_y[test_point_index])
# indices = np.argsort(-clf.feature_importances_)
# print("-"*50)
# for i in indices:
#     if i<9:
#         print("Gene is important feature")
#     elif i<18:
#         print("Variation is important feature")
#     else:
#         print("Text is important feature")
```

```
Predicted Class : 4
Predicted Class Probabilities: [[0.1285 0.03   0.1507 0.5808 0.0175 0.0
346 0.0061 0.0331 0.0187]]
Actual Class : 4
```

4.5.5.2. Incorrectly Classified point

In [111]:

```
test_point_index = 31
predicted_cls = sig_clf.predict(test_x_responseCoding[test_point_index]
.reshape(1,-1))
print("Predicted Class :", predicted_cls[0])
print("Predicted Class Probabilities:", np.round(sig_clf.predict_proba(
test_x_responseCoding[test_point_index].reshape(1,-1)),4))
print("Actual Class :", test_y[test_point_index])
# indices = np.argsort(-clf.feature_importances_)
# print("-"*50)
# for i in indices:
#     if i<9:
#         print("Gene is important feature")
#     elif i<18:
#         print("Variation is important feature")
#     else:
#         print("Text is important feature")
```

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0275 0.1311 0.2907 0.0231 0.023  0.0439 0.3708 0.0327 0.0573]]
Actual Class : 7
```

4.7 Stack the models

4.7.1 testing with hyper parameter tuning

In [112]:

```
# 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/
#------------------------------
```

```
# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
# --------------------------------
# default parameters
# SVC(C=1.0, kernel='rbf', degree=3, gamma='auto', coef0=0.0, shrinking=True, probability=False, tol=0.001,
# cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', random_state=None)

# Some of methods of SVM()
# fit(X, y, [sample_weight])    Fit the SVM model according to the given training data.
# predict(X)    Perform classification on samples in X.
# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/mathematical-derivation-copy-8/
# --------------------------------


# read more about support vector machines with linear kernals here http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
# --------------------------------
# default parameters
# sklearn.ensemble.RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2,
# min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,
# min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=False,
# class_weight=None)

# Some of methods of RandomForestClassifier()
# fit(X, y, [sample_weight])    Fit the SVM model according to the given training data.
# predict(X)    Perform classification on samples in X.
# predict_proba (X)    Perform classification on samples in X.
# some of attributes of  RandomForestClassifier()
# feature_importances_ : array of shape = [n_features]
# The feature importances (the higher, the more important the feature).

# --------------------------------
# video link: https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/random-forest-and-their-construction-2/
# --------------------------------


clf1 = SGDClassifier(alpha=0.001, penalty='l2', loss='log', class_weight='balanced', random_state=0)
clf1.fit(train_x_onehotCoding, train_y)
sig_clf1 = CalibratedClassifierCV(clf1, method="sigmoid")

clf2 = SGDClassifier(alpha=0.01, penalty='l2', loss='hinge', class_weight='balanced', random_state=0)
clf2.fit(train_x_onehotCoding, train_y)
sig_clf2 = CalibratedClassifierCV(clf2, method="sigmoid")


clf3 = MultinomialNB(alpha=1000)
clf3.fit(train_x_onehotCoding, train_y)
sig_clf3 = CalibratedClassifierCV(clf3, method="sigmoid")

sig_clf1.fit(train_x_onehotCoding, train_y)
print("Logistic Regression :  Log Loss: %0.2f" % (log_loss(cv_y, sig_clf1.predict_proba(cv_x_onehotCoding))))
sig_clf2.fit(train_x_onehotCoding, train_y)
print("Support vector machines : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf2.predict_proba(cv_x_onehotCoding))))
sig_clf3.fit(train_x_onehotCoding, train_y)
print("Naive Bayes : Log Loss: %0.2f" % (log_loss(cv_y, sig_clf3.predict_proba(cv_x_onehotCoding))))
print("-"*50)
alpha = [0.0001,0.001,0.01,0.1,1,10]
best_alpha = 999
for i in alpha:
    lr = LogisticRegression(C=i)
    sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3
], meta_classifier=lr, use_probas=True)
    sclf.fit(train_x_onehotCoding, train_y)
    print("Stacking Classifer : for the value of alpha: %f Log Loss: %
0.3f" % (i, log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))))
    log_error =log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))
    if best_alpha > log_error:
        best_alpha = log_error
```

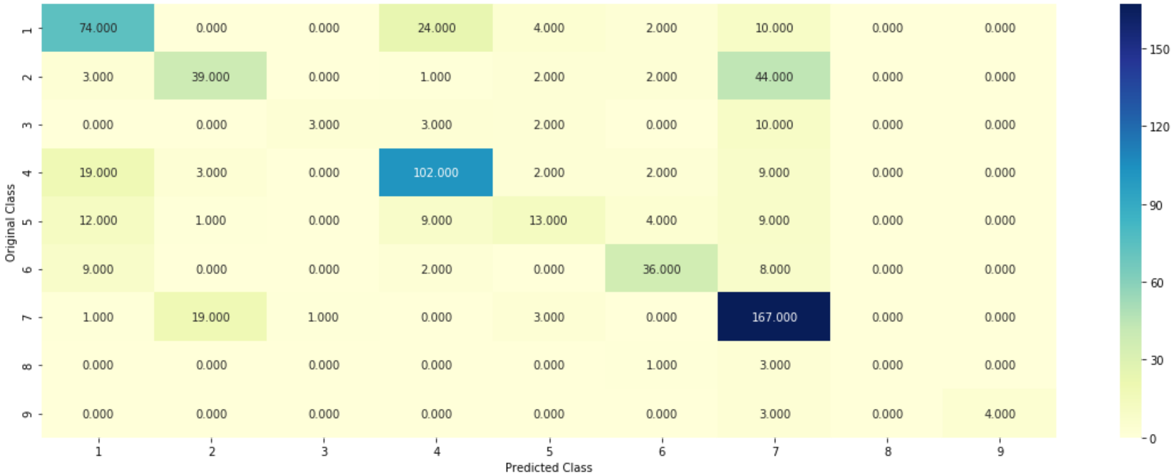
```
Logistic Regression :  Log Loss: 1.08
Support vector machines : Log Loss: 1.13
Naive Bayes : Log Loss: 1.19
--------------------------------------------------
Stacking Classifer : for the value of alpha: 0.000100 Log Loss: 2.173
Stacking Classifer : for the value of alpha: 0.001000 Log Loss: 1.994
Stacking Classifer : for the value of alpha: 0.010000 Log Loss: 1.406
Stacking Classifer : for the value of alpha: 0.100000 Log Loss: 1.096
Stacking Classifer : for the value of alpha: 1.000000 Log Loss: 1.238
Stacking Classifer : for the value of alpha: 10.000000 Log Loss: 1.570
```

4.7.2 testing the model with the best hyper parameters

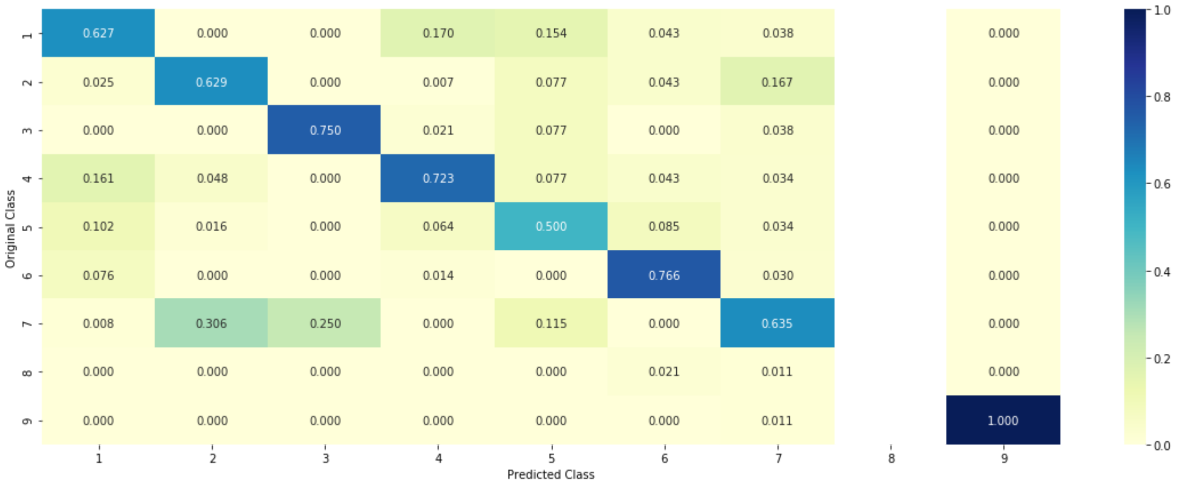
In [113]:

```
lr = LogisticRegression(C=0.1)
sclf = StackingClassifier(classifiers=[sig_clf1, sig_clf2, sig_clf3], m
eta_classifier=lr, use_probas=True)
sclf.fit(train_x_onehotCoding, train_y)

log_error = log_loss(train_y, sclf.predict_proba(train_x_onehotCoding))
print("Log loss (train) on the stacking classifier :",log_error)

log_error = log_loss(cv_y, sclf.predict_proba(cv_x_onehotCoding))
print("Log loss (CV) on the stacking classifier :",log_error)

log_error = log_loss(test_y, sclf.predict_proba(test_x_onehotCoding))
print("Log loss (test) on the stacking classifier :",log_error)

print("Number of missclassified point :", np.count_nonzero((sclf.predic
t(test_x_onehotCoding)- test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=sclf.predict(test_x_oneh
otCoding))
```

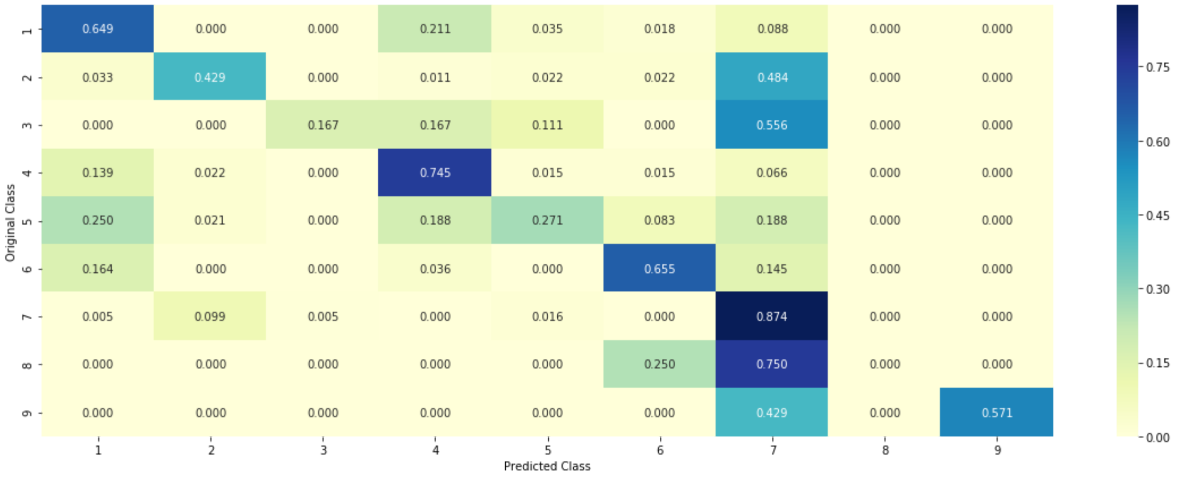
```
Log loss (train) on the stacking classifier : 0.61413274403486 54
Log loss (CV) on the stacking classifier : 1.09583371707431 09
Log loss (test) on the stacking classifier : 1.08784415014571 08
Number of missclassified point : 0.341353383458646 63

-------------------- Confusion matrix --------------------
```



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



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



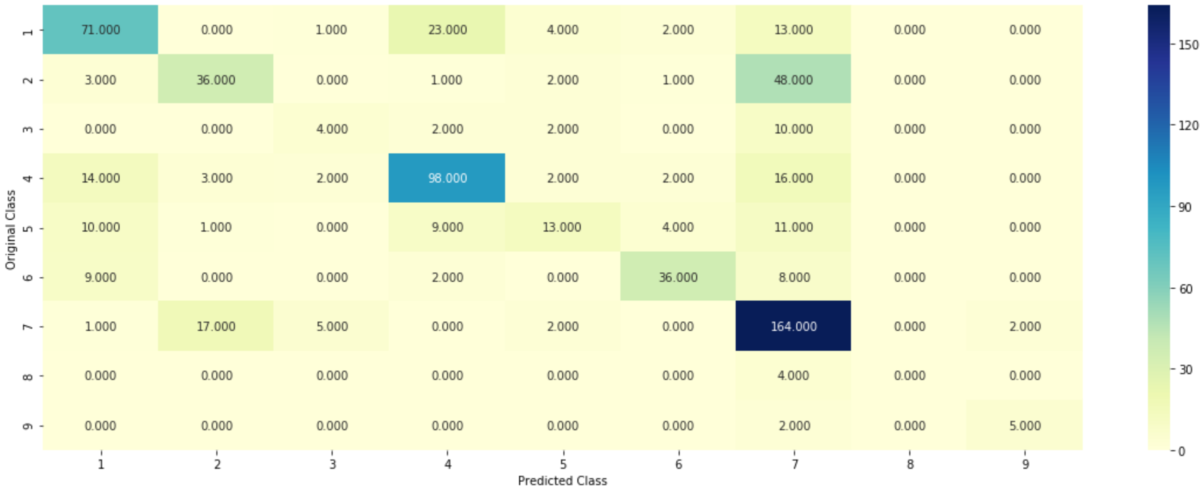
4.7.3 Maximum Voting classifier

In [114]:

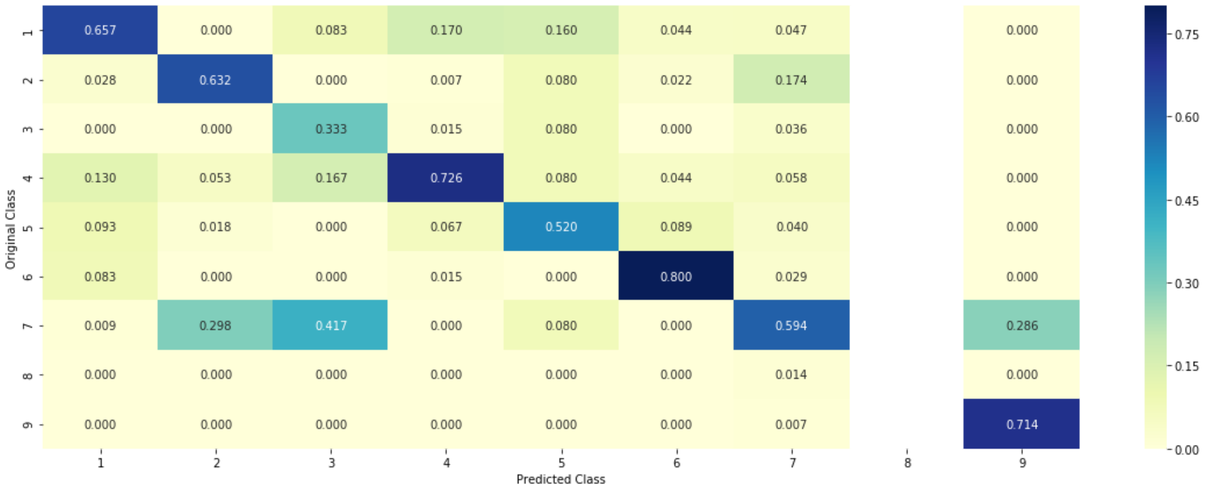
```
#Refer:http://scikit-learn.org/stable/modules/generated/sklearn.ensembl
e.VotingClassifier.html
```

```
from sklearn.ensemble import VotingClassifier
vclf = VotingClassifier(estimators=[('lr', sig_clf1), ('svc', sig_clf2
), ('rf', sig_clf3)], voting='soft')
vclf.fit(train_x_onehotCoding, train_y)
print("Log loss (train) on the VotingClassifier :", log_loss(train_y, v
clf.predict_proba(train_x_onehotCoding)))
print("Log loss (CV) on the VotingClassifier :", log_loss(cv_y, vclf.pr
edict_proba(cv_x_onehotCoding)))
print("Log loss (test) on the VotingClassifier :", log_loss(test_y, vcl
f.predict_proba(test_x_onehotCoding)))
print("Number of missclassified point :", np.count_nonzero((vclf.predic
t(test_x_onehotCoding)- test_y))/test_y.shape[0])
plot_confusion_matrix(test_y=test_y, predict_y=vclf.predict(test_x_oneh
otCoding))
```

```
Log loss (train) on the VotingClassifier : 0.7034960713763345
Log loss (CV) on the VotingClassifier : 1.0511947539385507
Log loss (test) on the VotingClassifier : 1.0285164583991668
Number of missclassified point : 0.35789473684210527

-------------------- Confusion matrix --------------------
```

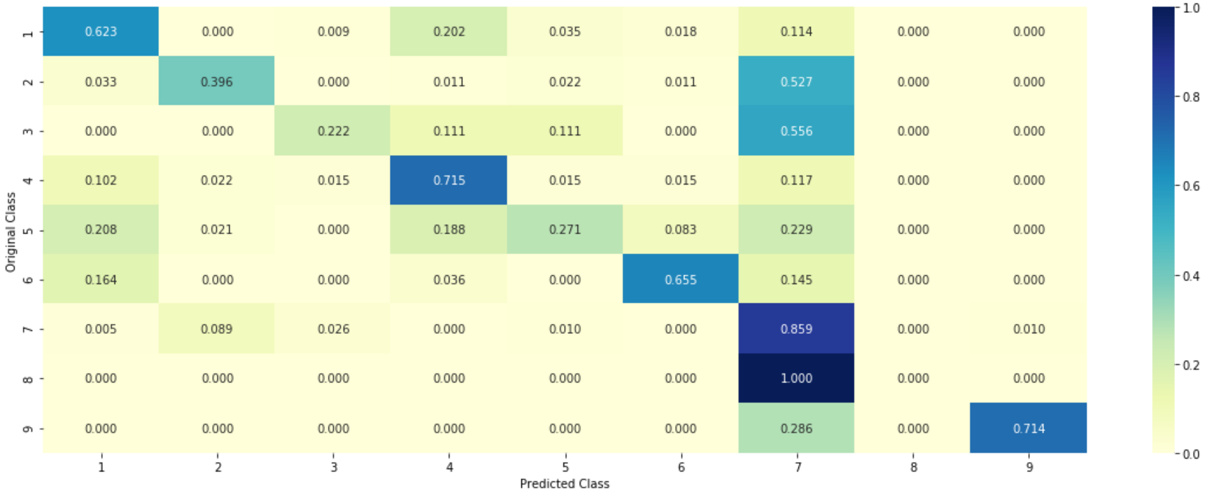


```
-------------------- Precision matrix (Columm Sum=1) -----------------
--
```



	1	2	3	4	5	6	7	8	9
1	0.657	0.000	0.083	0.170	0.160	0.044	0.047		0.000
2	0.028	0.632	0.000	0.007	0.080	0.022	0.174		0.000
3	0.000	0.000	0.333	0.015	0.080	0.000	0.036		0.000
4	0.130	0.053	0.167	0.726	0.080	0.044	0.058		0.000
5	0.093	0.018	0.000	0.067	0.520	0.089	0.040		0.000
6	0.083	0.000	0.000	0.015	0.000	0.800	0.029		0.000
7	0.009	0.298	0.417	0.000	0.080	0.000	0.594		0.286
8	0.000	0.000	0.000	0.000	0.000	0.000	0.014		0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.007		0.714

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



	1	2	3	4	5	6	7	8	9
1	0.623	0.000	0.009	0.202	0.035	0.018	0.114	0.000	0.000
2	0.033	0.396	0.000	0.011	0.022	0.011	0.527	0.000	0.000
3	0.000	0.000	0.222	0.111	0.111	0.000	0.556	0.000	0.000
4	0.102	0.022	0.015	0.715	0.015	0.015	0.117	0.000	0.000
5	0.208	0.021	0.000	0.188	0.271	0.083	0.229	0.000	0.000
6	0.164	0.000	0.000	0.036	0.000	0.655	0.145	0.000	0.000
7	0.005	0.089	0.026	0.000	0.010	0.000	0.859	0.000	0.010
8	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.286	0.000	0.714

Lets summarize above models before proceeding with the feature engineering approach.

In [115]:

```
print()
from prettytable import PrettyTable
ptable = PrettyTable()
ptable.title = "*** Model Summary *** [Performance Metric: Log-Loss]"
ptable.field_names=["Model Name","Train","CV","Test","% Misclassified P
oints"]
ptable.add_row(["Naive Bayes","0.92","1.27","1.19","41"])
ptable.add_row(["KNN","0.45","1.12","1.09","37"])
ptable.add_row(["Logistic Regression With Class balancing","0.58","1.1
1","1.03","37"])
ptable.add_row(["Logistic Regression Without Class balancing","0.58",
"1.16","1.07","38"])
ptable.add_row(["Linear SVM","0.71","1.19","1.11","38"])
ptable.add_row(["Random Forest Classifier With One hot Encoding","0.64"
,"1.18","1.14","40"])
ptable.add_row(["Random Forest Classifier With Response Coding","0.04",
"1.38","1.31","47"])
ptable.add_row(["Stack Models:LR+NB+SVM","0.92","1.15","1.06","32"])
ptable.add_row(["Maximum Voting classifier","0.92","1.09","1.03","33"])
print(ptable)
print()
```

```
+------------------------------------------------+-------+------+------
+------------------------+
|                   Model Name                   | Train |  CV  | Test
| % Misclassified Points |
+------------------------------------------------+-------+------+------
+------------------------+
|                  Naive Bayes                   |  0.92 | 1.27 | 1.19
|           41           |
|                      KNN                       |  0.45 | 1.12 | 1.09
|           37           |
|    Logistic Regression With Class balancing    |  0.58 | 1.11 | 1.03
|           37           |
|  Logistic Regression Without Class balancing   |  0.58 | 1.16 | 1.07
|           38            |
|                   Linear SVM                   |  0.71 | 1.19 | 1.11

|           38            |
| Random Forest Classifier With One hot Encoding |  0.64 | 1.18 | 1.14
|           40            |
| Random Forest Classifier With Response Coding  |  0.04 | 1.38 | 1.31
|           47            |
|             Stack Models:LR+NB+SVM             |  0.92 | 1.15 | 1.06
|           32            |
|           Maximum Voting classifier            |  0.92 | 1.09 | 1.03
|           33            |
+------------------------------------------------+-------+------+------
+------------------------+
```

From above summary table we can observed that 'Logistic Regression With Class balancing' is better fit than others.So we will try countVectorizer features with both unigrams and bigrams to see whether it will reduce the log loss further or not.

Logistic Regression With Class Balancing

Gene Feature

In [0]:

```
#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", x_train))
```

```
# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))

# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [0]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = CountVectorizer(ngram_range=(1, 2))
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(x_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(x_test['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(x_cv['Gene'])

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

Variation Feature

In [0]:

```
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_train))

# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_test))

# cross validation gene feature
```

```
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_cv))
```

In [0]:

```
# one-hot encoding of variation feature.
variation_vectorizer = CountVectorizer(ngram_range=(1, 2))
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(x_train['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(x_test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv['Variation'])


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

Text Feature

In [120]:

```
# 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(x_train['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_count
s))


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

```
Total number of unique words in train data : 769905
```

In [0]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)

# 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_responseCoding.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_t
ext_feature_responseCoding.sum(axis=1)).T
```

In [0]:

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

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

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

Stack above three features

In [0]:

```
# 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_onehotCoding))
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(y_train['Class']))

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

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


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
cv_gene_var_responseCoding = np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))
```

```
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 [124]:

```
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.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 772195)
(number of data points * number of features) in test data =  (665, 772195)
(number of data points * number of features) in cross validation data = (532, 772195)
```

In [125]:

```
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.shape)
```

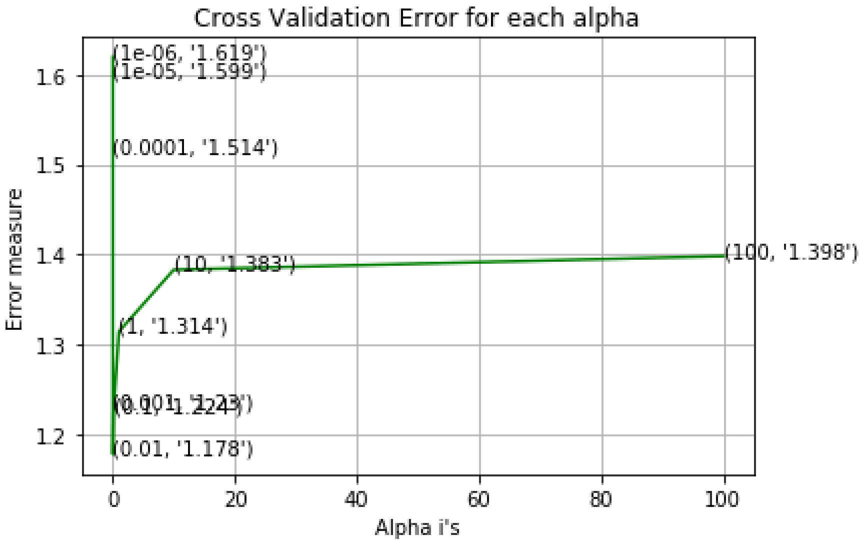
```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

Lets apply Logistic Regression

In [126]:

```
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], p
enalty='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))
```

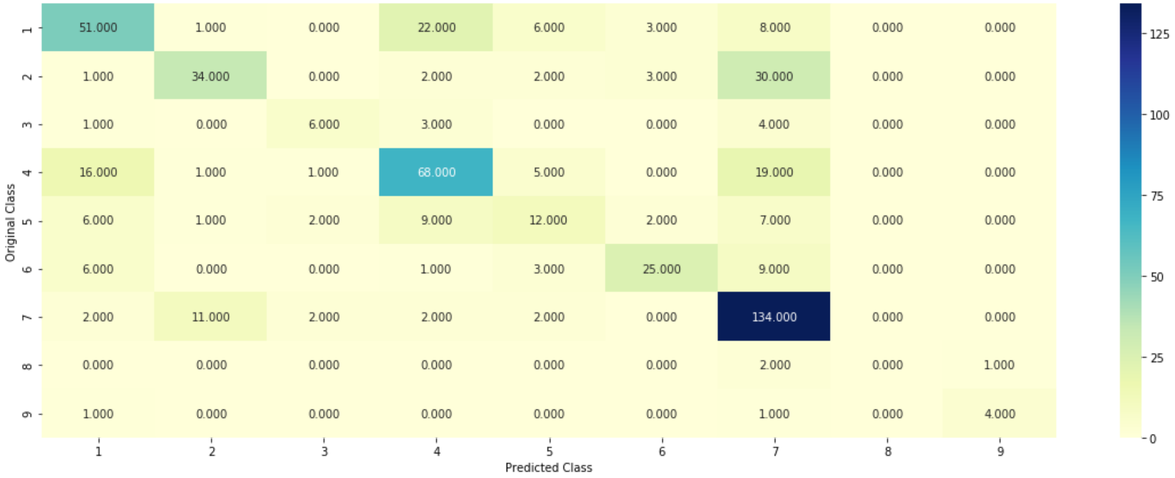
```
for alpha = 1e-06
Log Loss : 1.618979570864687
for alpha = 1e-05
Log Loss : 1.5985078398453019
for alpha = 0.0001
Log Loss : 1.5140178317499502
for alpha = 0.001
Log Loss : 1.2298290599548007
for alpha = 0.01
Log Loss : 1.1778765288585393
for alpha = 0.1
Log Loss : 1.2244858225045734
for alpha = 1
Log Loss : 1.3140400594998993
for alpha = 10
Log Loss : 1.3829604689253705
for alpha = 100
Log Loss : 1.3981993545248945
```



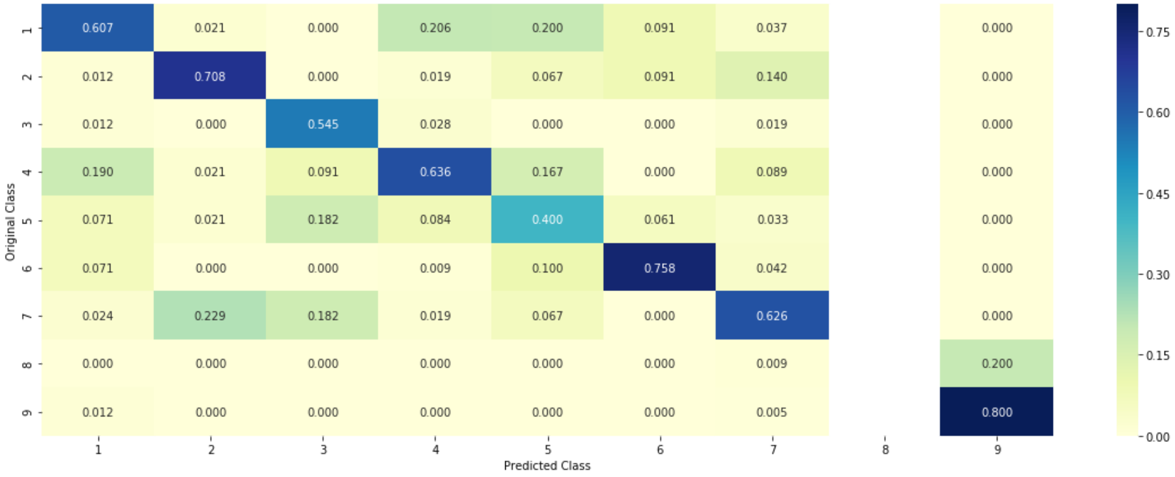
```
For values of best alpha =  0.01 The train log loss is: 0.75083584841
97278
For values of best alpha =  0.01 The cross validation log loss is: 1.
1778765288585393
For values of best alpha =  0.01 The test log loss is: 1.118561475652
6412
```

In [127]:
```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
enalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_o
nehotCoding, cv_y, clf)
```

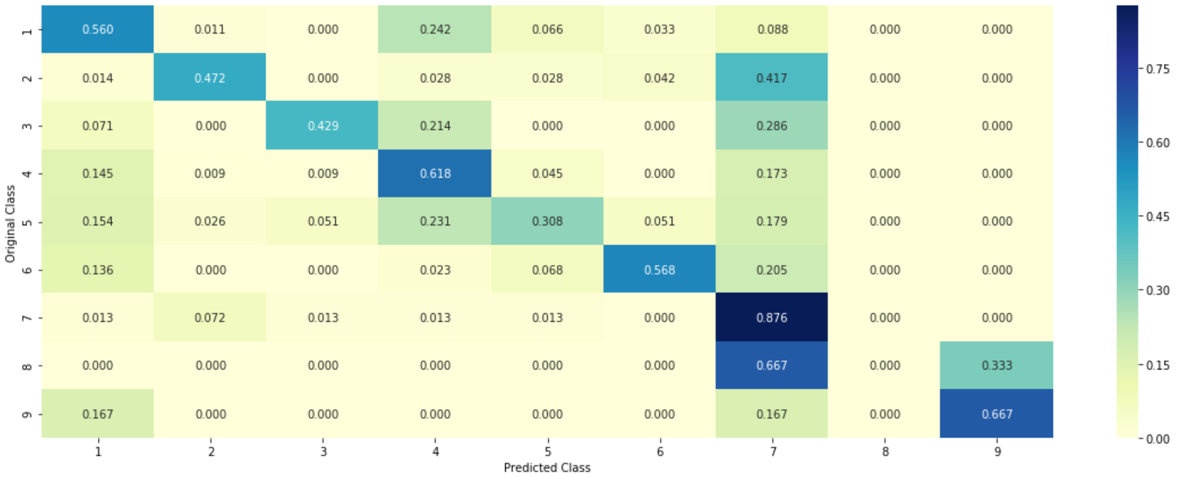
```
Log loss : 1.1778765288585393
Number of mis-classified points : 0.37218045112781956

-------------------- Confusion matrix --------------------
```



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



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



Still model does not decreases log loss values after using unigram and bigram features

---

Lets doing some featurization on the data and then apply logistic regression again

Lets merge gene and variation data into one list and apply TFidfVectorizer on top of it.

Gene Feature

In [0]:

```
result = pd.merge(data_variants, data_text,on='ID', how='left')
result.loc[result['TEXT'].isnull(),'TEXT'] = result['Gene'] +' '+result
['Variation']
y_true = result['Class'].values
```

```
result.Gene = result.Gene.str.replace('\s+', '_')
result.Variation = result.Variation.str.replace('\s+', '_')

x_train, x_test, y_train, y_test = train_test_split(result, y_true, stratify=y_true, test_size=0.2)

x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, stratify=y_train, test_size=0.2)
```

In [0]:

```
# get_gv_fea_dict: Get Gene varaition Feature Dict
def get_gv_fea_dict(alpha, feature, df):
    value_count = x_train[feature].value_counts()
    gv_dict = dict()
    for i, denominator in value_count.items():
        vec = []
        for k in range(1,10):
            cls_cnt = x_train.loc[(x_train['Class']==k) & (x_train[feature]==i)]
            vec.append((cls_cnt.shape[0] + alpha*10)/ (denominator + 90*alpha))
        gv_dict[i]=vec
    return gv_dict

# Get Gene variation feature
def get_gv_feature(alpha, feature, df):
    gv_dict = get_gv_fea_dict(alpha, feature, df)
    value_count = x_train[feature].value_counts()
    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])
    return gv_fea
```

In [0]:

```
#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", x_train))

# test gene feature
test_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_test))

# cross validation gene feature
cv_gene_feature_responseCoding = np.array(get_gv_feature(alpha, "Gene", x_cv))
```

In [0]:

```
# one-hot encoding of Gene feature.
gene_vectorizer = TfidfVectorizer()
train_gene_feature_onehotCoding = gene_vectorizer.fit_transform(x_train['Gene'])
test_gene_feature_onehotCoding = gene_vectorizer.transform(x_test['Gene'])
cv_gene_feature_onehotCoding = gene_vectorizer.transform(x_cv['Gene'])
```

Variation Feature

In [0]:

```
# alpha is used for laplace smoothing
alpha = 1

# train gene feature
train_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_train))

# test gene feature
test_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_test))

# cross validation gene feature
cv_variation_feature_responseCoding = np.array(get_gv_feature(alpha, "Variation", x_cv))
```

In [0]:

```
# one-hot encoding of variation feature.
variation_vectorizer = TfidfVectorizer()
train_variation_feature_onehotCoding = variation_vectorizer.fit_transform(x_train['Variation'])
test_variation_feature_onehotCoding = variation_vectorizer.transform(x_test['Variation'])
cv_variation_feature_onehotCoding = variation_vectorizer.transform(x_cv['Variation'])
```

Text Feature

In [0]:

```
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


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.get(word,0)+90)))
            text_feature_responseCoding[row_index][i] = math.exp(sum_prob/len(row['TEXT'].split()))
            row_index += 1
    return text_feature_responseCoding
```

In [135]:

```
# building a CountVectorizer with all the words that occured minimum 3 times in train data
text_vectorizer = TfidfVectorizer()
```

```
train_text_feature_onehotCoding = text_vectorizer.fit_transform(x_train
['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 num
ber of times it occured
text_fea_dict = dict(zip(list(train_text_features),train_text_fea_count
s))


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

```
Total number of unique words in train data : 127468
```

In [0]:

```
dict_list = []
# dict_list =[] contains 9 dictoinaries each corresponds to a class
for i in range(1,10):
    cls_text = x_train[x_train['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 = extract_dictionary_paddle(x_train)


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[i]+90))
    confuse_array.append(ratios)
confuse_array = np.array(confuse_array)
```

In [0]:

```
#response coding of text features
train_text_feature_responseCoding  = get_text_responsecoding(x_train)
test_text_feature_responseCoding  = get_text_responsecoding(x_test)
cv_text_feature_responseCoding  = get_text_responsecoding(x_cv)

# 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_responseCoding.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_t
ext_feature_responseCoding.sum(axis=1)).T
```

In [0]:

```
test_text_feature_onehotCoding = text_vectorizer.transform(x_test['TEX
T'])
cv_text_feature_onehotCoding = text_vectorizer.transform(x_cv['TEXT'])
```

Features after feature engineering

In [0]:

```
# Collecting all the genes and variations data into a single list
gene_variation = []

for gene in data_variants['Gene'].values:
    gene_variation.append(gene)

for variation in data_variants['Variation'].values:
    gene_variation.append(variation)
```

In [0]:

```
tfidfVectorizer = TfidfVectorizer(max_features=1000)
text2 = tfidfVectorizer.fit_transform(gene_variation)
gene_variation_features = tfidfVectorizer.get_feature_names()

train_text = tfidfVectorizer.transform(x_train['TEXT'])
```

```
test_text = tfidfVectorizer.transform(x_test['TEXT'])
cv_text = tfidfVectorizer.transform(x_cv['TEXT'])
```

Stack above three features

In [0]:

```
train_gene_var_onehotCoding = hstack((train_gene_feature_onehotCoding,train_variation_feature_onehotCoding))
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))

# Adding the train_text feature
train_x_onehotCoding = hstack((train_gene_var_onehotCoding, train_text))
train_x_onehotCoding = hstack((train_x_onehotCoding, train_text_feature_onehotCoding)).tocsr()
train_y = np.array(list(x_train['Class']))

# Adding the test_text feature
test_x_onehotCoding = hstack((test_gene_var_onehotCoding, test_text))
test_x_onehotCoding = hstack((test_x_onehotCoding, test_text_feature_onehotCoding)).tocsr()
test_y = np.array(list(x_test['Class']))

# Adding the cv_text feature
cv_x_onehotCoding = hstack((cv_gene_var_onehotCoding, cv_text))
cv_x_onehotCoding = hstack((cv_x_onehotCoding, cv_text_feature_onehotCoding)).tocsr()
cv_y = np.array(list(x_cv['Class']))


train_gene_var_responseCoding = np.hstack((train_gene_feature_responseCoding,train_variation_feature_responseCoding))
test_gene_var_responseCoding = np.hstack((test_gene_feature_responseCoding,test_variation_feature_responseCoding))
cv_gene_var_responseCoding = np.hstack((cv_gene_feature_responseCoding,cv_variation_feature_responseCoding))
```

```
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 [142]:

```
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.shape)
```

```
One hot encoding features :
(number of data points * number of features) in train data =  (2124, 130673)
(number of data points * number of features) in test data =  (665, 130673)
(number of data points * number of features) in cross validation data = (532, 130673)
```

In [143]:

```
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.shape)
```

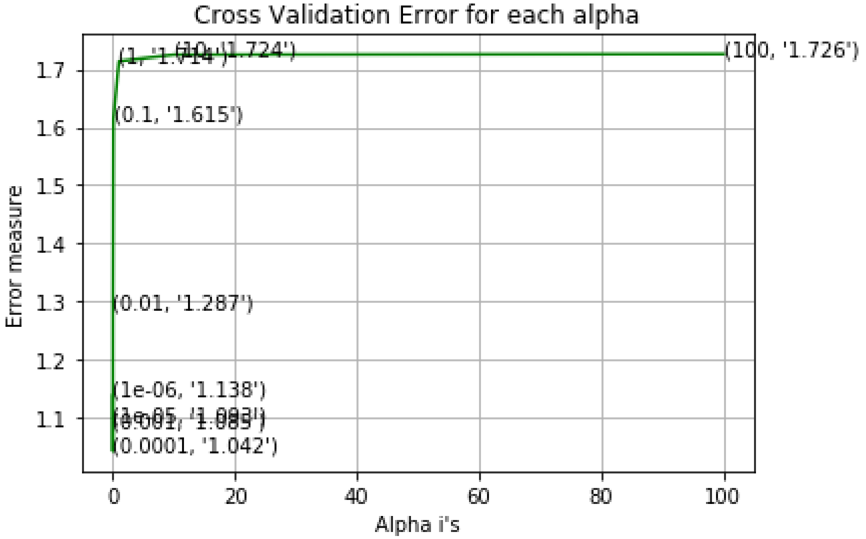
```
 Response encoding features :
(number of data points * number of features) in train data =  (2124, 27)
(number of data points * number of features) in test data =  (665, 27)
(number of data points * number of features) in cross validation data = (532, 27)
```

In [144]:

```
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], p
enalty='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 : 1.1379674427236748
for alpha = 1e-05
Log Loss : 1.0934758287879884
for alpha = 0.0001
Log Loss : 1.0415164684426224
for alpha = 0.001
Log Loss : 1.0846899485185262
for alpha = 0.01
Log Loss : 1.2873442003072824
for alpha = 0.1
Log Loss : 1.6149141635399866
for alpha = 1
Log Loss : 1.7138276615475843
for alpha = 10
Log Loss : 1.7244652901116457
for alpha = 100
Log Loss : 1.7255947796639737
```

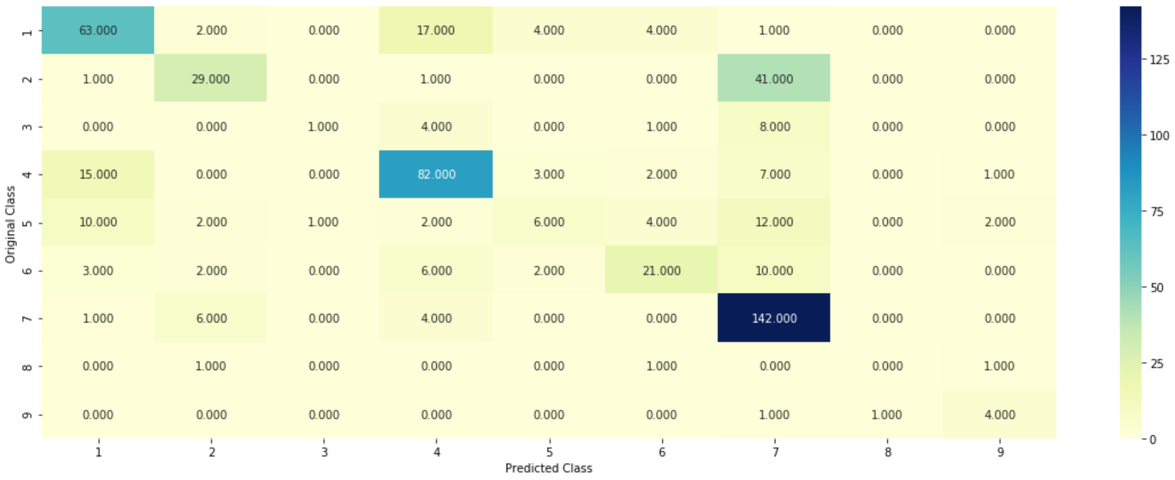


```
For values of best alpha =  0.0001 The train log loss is: 0.442949873
7224747
For values of best alpha =  0.0001 The cross validation log loss is:
1.0415164684426224
For values of best alpha =  0.0001 The test log loss is: 0.9712830857
383803
```

In [145]:

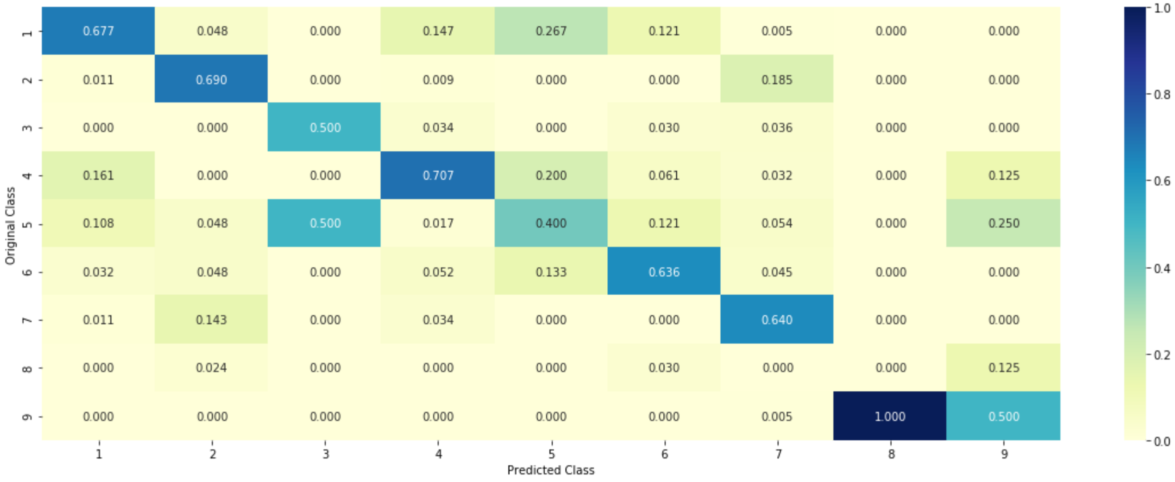
```
clf = SGDClassifier(class_weight='balanced', alpha=alpha[best_alpha], p
enalty='l2', loss='log', random_state=42)
predict_and_plot_confusion_matrix(train_x_onehotCoding, train_y, cv_x_o
nehotCoding, cv_y, clf)
```

```
Log loss : 1.0415164684426224
Number of mis-classified points : 0.3458646616541353

-------------------- Confusion matrix --------------------
```



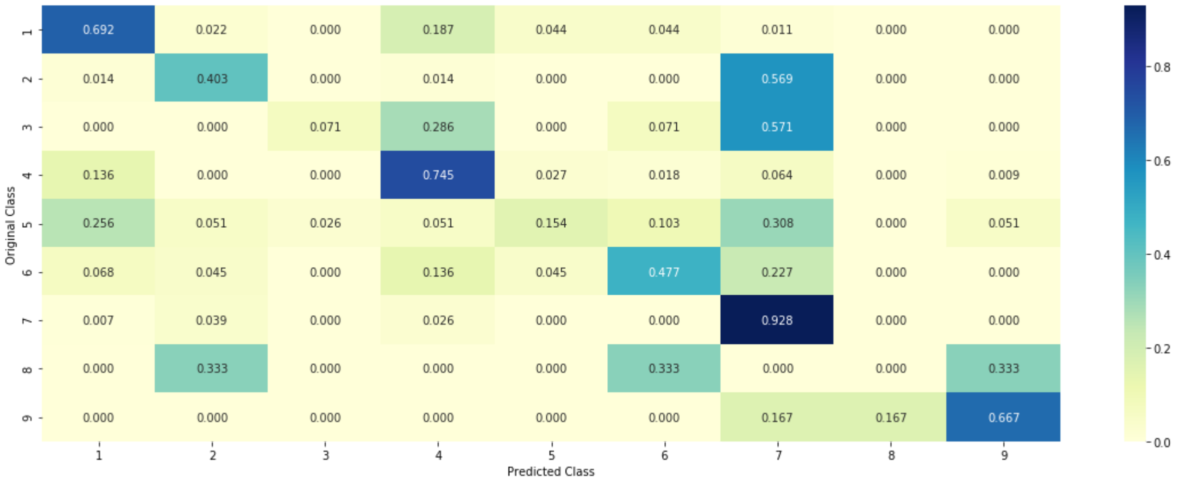
	1	2	3	4	5	6	7	8	9
1	63.000	2.000	0.000	17.000	4.000	4.000	1.000	0.000	0.000
2	1.000	29.000	0.000	1.000	0.000	0.000	41.000	0.000	0.000
3	0.000	0.000	1.000	4.000	0.000	1.000	8.000	0.000	0.000
4	15.000	0.000	0.000	82.000	3.000	2.000	7.000	0.000	1.000
5	10.000	2.000	1.000	2.000	6.000	4.000	12.000	0.000	2.000
6	3.000	2.000	0.000	6.000	2.000	21.000	10.000	0.000	0.000
7	1.000	6.000	0.000	4.000	0.000	0.000	142.000	0.000	0.000
8	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000
9	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	4.000

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



	1	2	3	4	5	6	7	8	9
1	0.677	0.048	0.000	0.147	0.267	0.121	0.005	0.000	0.000
2	0.011	0.690	0.000	0.009	0.000	0.000	0.185	0.000	0.000
3	0.000	0.000	0.500	0.034	0.000	0.030	0.036	0.000	0.000
4	0.161	0.000	0.000	0.707	0.200	0.061	0.032	0.000	0.125
5	0.108	0.048	0.500	0.017	0.400	0.121	0.054	0.000	0.250
6	0.032	0.048	0.000	0.052	0.133	0.636	0.045	0.000	0.000
7	0.011	0.143	0.000	0.034	0.000	0.000	0.640	0.000	0.000
8	0.000	0.024	0.000	0.000	0.000	0.030	0.000	0.000	0.125
9	0.000	0.000	0.000	0.000	0.000	0.000	0.005	1.000	0.500

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



Original Class \ Predicted Class	1	2	3	4	5	6	7	8	9
1	0.692	0.022	0.000	0.187	0.044	0.044	0.011	0.000	0.000
2	0.014	0.403	0.000	0.014	0.000	0.000	0.569	0.000	0.000
3	0.000	0.000	0.071	0.286	0.000	0.071	0.571	0.000	0.000
4	0.136	0.000	0.000	0.745	0.027	0.018	0.064	0.000	0.009
5	0.256	0.051	0.026	0.051	0.154	0.103	0.308	0.000	0.051
6	0.068	0.045	0.000	0.136	0.045	0.477	0.227	0.000	0.000
7	0.007	0.039	0.000	0.026	0.000	0.000	0.928	0.000	0.000
8	0.000	0.333	0.000	0.000	0.000	0.333	0.000	0.000	0.333
9	0.000	0.000	0.000	0.000	0.000	0.000	0.167	0.167	0.667

by using logistic regression using penalty l2 and keeping range of alpha (-6,3) we could manage to reduce test log loss less than unity value i.e 0.97